ENCYCLOPÆDIA
BRITANNICA

MACROPÆDIA

The Encyclopædia Britannica
is published with the editorial advice
of the faculties of the University of Chicago;
a committee of persons holding
academic appointments at the universities
of Oxford, Cambridge, London, and Edinburgh;
a committee at the University of Toronto;
and committees drawn from members of the faculties
of the University of Tokyo
and the Australian National University.

THE UNIVERSITY OF CHICAGO

"Let knowledge grow from more to more
and thus be human life enriched."

The New Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 12

Knowledge in Depth

FOUNDED 1768
15TH EDITION

Encyclopædia Britannica, Inc.
William Benton, Publisher, 1943–1973
Helen Hemingway Benton, Publisher, 1973–1974
Chicago/London/Toronto/Geneva/Sydney/Tokyo/Manila/Seoul

©1979
by Encyclopædia Britannica, Inc.
Copyright under International Copyright Union
All rights reserved under Pan American and
Universal Copyright Conventions
by Encyclopædia Britannica, Inc.

Printed in U.S.A.

Library of Congress Catalog Card Number: 77-94292
International Standard Book Number: 0-85229-339-9

Metamorphic Rocks

Metamorphic rocks are rocks that have recrystallized as a result of changes in the physical environment. Such changes may occur by reactions involving only the solid state (*i.e.*, the mineral grains) or, more commonly, by reaction in a fluid medium, which makes up a very small percentage of the volume of the rock at any given instant. Commonly, the fluid medium is an aqueous film present in rock pores and on the boundaries between mineral grains. In a sense, metamorphic rocks are the most common rock types of the solid Earth, primarily because the Earth is a dynamic system whose temperatures and pressures tend to fluctuate in space and time. Because the pressure–temperature (*P–T*) conditions to which rocks of all igneous and sedimentary types may be subjected are almost infinite, the variety of metamorphic rock types is very large indeed.

Mineral response to metamorphism

A very simple mineralogical system and its response to changing pressure and temperature provide a good illustration of what occurs in metamorphism. An uncomplicated sediment at the Earth's surface, a mixture of the clay mineral kaolinite [$Al_4Si_4O_{10}(OH)_8$] and the mineral quartz (SiO_2), provides a good example. Most sediments have small crystals or grain sizes but great porosity and permeability, and the pores are filled with water. As time passes, more sediments are piled on top of the surface layer, and it becomes slowly buried. Accordingly, the pressure to which it is subjected increases because of the load on top, or overburden. For rocks with a density of two to three grams per cubic centimetre, the pressure will increase by 200 to 300 bars (one bar equals atmospheric pressure at an altitude of about 100 metres [300 feet] above sea level) for each kilometre of overburden. At the same time, the temperature will increase because of radioactive heating within the sediment and heat flow from deeper levels within the Earth. On the average, the temperature increases by about 30° C for each kilometre (87° F per mile) of burial.

In the first stages of incremental burial and heating, few chemical reactions will occur in the sediment layer, but the porosity decreases, and the low-density pore water is squeezed out. This process will be virtually complete by the time the layer is buried by five kilometres of overburden. There will be some increase in the size of crystals; small crystals with a large surface area are more soluble and less stable than large crystals, and throughout metamorphic processes there is always a tendency for crystals to grow in size with time, particularly if temperature is rising, because it increases the speed of reaction.

Eventually, when the rock is buried to a depth at which temperatures of about 300° C (600° F) obtain, a chemical reaction sets in, and the kaolinite and quartz are transformed to pyrophyllite and water:

kaolinite + quartz → pyrophyllite + water

$$Al_4Si_4O_{10}(OH)_8 + 4SiO_2 \rightarrow 2Al_2Si_4O_{10}(OH)_2 + 2H_2O.$$

The exact temperature at which this occurs depends on the fluid pressure in the system, but in general the fluid and rock-load pressures tend to be rather similar during such reactions. The water virtually fights its way out by lifting the rocks. Thus, the first chemical reaction is a dehydration reaction leading to the formation of a new hydrate. The water released is itself a solvent for silicates and promotes the crystallization of the product phases.

If heating and burial are continued, another dehydration reaction sets in at about 400° C, in which the pyrophyllite is transformed to andalusite and quartz and water:

A second dehydration reaction

pyrophyllite → andalusite + quartz + water

$$Al_2Si_4O_{10}(OH)_2 \rightarrow Al_2SiO_5 + 3SiO_2 + H_2O.$$

After the water has escaped, the rock becomes virtually anhydrous, containing only traces of fluid in minute and small inclusions in the product crystals. Both of these dehydration reactions tend to be fast, because water, a good silicate solvent, is present.

Although the mineral andalusite is indicated as the first product of dehydration of pyrophyllite, there are three minerals with the chemical composition Al_2SiO_5. Each has unique crystal structures, and each is stable under definite *P–T* conditions (Figure 1). Such differing forms

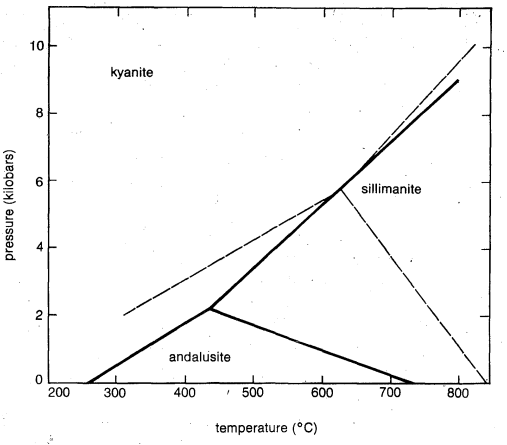


Figure 1: Pressure–temperature regions where the three polymorphic modifications of Al_2SiO_5 are stable. The dashed and solid lines are boundaries provided by different laboratories (see text).

with identical composition are called polymorphs. If pyrophyllite is dehydrated under high-pressure conditions, the polymorph of Al_2SiO_5 formed would be the mineral kyanite (the most dense polymorph). On the other hand, if the original temperature gradient persists, then at a depth of burial corresponding to about 700° C (1,300° F) the polymorphic transformation from andalusite to sillimanite will occur:

andalusite → sillimanite

$$Al_2SiO_5 \rightarrow Al_2SiO_5$$

Sillimanite is more stable than andalusite at high temperatures, but, unless a small amount of water is present in the rock, this reaction may not go to completion even in geological time. If sillimanite does form, however, then the temperature range within the Earth's crust will preserve the sillimanite–quartz assemblage unchanged.

If the forces leading to burial and sinking are reversed when the base of the sedimentary column has reached the sillimanite stage, then the thick column may be pushed up into a mountain range, permitting its observation. The reactions that proceeded during burial tend not to be reversed: with the water of the original sediments gone, the hydrates cannot reform, and chemical reaction rates are always faster in response to rising than to lowering temperatures. The re-exposed column would reveal the metamorphic history of the pile of sediments.

The types of reaction cited here are typical of all metamorphic changes. Gases are lost (hydrates lose water, carbonates lose carbon dioxide), and mineralogical phases undergo polymorphic or other structural changes; low-volume, dense mineral species are formed by high pressures, and less dense phases are favoured by high temperatures. Considering the immense chemical and mineralogical complexity of the Earth's crust, it is clear that the number of possible reactions is vast. In any given complex column of crustal materials some chemical reaction is likely for almost any incremental change in pressure and temperature. This is a fact of immense importance in unravelling the history and mechanics of the Earth, for such changes constitute a vital record and are perhaps the primary reason for the study of metamorphic rocks.

Observations show that pressure is only rarely hydrostatic (equal in all directions) at any point within the Earth's crust. In real cases, consequently, stresses operate that may lead to flow or fracture of materials. Such occurrences produce certain characteristic fabrics or structures in metamorphic rocks that may be observed at the level of the orientation of small crystals in a rock or as a pattern of folds in a mountain range. One of the principal characteristics of most metamorphic rocks is that the arrangement of crystals is not isotropic, or random, but that there is a strong preferred orientation related to the direction of stress components of pressure.

Economic aspects of metamorphic rocks

Many metamorphic reactions result in the gain or loss of water and other volatile compounds, such as the gaseous forms of hydrogen sulfide, carbon dioxide, and hydrochloric acid. At moderate temperatures and pressures, a good solvent for silicates is water; for carbonates, carbon dioxide in water; and for sulfides and elements such as gold, sodium chloride in water. Because large quantities of fluids may take part (perhaps 5 percent by weight of the rocks) in metamorphism, these fluids may transport economically important quantities of other minerals, which either may form deposits in veins or be disseminated in more permeable rocks. Various types of veins often have a very simple mineralogy and are characteristic of metamorphic rocks.

A large number of ore deposits are formed in metamorphic rocks. Practically all gold concentration occurs during metamorphism, even though sedimentary processes may lead to reworking and further concentration. Most important among the many other elements characteristically associated with metamorphic processes are copper, tungsten, zinc, lead, and mercury.

Other metamorphic rocks have useful application as building materials, often because of their peculiar grain sizes and fabrics. Thus, slates characterized by their ability to split or show slaty cleavage have applications wherever erosion- and corrosion-resistant thin materials are wanted. The uses of marble are associated with the increased grain size resulting from the metamorphism of limestones. Serpentines find application because of their beauty when polished; they form during metamorphism by the hydration of certain classes of basic igneous rocks (peridotites and dunites). At times, nickel deposits form during this process. Some metamorphic minerals are of value because of their physical properties or purity. Thus, garnet is often used in abrasives and kyanite as a source of pure aluminum silicate in certain refractory materials.

This article includes treatment of the formation of metamorphic rocks, their physical and chemical characteristics, and their occurrence and distribution as metamorphic facies—*i.e.*, rocks that are associated with a particular set of formative conditions that has given rise to distinctive mineral assemblages or other rock characteristics. For further information on the properties of rocks and minerals and the processes of metamorphism, see MINERALS; SILICATE MINERALS; ROCKS, PHYSICAL PROPERTIES OF; ROCK METAMORPHISM, PRINCIPLES OF; and GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES. See also IGNEOUS ROCKS and SEDIMENTARY ROCKS for an overview of rock types that become subject to metamorphism; and ROCK DEFORMATION; MOUNTAIN-BUILDING PROCESSES; and ORE DEPOSITS for coverage of topics intimately related to metamorphic processes.

FORMATION AND CLASSIFICATION OF METAMORPHIC ROCKS

The common metamorphic rocks observed on the Earth's surface, in mountain ranges and areas of deep erosion, are formed from materials within the crust. The average thickness of the crust is around 30 kilometres (20 miles) under the continents and six to eight kilometres (four to five miles) under the oceans. These thicknesses correspond to load pressures on the order of ten kilobars and two to three kilobars, respectively (one kilobar is very nearly 1,000 atmospheres pressure). Occasionally, crustal thickness may approach 60 kilometres (40 miles), and, hence, pressures may approach 20 kilobars. From place to place on the Earth's surface, thermal gradients are quite variable. The average figure quoted is about 30° C per kilometre of depth, but in some regions the thermal gradient is much less (about 10° C per kilometre), and in some active hydrothermal regions it may exceed 100° C per kilometre (300° F per mile) (see EARTH, HEAT FLOW IN). If the temperature at any point in the crust becomes very high, melting will commence, and metamorphic processes will give way to igneous processes. Thus, the melting temperatures of common rocks (800° C [1,500° F] for granites; 1,200° C [2,200° F] for basalts) represent the upper limits of metamorphic temperatures, and in a general way the limits of formation of common metamorphic rocks can be considered to be 100°–1,200° C (200°–2,200° F) and 1–20,000 bars.

Thermal gradients and pressures

Because most of the Earth's mantle (the region beneath the crust) is solid, metamorphic processes may also occur there. Mantle rocks are seldom observed at the surface, because they are too dense to rise, but occasionally a glimpse is presented by their inclusions in solid volcanic materials and in rapid gaseous volcanic extrusions. Such rocks may represent samples from a depth of a few hundred kilometres, where pressures of about 100 kilobars may be operative. The class of rocks known as kimberlites, which contains diamond, the high-pressure form of carbon, is an example. Experiments at high pressure have shown that few of the common minerals that occur at the surface will survive at depth within the mantle without changing to new high-density phases in which atoms are packed more closely together. Thus, the common form of SiO_2, quartz, with a density of 2.65 transforms to a new phase, stishovite, with a density of 4.29. Such changes are of critical significance in the geophysical interpretation of the Earth's interior.

Geologists believe the Earth is about 4,600,000,000 years old. This is also the age of meteoric materials and probably the age of the Moon. The oldest rocks on Earth are found on the continents; ages of about 3,500,000,000 years from Africa and the Soviet Union have been substantiated. These ages were obtained from analyses of igneous rocks, but the rocks themselves intrude metamorphosed sedimentary rocks. The oldest rocks found on earth at this time are metamorphic rocks from Greenland with an age of 3,800,000,000 years. It may be argued, therefore, that the oldest rocks on Earth are metamorphic and that metamorphic processes of the same type as those existing today have been operating since the earliest times, when the Earth was cool enough to allow fragments of the crust to survive. Thus, metamorphic processes affect almost all rocks except those that are being formed at the surface today.

The Earth's oldest rocks

Because metamorphism represents a response to changing physical conditions, those regions of the Earth's surface where dynamic processes are most active will also be regions where metamorphic processes are most intense and easily observed. The vast region of the Pacific margin, for example, with its seismic and volcanic activity, is also a region in which materials are being buried and metamorphosed intensely. In fact, the margins of continents and regions of mountain building are also regions where metamorphic processes proceed with intensity. But in quiet places, where sediments may accumulate at slow rates, less spectacular changes occur; these record changing conditions of pressures and temperatures that act upon each mineral grain. Metamorphic rocks are therefore distributed throughout the geologic column.

Types of metamorphism. It is convenient to distinguish several general types of metamorphism in order to simplify the description of the various metamorphic phenomena. Recognized here are contact, regional, hydrothermal, dynamic, and retrograde metamorphism, each of which will be described in turn.

Contact metamorphism. Whenever the crust is invaded at any level by silicate melts (magmas, from which igneous rocks crystallize within the Earth), they perturb the normal thermal regime and cause a heat increase in the vicinity. If a mass of basaltic liquid coming from the upper mantle is trapped in the crust and crystallizes there, it will heat up the surroundings; and the amount of heating and its duration will be a direct function of the mass of igneous material and its shape. Contact-metamorphic phenomena thus occur in the vicinity of hot igneous materials and at any depth. Under such circumstances pressure and temperature are not simply correlated. Thermal gradients are often very steep unless the igneous mass is very large. Contact aureoles—the surrounding zones of rock that become altered or metamorphosed—vary in thickness from inches (around tabular bodies such as dikes and thin sills) to several kilometres (around large granitic intrusions).

Contact-aureole thickness

If small fragments of rock are totally enclosed in a magma, they may be heated to the temperature of the magma itself. Their metamorphism represents an upper limit to temperature and is sometimes called pyro-metamorphism.

Regional metamorphism. The general term applied to large-scale metamorphism that affects either sedimentary or igneous rocks that are subject to burial is regional metamorphism. Normally there is a simple relationship between depth of burial, pressure, and temperature. Metamorphic rocks are developed on the scale of a mountain range, but among systems (*e.g.*, the Alps or the Urals) the pattern of thermal gradients may differ. Stress is normally operative, and the rocks produced by regional metamorphism have a well-developed fabric.

When low pressures are associated with regional metamorphism, the term burial metamorphism is applied. It occurs on a large scale, and the general distinguishing feature is the presence of a rather low-temperature mineral assemblage and often a lack of any pronounced mineral fabric.

Hydrothermal metamorphism. Changes that occur in rocks near the surface, where there is intense activity of hot water, are categorized as hydrothermal metamorphism. Such areas include Yellowstone National Park, United States; Wairakei, New Zealand; and the Salton Sea, California. It is now generally recognized that the circulating groundwaters that often become heated by proximity to igneous materials produce the metamorphism. Migration of chemical elements, vein formation, and other kinds of mineral concentration may be extreme on account of the large volumes of water circulated.

Dynamic metamorphism. When directed pressure or stress is the dominant agent of metamorphism, it is termed dynamic; other terms are dislocation, kinematic, and mechanical metamorphism. Mineralogical changes occurring on a fault plane provide an obvious example. In some such cases, the action may simply be a grinding up of existing grains or realignment of minerals that have non-equidimensional crystals. If the action is intense, friction may even lead to melting.

Retrograde metamorphism. Two reasons explain why metamorphic reactions that occur in response to rising pressure and temperature are not reversed when the rocks ultimately are returned to the Earth's surface, when pressures and temperatures are lower. First, if prograde (initial) reactions involve loss of volatile constituents such as water and carbon dioxide, then, unless these can be supplied again during unloading (erosional stripping away of the overlying rocks), the changes cannot be reversed. This is the most general case. Second, reaction rates generally increase with temperature; thus, prograde reactions are faster than retrograde reactions. Nevertheless, there are few metamorphic rocks that do not show at least some traces of retrograde processes, and these traces may record details of the unloading history.

The general irreversibility of metamorphic reactions

Metamorphic grade refers to the pressure–temperature relations that are associated with particular metamorphic minerals or mineral assemblages. High-grade metamorphism involves minerals produced under high temperatures and pressures; low-grade metamorphism, the reverse. When rocks are subjected to more than one metamorphic event, the first event may be of a higher or lower grade than subsequent events. If the first metamorphic event affected a given rock at a higher grade than conditions of a later event, the rock may be quite unresponsive chemically, even though new deformation structures may appear. But, if the later metamorphic phase carries the rock into higher grades, the first event may be obliterated, totally or partly. Careful studies of rock textures may reveal the primary metamorphism. Rocks that are products of dry metamorphism, such as eclogites and granulites, are highly susceptible to later events. Low-grade facies such as glaucophane–lawsonite schists are also likely to be altered by later, higher grade changes.

Types of metamorphic rocks. Because of the diverse chemistry, mineralogy, and primary origin of metamorphic rocks and because of the diverse fabrics or textures that may develop depending on the stresses that may operate during their formation, there is no simple, universally used classification of these rocks. In addition, different countries may have their own special terms. In general, any classification of metamorphic rocks tends to stress either their fabric, mineralogy, or primary origin. Some of the most common metamorphic rock types will be described here. Rocks in which metamorphic minerals are easily seen by eye or hand lens and in which the mineral grains have a highly orientated fabric are called schists (Figure 2). Grains of acicular (needlelike) or platy minerals (amphiboles and micas) tend to lie with their long directions parallel or their planar directions parallel. Often the rocks show a pronounced mineralogical layering; quartz layers a few millimetres or centimetres in thickness may lie between mica layers, for example. Other words often qualify schist: greenschist is a schist rich in the green mineral chlorite; blueschist is rich in the blue amphibole, glaucophane; mica-schist is rich in mica; and a graphite-schist is rich in graphite. Schists that are rich in the amphibole hornblende and are often derived by metamorphism of common igneous rocks of the basalt–gabbro type are called amphibolites.

Rock types according to structures and fabrics

A very fine grained metamorphic rock (usually developed from clay-rich sediments) exhibiting perfect planar layering and perfection of splitting into layers (slaty cleavage) is slate. Such rocks are normally rich in micas

By courtesy of W.S. Fyfe

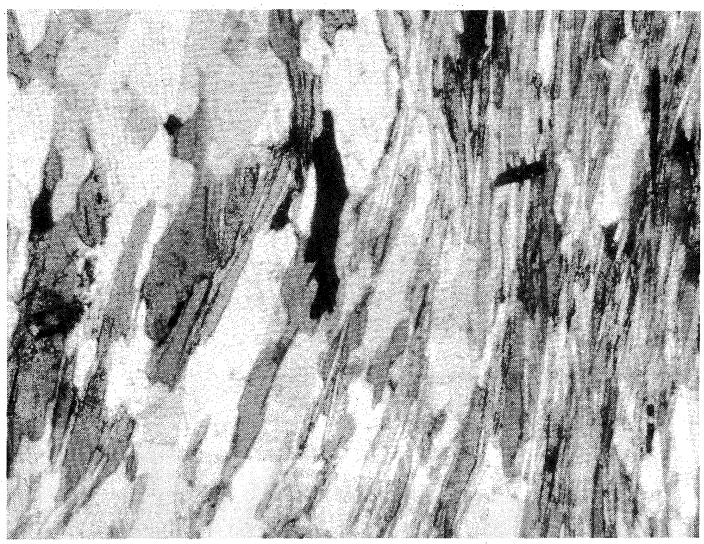


Figure 2: Photomicrograph of a staurolite–biotite schist from Scotland showing strong orientation of biotite crystals (dark) in a matrix of quartz and feldspars (magnification 50 X).

and chlorites. As the intensity of metamorphism increases, a few large crystals may grow (porphyroblasts); such slates are sometimes termed spotted slates. As metamorphism proceeds, the average crystal size increases, and mineral segregation develops; the rock then may be termed a phyllite (Figure 3).

By courtesy of W.S. Fyfe

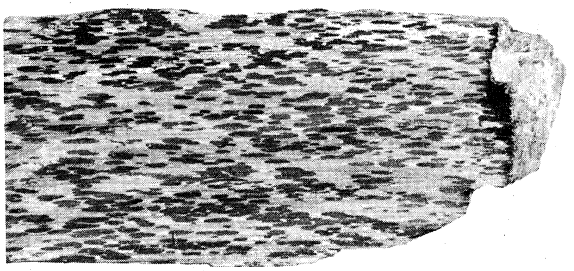


Figure 3: Cross-sectional view of a spotted phyllite from Cornwall, showing the intense stretching of spots (chlorite) in the quartz-rich (white) matrix.

A gneiss is produced by intense metamorphism, at high temperature and pressure. The grain size is coarser than that in schists, and layering is often well developed; mineral orientation is less perfect than in schists, however. Very common granitic gneisses (Figure 4) of Precambrian (more than 570,000,000 years in age) areas have been derived from metamorphism of granitic igneous rocks.

Hornfelses, formed by contact metamorphism, often show little sign of the action of directed pressure. They are fine-grained rocks in which crystals show little orientation. Granulites are products of ultrametamorphism (perhaps often the residue from partial fusion and melting). They tend to be coarse-grained and lack minerals such as amphiboles and micas capable of exhibiting strong orientation. Often they may resemble igneous rocks in appearance.

Rocks derived from the metamorphism of carbonate sediments containing calcite or dolomite are marbles (*q.v.*). The main result of metamorphism is an increase in grain size. Because of the rather equidimensional habit of calcite and dolomite crystals, they rarely appear schistose unless they contain other minerals such as mica.

Mylonites and cataclasites are rocks in which the texture is the result of mechanical shattering of grains. Often they show only slight if any development of new minerals. They form on fault planes or in zones of intense shearing. If the crustal rocks have an appropriate composition, phyllonites may develop where new mica crystals grow parallel to the shearing direction. If shearing is extreme, melting may occur, locally producing a pseudotachylite. Tachylite is a term applied to certain types of glasses formed by rapid cooling of molten rocks.

By courtesy of W.S. Fyfe

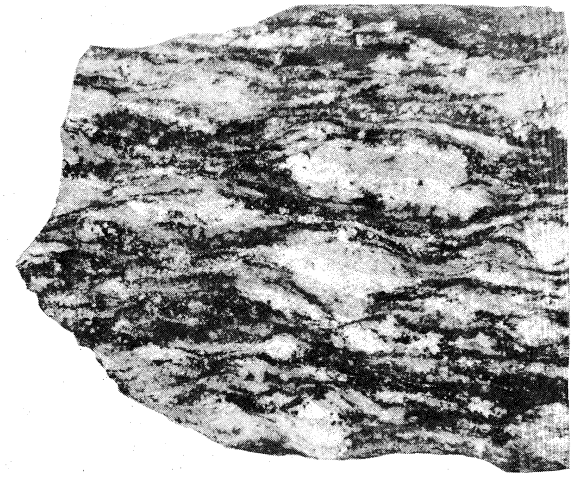


Figure 4: Typical granitic gneiss from Africa with large crystals of feldspar and a matrix containing dark biotite flowing around them.

Most of the above terms indicate structural or fabric classification of metamorphic rocks. Sometimes terms are used to indicate chemical features. Several types of schists, for example, include the following: pelitic schists contain much Al_2O_3 and often are derivatives of clay-rich sediments; quartzofeldspathic schists are high in SiO_2 and feldspars and often are derivatives of sandstones or quartz-rich igneous rocks; calcareous schists have a high content of lime (CaO) and often are derivatives of impure limestones, dolomites, or calcareous muds; and basic schists contain the elements of basic igneous rocks, namely, calcium, magnesium, and iron.

Rock types according to chemistry and primary origins

In addition to the fabric and chemical features of rocks, the primary origin of the material that has been metamorphosed serves as the basis of such names as meta-graywacke, the metamorphic product of the voluminous sediment type called graywacke (a sandstone with muddy matrix and profuse rock fragments); meta-basalt, which is derived from basalt; and meta-sediment, a general term stressing the sedimentary origin of the metamorphic rocks.

PHYSICAL AND CHEMICAL CHARACTERISTICS

Metamorphic minerals. It is clear that many more metamorphic minerals than igneous or sedimentary minerals might be expected to exist, simply because the conditions of formation of metamorphic minerals cover a much wider range of temperatures and pressures than those of igneous or sedimentary environments. The most common characteristic minerals of metamorphic rocks are listed in the accompanying Table. Generally, if such minerals are found well developed in a rock, then the rock will be metamorphic. Under some conditions, however, a sediment may contain such mineralogical debris, as from metamorphic rocks that are being weathered and eroded. In such a case, fabric and microscopic study may be needed to identify the rock as metamorphic. To stress changes that occur in response to increasing temperature and pressure, the common metamorphic minerals are listed in that order—*e.g.*, quartz is associated with the lowest temperatures and pressures. Minerals marked with an asterisk may have a wide range of stability; quartz, for example, may be stable in sediments, igneous melts, or almost any type of metamorphic rock. Phases marked with a dagger (†) are minerals of a composition for which more than one form (polymorph) may be found. In some cases a single mineral may be characteristic of a metamorphic facies, but more commonly an assemblage is needed to indicate conditions of metamorphism.

Metamorphic minerals display the complete range of silicate structural types from low-density, open-type structures (*e.g.*, zeolites) to high-density phases, such as garnets and jadeite. Changes in the atomic structural arrangement reflect the geothermal regime quite clearly.

Rock composition. Common metamorphic rock types have essentially the same chemical composition as what must be their equally common igneous or sedimentary precursors. Common greenschists have essentially the same compositions as basalts; marbles are like limestones; slates are similar to mudstones or shales; and many gneisses are like granodiorites. In general, then, the chemical composition of a metamorphic rock will closely reflect the primary nature of the material that has been metamorphosed. If there are significant differences in composition between a parent material and the metamorphic equivalent, such differences tend to affect only the most mobile (soluble) or volatile elements; water and carbon dioxide contents change significantly, for example. An impure limestone in the system $CaCO_3$–SiO_2 (calcium carbonate–silica) may end up in the system CaO–SiO_2 (calcium oxide–silica). An extreme example is provided by buried salt deposits. In their primary state the chemistry of evaporites (*q.v.*) commonly falls in the system $CaCO_3$–NaCl–$CaSO_4$–H_2O (calcium carbonate–sodium chloride–calcium sulfate–water). They are not represented in the more advanced stages of metamorphism except perhaps by rocks in the system $CaCO_3$–$CaSO_4$ (calcium carbonate–calcium sulfate). Often the greatest part of the rock has been dissolved by metamor-

Relation of metamorphic and original rock composition

Common Minerals of Metamorphic Rocks

mineral	composition
Quartz*†	SiO_2
Kaolinite	$Al_4Si_4O_{10}(OH)_8$
Montmorillonite	$(\frac{1}{2}Ca, Na)_{0.7}(Al, Mg, Fe)_4[(Si, Al)_8O_{20}](OH)_4 \cdot nH_2O$
Albite*	$NaAlSi_3O_8$
Calcite*†	$CaCO_3$
Dolomite*	$CaMg(CO_3)_2$
Analcite	$NaAlSi_2O_6 \cdot H_2O$
Laumontite*	$CaAl_2Si_4O_{12} \cdot 4H_2O$
Adularia†	$KAlSi_3O_8$
Chlorite*	$(Mg, Al, Fe)_{12}[(Si, Al)_8O_{20}](OH)_{16}$
Prehnite	$Ca_2Al(AlSi_3O_{10})(OH)_2$
Pumpellyite	$Ca_4(Mg, Fe^2)(Al, Fe^3)_5O(OH)_3(Si_2O_7)_2(SiO_4)_2 \cdot 2H_2O$
Stilpnomelane	$(K, Na, Ca)_{0\text{-}1.4}(Fe^3, Fe^2, Mg, Al, Mn)_{5.9\text{-}8.2}(Si_8O_{20})(OH)_4(O, OH, H_2O)_{3.6\text{-}8.5}$
Epidote*	$Ca_2Fe^3Al_2O(OH)(Si_2O_7)(SiO_4)$
Muscovite*	$KAl_2(AlSi_3O_{10})(OH)_2$
Magnetite*	Fe_3O_4
Hematite*	Fe_2O_3
Pyrite	FeS_2
Glaucophane	$Na_2Mg_3Al_2(Si_8O_{22})(OH, F)_2$
Jadeite	$NaAlSi_2O_6$
Omphacite	a complex solid solution mainly of $NaAlSi_2O_6 - CaMgSi_2O_6 - NaFeSi_2O_6$
Lawsonite	$CaAl_2(Si_2O_7)(OH)_2 \cdot H_2O$
Aragonite†	$CaCO_3$
Magnesite	$MgCO_3$
Tremolite	$Ca_2Mg_5(Si_8O_{22})(OH)_2$
Actinolite	$Ca_2(Mg, Fe)_5(Si_8O_{22})(OH)_2$
Talc	$Mg_3(Si_4O_{10})(OH)_2$
Serpentine†	$Mg_3(Si_2O_5)(OH)_4$
Hornblende*	$(Na, K)_{0\text{-}1}Ca_2(Mg, Fe^2, Fe^3, Al)_5(Si_{6\text{-}7}Al_{2\text{-}1}O_{22})(OH, F)_2$
Plagioclase*	$(NaAlSi_3O_8, CaAl_2Si_2O_8)$
Kyanite†	Al_2SiO_5
Sillimanite†	Al_2SiO_5
Andalusite†	Al_2SiO_5
Microcline†	$KAlSi_3O_8$
Garnet*	a complex solid solution of pyrope $Mg_3Al_2Si_3O_{12}$ almandine $Fe^2{}_3Al_2Si_3O_{12}$ spessartite $Mn_3Al_2Si_3O_{12}$ grossularite $Ca_3Al_2Si_3O_{12}$ andradite $Ca_3Fe^3{}_2Si_3O_{12}$
Staurolite	$(Fe^2, Mg)_2(Al, Fe^3)_9O_6(SiO_4)_4(O, OH)_2$
Chloritoid	$(Fe^2, Mg, Mn)_2(Al, Fe^3)Al_3O_2(SiO_4)_2(OH)_4$
Anthophyllite	$Mg_7Si_8O_{22}(OH)_2$
Orthoclase†	$KAlSi_3O_8$
Anorthite	$CaAl_2Si_2O_8$
Diopside	$CaMgSi_2O_6$
Phlogopite*	$KMg_3(AlSi_3O_{10})(OH)_2$
Scapolite	$(Na, Ca, K)_4[Al_3(Al, Si)_3Si_6O_{24}](Cl, SO_4, CO_3OH)$
Biotite*	$K_2(Mg, Fe^2)_{6\text{-}4}(Fe^3, Al, Ti)_{0\text{-}2}(Si_{6\text{-}5}Al_{2\text{-}3}O_{20})(OH, F)_4$
Cordierite	$Al_3(Mg, Fe^2)_2(Si_5AlO_{18})$
Wollastonite	$CaSiO_3$
Forsterite	Mg_2SiO_4
Hypersthene	$(Mg, Fe)SiO_3$
Enstatite	$MgSiO_3$
Tridymite†	SiO_2
Corundum	Al_2O_3
Periclase	MgO
Mullite	$3Al_2O_3 \cdot 2SiO_2$
Sanidine†	$KAlSi_3O_8$
Larnite	Ca_2SiO_4

*Indicates wide range of stability. †Indicates more than one form (polymorph) exists.

phic fluids, which have carried it to the surface. These fluids, rich in halides, or salts, may play a key role in the transport of ore minerals.

Metamorphic rocks only rarely exhibit a chemical composition that is characteristically "metamorphic." This statement is equivalent to saying that diffusion of materials in metamorphism is a slow process, and various chemical units do not mix on any large scale. But occasionally, particularly during contact metamorphism, diffusion may occur across a boundary of chemical dissimilarity leading to rocks of unique composition. If a granite is emplaced into a limestone, the contact region may be flooded with silica and other components leading to the formation of a metasomatic rock. Often such contacts are chemically zoned. A simple example is provided by the metamorphism of magnesium-rich igneous rocks in contact with quartz-rich sediments. A zonation of the type serpentine–talc–quartz may be found such as:

$$Mg_6(Si_4O_{10})(OH)_8/Mg_3(Si_4O_{10})(OH)_2/SiO_2.$$

In this case the talc zone has grown by silica diffusion into the more silica-poor environment of the serpentine. Economic deposits are not uncommon in such situations; *e.g.*, the formation of the $CaWO_4$ (calcium tungstate) scheelite when tungstate in the form of WO_3 moves from a granite into a limestone contact. The reaction can be expressed as:

$$\underset{\text{calcite}}{CaCO_3} + WO_3\ (\text{solution}) \rightarrow \underset{\text{scheelite}}{CaWO_4} + CO_2\ (\text{gas}).$$

Metamorphic facies. Metamorphic petrologists working on contact metamorphism early in the 20th century introduced the idea of metamorphic facies to correlate metamorphic events. The concept was first defined in 1914 by a Finn, Pentti Eelis Eskola, as any rock of a metamorphic formation that has attained chemical equilibrium through metamorphism at constant temperature and pressure conditions, with its mineral composition controlled only by the chemical composition. In current usage, a metamorphic facies (Figure 5) is a set of metamorphic mineral assemblages, repeatedly associated in space and time, such that there is a constant and therefore predictable relation between mineral composition and chemical composition.

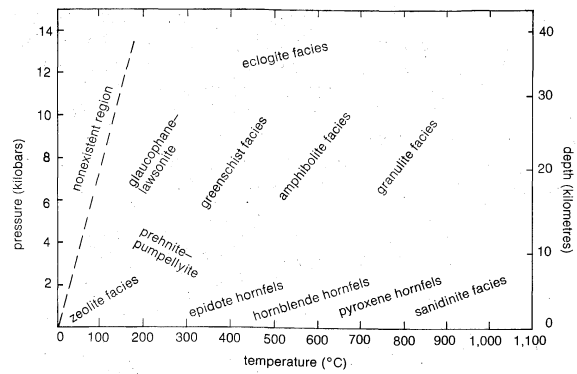


Figure 5: Approximate pressure–temperature regions of the major metamorphic facies (see text).

The facies concept is basically observational. In a layered sequence of sediments, for example, if each layer exhibits a different chemical composition, then it is reasonable to suppose that the different minerals within any single layer represent a response to similar conditions. In other words, the rocks belong to the same facies. In another region, in which the composition is equivalent to that in only one of the sediment layers in the first sequence, the facies concept permits correlation with the more complex assemblages. A facies is defined by a set of rock types, but the naming of facies tends to follow some single common type. The greenschist facies is named after schists formed from basaltic compositions; other facies, such as the zeolite facies, are named after some characteristic mineral type in the rock; still others are named after a textural feature. The definitions of facies suggest that during most metamorphic processes there is a close approach to some equilibrium state, and experimental studies of metamorphic processes substantiate this deduction.

Grades, zones, and facies series

The intensity of metamorphism is called metamorphic grade. A line on a map where the grade of metamorphism is similar would be called an isograd. During regional mapping of metamorphic rocks it is common to record where a particular mineral appears in progressive metamorphism. If the mineral was biotite or garnet in rocks of the greenschist facies, this would be recorded as the biotite or garnet zone. Zones of development of various minerals in contact metamorphism are often mapped and indicate the type of thermal gradient around an igneous intrusion.

If the thermal gradient is very low (high pressure at low temperature), transitions will occur from zeolite facies to glaucophane-schist facies with increasing burial. If the gradient is moderate, the transitions involve zeolite facies that lead to greenschist facies that lead to amphibolite facies that lead to facies of still higher grades. The concept of facies series arises from the attempt to describe and emphasize such major geothermal trends, which have

great significance in the major features of Earth dynamics. Thus, the zeolite–glaucophane-schist trend is found only when the rate of downbuckling of the crust is very fast. This type of motion occurs on continental margins and is associated with continental drift.

Textures and fabrics. The study of the fabric of metamorphic rocks has become a highly specialized subject. Modern work has been based on the classic investigation in 1930 of Bruno Sander of Innsbruck, Austria. The study of fabric or structure may reveal the nature and direction of forces acting during dynamic processes within the Earth. If hydrostatic forces prevailed, there would be no reason for the crystals in a rock to show pronounced preferred orientation of their axes, but such preferred orientation of crystals and mineral grains is perhaps the most striking difference between a metamorphic rock and other rock types.

The most obvious features of a metamorphic rock are certain planar features that are often termed s-surfaces. The simplest planar features may be primary bedding (akin to the layering in sedimentary rocks). As the rock crystallizes or recrystallizes under directed pressure, new crystals may grow in some preferred direction, sometimes subparallel to the primary bedding but often at new angles defining new planar structures. At the same time, folding of layers may occur, leading to folds on scales with amplitude of kilometres or millimetres. Fabric symmetry may be represented by nature of deformed fossils, pebbles in a conglomerate, or any objects with a known shape prior to deformation (Figure 6).

By courtesy of J. Ramsay, Imperial College, London

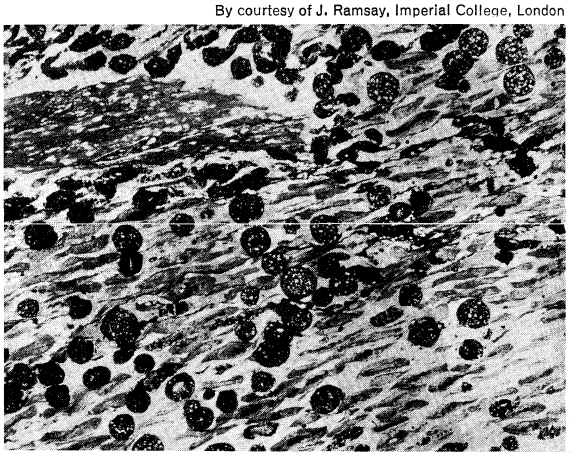


Figure 6: A deformed dolomite–calcite marble from Sardinia containing spherical fossil objects (oolites) about three millimetres in size. Oolites made of dolomite (black circles) are not deformed. Oolites made of calcite (gray), a weaker solid, are smeared out in S forms.

A few terms that commonly are used to describe several types of preferred orientation in metamorphic rocks include foliation, a general term describing any type of s-surface, bedding, or crystal orientation; slaty cleavage, a planar structure leading to facile cleavage that is normally caused by the preferred orientation of mica crystals; schistosity, a term used to describe repetitive and pronounced foliation of the type that is present in schists; and lineation, which is any linear structure, such as the axis of the fold, grooves on a fault plane, or the direction of stretching of pebbles.

Crystalloblastic fabric and laminations

The various mineral phases of a metamorphic rock have different physical properties and symmetries. When a rock is subjected to recrystallization in a stress field, different substances will behave differently according to such physical properties and symmetries. Some crystals always tend to grow in better formed crystals than others; rates of nucleation may differ, and this can lead to different patterns of growth of crystals—there may be a few large crystals or a mass of small crystals. Minerals can be arranged in order of their tendency to form crystals showing planar surfaces, namely, magnetite, garnet, epidote, mica, calcite, quartz, and feldspar. Crystals that tend to form large single crystals (*e.g.*, garnet) are termed porphyroblasts. Porphyroblastic crystals and their contained inclusions often record details of the mechanism of deformation and flow. A spectacular example is provided by the so-called snowball garnets, which have spiral trails of inclusions that indicate rotation during growth (Figure 7).

By courtesy of J. Ramsay, Imperial College, London

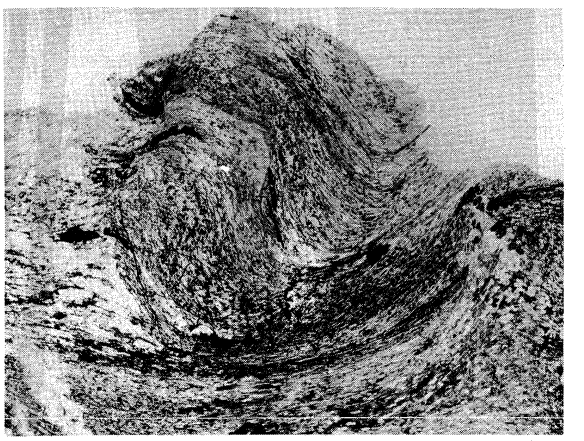


Figure 7: A large garnet crystal (enlarged; one centimetre) growing in a quartz–mica matrix. The garnet contains trails of inclusions and has rotated during growth in the flowing rock medium.

There is a distinct tendency for many types of metamorphic rocks to become laminated, and the separate laminae may have distinct chemical compositions. A rather homogeneous sediment may become inhomogeneous on a scale of millimetres or centimetres. When graywackes are metamorphosed within the greenschist facies, for example, laminae rich in quartz and feldspar alternate with others rich in epidote, chlorite, and muscovite. The precise causes of this process are not well-known, but it must result from a combination of extensive deformation accompanied by recrystallization. In a sense, it is a type of flow unmixing. It is often important to recognize that this structure need have no relation to original bedding in the unmetamorphosed sediments.

ROCKS OF THE PRINCIPAL FACIES

Facies associated with contact metamorphism. Possibly the simplest of all types of metamorphic processes to envisage are those in which hot silicate liquids are intruded into rocks at rather shallow depths. The hot liquids heat up the originally cool rocks, and the extent of the aureole that is formed depends on the size of the intruding igneous body and its temperature. The contact metamorphic rocks of the aureole zone often lack any obvious schistosity or foliation. The facies associated with contact metamorphism include the sanidinite, pyroxenite–hornfels, hornblende–hornfels, and albite–epidote–hornfels facies.

Sanidinite facies. Rocks of the sanidinite facies are represented by small fragments of aureole materials that have often been totally immersed in silicate liquids or by the aureole rocks surrounding volcanic pipes. Very high temperatures are attained, often at very low pressures. The dominant feature of the mineralogy of this facies is an almost complete lack of minerals containing water or carbon dioxide. Many of the minerals show similarity to those of igneous rocks themselves. If the duration of heating is short, adjustment to the imposed temperature is often imperfect.

Typical pelitic rocks

Pelitic rocks (high in Al_2O_3) contain minerals such as mullite, sillimanite, sanidine, cordierite, spinel, hypersthene, anorthite, tridymite, and even glass. One of the classic localities of such rocks is the island of Mull, off the west coast of Scotland, but in most regions of volcanism such rocks can be found.

Calcareous rocks (originally impure limestones or dolomites) tend to lose almost all their carbon dixoide, but pure calcite may survive. Typical metamorphic minerals are: quartz, wollastonite, anorthite, diopside, periclase,

and in some places (the classic is Scawt Hill in Northern Ireland) an array of complex calcium silicates such as spurrite, larnite, rankinite, melilite, merwinite, and monticellite. These minerals result from the addition of varying amounts of silica to impure mixtures of calcite and dolomite. In a general way the minerals of this facies are reminiscent of those of industrial slags.

Pyroxene–hornfels facies. Rocks of the pyroxene–hornfels facies are characteristically formed near larger granitic or gabbroic bodies at depths of a few kilometres or at pressures of a few hundred bars. The mineral assemblages are again largely anhydrous, but, unlike the sanidinite facies, the minerals reflect distinctly lower temperatures. One of the classic descriptions of such rocks is from the Oslo district of Norway.

In pelitic rocks, minerals such as quartz, orthoclase, andalusite, sillimanite, cordierite, hypersthene, and plagioclase occur. Sometimes the hydrate biotite is developed. In calcareous rocks the minerals found include plagioclase, diopside, grossularite, vesuvianite (a hydrate), wollastonite, and sometimes the more complex calcium silicates monticellite, melilite, spurrite, tilleyite, and clinohumite.

Hornblende–hornfels facies. A generally deeper level of contact metamorphism where pressures of a few kilobars may be active is represented by the hornblende–hornfels facies. Hydrated phases become stable, and the transition to regional metamorphism becomes apparent. Because of the generally greater depth, this type of aureole is often superposed on a more normal metamorphism, and the rocks may appear schistose and show the development of new thermally generated minerals on a pre-existing assemblage. This type of metamorphism develops the classic "spotted" texture in which new porphyroblasts grow in slates and phyllites of a previous episode of metamorphism. Typically, these rocks are developed near most of the world's large granite batholiths, where these have moved to higher levels in the Earth's crust.

Typical minerals of pelitic assemblages include quartz, muscovite, biotite, andalusite, sillimanite, cordierite, plagioclase, microcline, and staurolite. Calcareous assemblages include calcite, quartz, diopside, grossularite, plagioclase, wollastonite, brucite, talc, forsterite, tremolite, and clinozoisite. Basaltic compositions include plagioclase, hornblende, diopside, quartz, biotite, and almandine garnet.

Formation of spectacular skarns

When rather pure limestone and dolomite come into direct contact with granitic rocks, elements such as silicon, iron, magnesium, and aluminum diffuse into the limestone, forming spectacular rocks termed skarns. These rocks often consist of large garnet crystals (grossularite) with green diopside and vesuvianite or epidote.

Albite–epidote–hornfels facies. Rocks of the albite–epidote–hornfels facies are characteristically found as the outer zones of contact aureoles where the thermal episode fades out and the rocks pass into their regional grade of metamorphism. The mineral assemblages are quite similar to those found in regional greenschist-facies metamorphism, except for the presence of low-pressure phases such as andalusite. Characteristic minerals include quartz, muscovite, biotite, chlorite, andalusite, actinolite, calcite, dolomite, albite, and epidote.

Conditions of formation of the contact facies. The mineralogical reactions included in the rocks described above are varied, and laboratory studies have provided a rather exact picture of the conditions of formation of these facies. It should be stressed that all correspond to conditions of pressure and temperature such that a thermal accident or abnormal thermal gradient is necessary to produce the observed mineral phases. This accident is the rise of silicate melt from depth.

Facies associated with regional metamorphism. Regional metamorphism is associated with the major events of Earth dynamics, and the vast majority of metamorphic rocks are so produced. They are the rocks involved in the cyclic processes of erosion, sedimentation, burial, metamorphism, and mountain building, events that are all related to major convective processes in the Earth's mantle. Two particular trends of facies can be noted: a cold trend, in which the geothermal gradient is low (about 10°–15° C per kilometre [30°–45° F per mile]), leading from zeolite facies to glaucophane schist facies; and the more general trend for normal gradients (20°–30° C per kilometre [57°–87° F per mile]), from zeolite facies to greenschist facies to amphibolite facies to a zone of crustal fusion. In almost all such crustal events at different times and places, however, there is uniqueness as well as conformity to a general pattern. Metamorphic events in the European Alps, the Urals, and the Himalayas all show specific differences: to unravel such differences and their significance is a great task of metamorphic petrology.

Zeolite facies. In the zeolite facies, sediments and volcanic debris show the first major response to burial. Reactions are often not complete, and typical metamorphic fabrics may be poorly developed or not developed at all. This is the facies of burial metamorphism.

The facies of burial metamorphism

The zeolite facies was first described from southern New Zealand, but similar rocks have now been described from the rocks of many younger mountain regions of the Earth, particularly around the Pacific margin and the European Alps. Typically, the rocks are best developed where reactive volcanic materials (often partly glassy) are common and the characteristic minerals include zeolites, which are low-density, hydrated silicates, stable at temperatures rarely exceeding 300° C (600° F). Typical mineral assemblages include heulandite, analcite, quartz with complex clay minerals (montmorillonite), micaceous phases such as chlorite and celadonite, and the potassium feldspar, adularia. At higher grades of metamorphism, the zeolite laumonite and the feldspar albite dominate the mineral assemblage. In New Zealand these are developed in a rock column that is about 15 kilometres (nine miles) thick. Calcareous rocks (impure limestones) show very little response to this grade of metamorphism.

Prehnite–pumpellyite facies. Along with the zeolite facies, the prehnite–pumpellyite facies received little attention until about 1950. The first rocks of the facies were described in New Zealand and the Celebes. The facies is transitional, bridging the path to the glaucophane–lawsonite facies or the greenschist facies. It is particularly well developed in graywacke-type sediments. The two minerals prehnite and pumpellyite replace the zeolite minerals of the zeolite facies and are themselves replaced by epidote minerals in the greenschist facies and by lawsonite and pyroxenes in the glaucophane–lawsonite facies. Typical minerals in this facies are quartz, albite, prehnite, pumpellyite, chlorite, stilpnomelane, muscovite, and actinolite. Almost all the minerals are hydrated, and, except for chlorite, they bear little resemblance to the minerals of sediments. Again, the facies has been most described from younger mountain ranges of the Pacific margin.

Glaucophane–lawsonite schist facies. Rocks of the glaucophane–lawsonite schist facies are known to represent deep metamorphism under conditions of a low thermal gradient. The characteristic locale for this type of metamorphism appears to be along a continental margin being underthrust by an oceanic plate. Regions in which glaucophane schists are to be found are also regions of great seismic and volcanic activity, such as the Pacific margin. The best described examples of this class of metamorphism come from California, Japan, New Caledonia, Celebes, the Alps, and the Mediterranean region. At present there are no known examples of glaucophane schists predating the Paleozoic Era (570,000,000 to 225,000,000 years ago). Because of the common presence of the blue amphibole glaucophane (this facies is also sometimes termed the blueschist facies) and minerals such as garnet and jadeite, these schists are among the most attractive of metamorphic rocks.

Characteristic minerals of the facies include quartz, glaucophane, lawsonite, jadeite, omphacite, garnet, albite, chlorite, muscovite, paragonite, epidote, and kyanite. In calcareous rocks, calcite may be replaced by the high-pressure polymorph aragonite. Lawsonite, aragonite, and jadeite are found in no other metamorphic facies. In general, the facies is characterized by many high-density minerals reflecting a high pressure of formation.

Greenschist facies. The greenschist facies was once considered the first major facies of metamorphism proper. The name comes from the abundance of the green mineral chlorite in such rocks (Figure 8). Because chlorite and muscovite are ubiquitous and because both exhibit a platy crystal habit, these rocks normally show a highly developed foliation and often exhibit strong metamorphic differentiation. They have been described from practically every metamorphic terrain on Earth from earliest Precambrian to the young mountain regions. In fact, many of the Earth's oldest rocks (about 3,000,000,000 years old) of the continental-shield areas are in this facies, classic examples of which are in the Appalachians, Scottish Highlands, New Zealand, the European Alps, Japan, and Norway.

By courtesy of W.S. Fyfe

Figure 8: A typical greenschist from Scotland; the gray minerals are biotite and chlorite showing strong preferred orientation. The black crystals are magnetite, and the white areas are albite–quartz. Epidote and calcite are also present (magnification 50 X).

The dominant minerals of greenschists formed from silicate-rich sediments include quartz, albite, muscovite, chlorite, epidote, calcite, actinolite, magnetite, biotite, and paragonite. Minerals less common include the manganese-rich garnet spessartite, stilpnomelane, kyanite, rutile, sphene, pyrophyllite, and chloritoid. Calcareous rocks are dominated by calcite, dolomite, and quartz; the major carbonate minerals are thermally stable. It is only when large quantities of water flush away carbon dioxide or keep its partial pressure low that carbonate–silicate reactions take place liberating carbon dioxide. The typical minerals of this facies have low water contents as compared to the zeolite facies minerals.

Amphibolite facies. The amphibolite facies is the common high-grade facies of regional metamorphism, and, like the greenschist facies, such rocks are present in all ages from all over the world. Their characteristic feature is the development of the most common amphibole, hornblende, in the presence of a plagioclase feldspar and garnet. The rocks are normally highly foliated or schistose. Many zones or isograds subdividing the facies have been recognized, and classic studies have been made in the Scottish Highlands, New Hampshire, Vermont, Switzerland, and the Himalayas.

Hornblende, the most common amphibole

Characteristic minerals derived from pelitic rocks are quartz, muscovite, biotite, garnet, plagioclase, kyanite (sillimanite), staurolite, and orthoclase. Minerals derived from basaltic rocks include hornblende, plagioclase, garnet, epidote, and biotite. Those derived from calcareous rocks are calcite, diopside, grossularite (garnet), zoisite, actinolite (hornblende), scapolite, and phlogopite. Minerals from magnesium-rich ultrabasic rocks are chlorite, anthophyllite, and talc. In most common types, water is present only in minerals of the mica and amphibole families, and, with their water contents of only about 1 to 3 percent, dehydration is nearing its metamorphic climax.

Conditions of formation of the regional metamorphic facies. Most workers on the metamorphic facies of regional metamorphism considered above agree that progressive metamorphic reactions occur in a regime in which fluid pressure is about the same as the lithostatic (rock) pressure. Furthermore, because water is the common fluid phase and will be diluted by CO_2 (carbon dioxide) only in exceptional circumstances, the fluid pressure will approximately equal the pressure of water. Current thought on the regions of formation of metamorphic facies is shown in Figure 5, where the coordinates are the pressure of water and temperature. No matter how imperfect these estimates are in absolute terms, there is little doubt about their relative significance: they emphasize the uniqueness of metamorphic events.

The eclogite and granulite facies. The fact that the eclogite and granulite facies have no hydrated phases has led to considerable debate as to the place of such rocks among other metamorphic facies. They could represent dry equivalents of other metamorphic rocks, or their conditions of formation may be unique. Experimental studies have shown the extreme thermal stability of hornblende in rocks of basaltic composition. If a crustal rock contains this mineral, then at moderate pressures its dehydration reactions may lead to partial melting of the crust and the onset of igneous phenomena.

Eclogite facies. Eclogite facies are recognized only in rocks whose composition is near that of basalt. Two minerals dominate the mineralogy—omphacite pyroxene and garnet. The garnet is rich in the high-pressure species pyrope, and the omphacite is rich in the high-pressure pyroxene jadeite. Small amounts of minerals such as kyanite, the hydrate zoisite, and hornblende may be present. The rocks are of high density and frequently show little or no schistosity.

Because of the high density and composition, it was proposed long ago that part of the upper mantle might be made of eclogite. Such a view is supported by eclogitic intrusions in volcanic rocks and by eclogitic inclusions in diamond-bearing kimberlite, which must come from the upper mantle. Some workers also think that eclogites found in metamorphic terrains in Norway, California, and the European Alps could also come from the mantle by tectonic processes (crustal movements). Some eclogites are known to have been produced within the crust from former surface materials (*e.g.*, basalt lava flows). It also is known that eclogites form over a wide range of temperatures, perhaps from 400° to 1,000° C (750° to 1,800° F). These conditions overlap greenschist, amphibolite, and granulite facies temperatures.

Upper mantle as source of eclogite

Experimental studies have demonstrated that eclogites can be stable (except under certain upper-mantle conditions) only if water pressure is much lower than load pressure. The facies represents dry, high-pressure metamorphism of basaltic materials. The exact mechanism by which these metamorphic reactions take place is still a matter of considerable argument.

Granulite facies. The granulite facies is an anhydrous facies that develops gradationally from amphibolites but without any suggestion of a desiccated environment. Granulite-facies rocks are often found best developed in ancient Precambrian-shield areas of the continents. No large-scale development of such rocks has been observed in post-Cambrian metamorphism.

Rocks of this facies frequently have a granular texture quite similar to plutonic igneous rocks. Schistosity is only weakly developed. Typical minerals of the facies are quartz, alkali feldspar, garnet, plagioclase, cordierite, kyanite, sillimanite, and hypersthene. In calcareous members, dolomite, calcite, diopside, and forsterite occur; and it is in this facies that minerals of the scapolite family are best developed. Small amounts of hornblende are often present. A rare mineral occurring in this facies is sapphirine. The rock type charnockite (from Madras, India), essentially a hypersthene granite, is normally included in this facies.

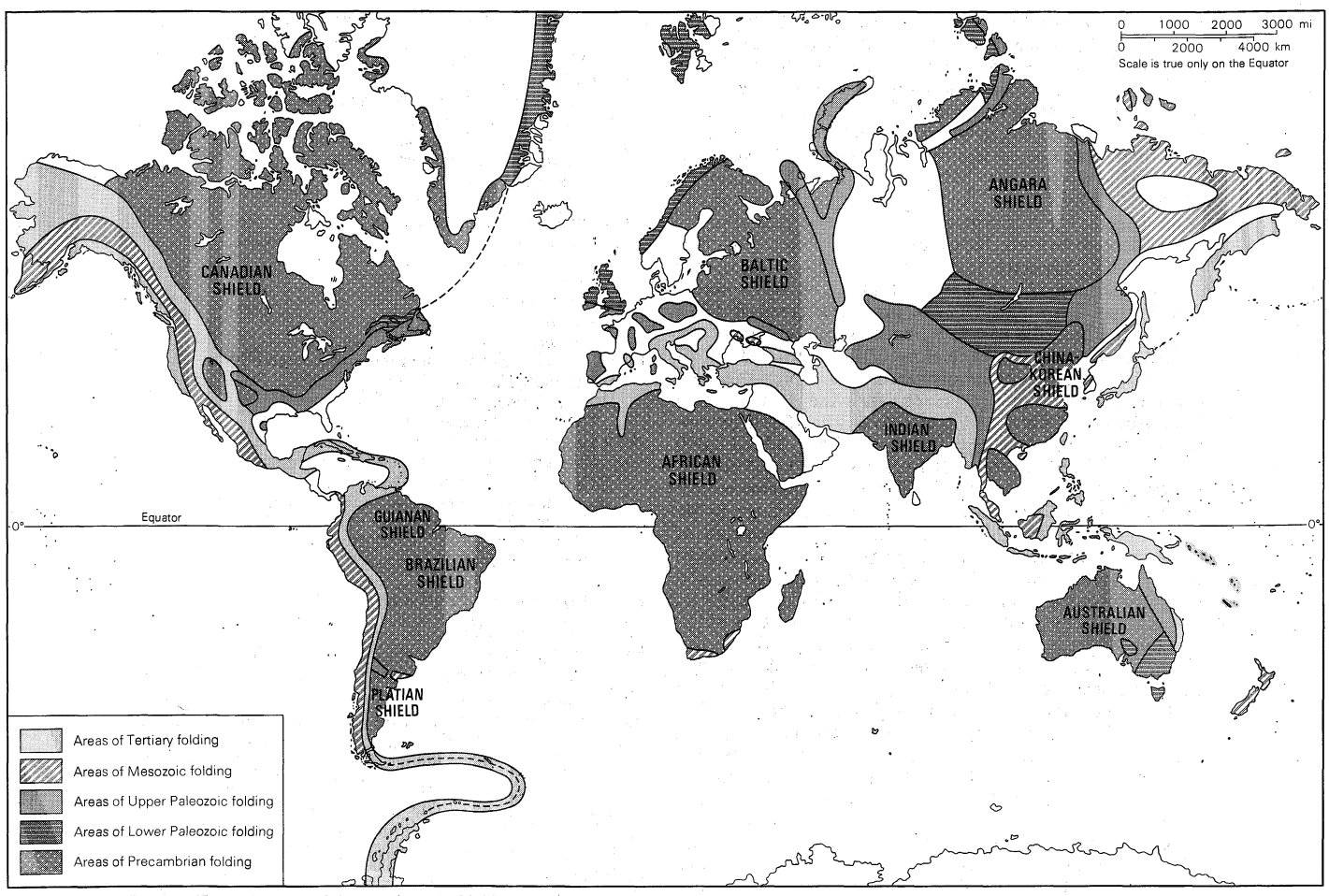


Figure 9: Tectonic units of the continents: shields and orogenic (fold) belts.
From J. Umbgrove, *The Pulse of the Earth*

It appears from experimental studies that during ultrametamorphism, when melting starts, the basic reactions to take place are of the type:

biotite + other minerals → melt + residue
hornblende + other minerals → melt + residue.

The first melts to form are partly wet granitic or granodioritic melts, and phases such as biotite and hornblende break down by producing a partly wet melt from the least refractory phases in the rocks. They would persist to much higher temperatures in systems of their own composition. The residue in the above equations is a granulite-facies metamorphic rock containing phases such as pyroxene and sillimanite or kyanite. Thus it is probable but certainly not universally proved that most granulites are formed only in the presence of a silicate liquid. This liquid may, of course, move to higher crustal levels. As with eclogites, this facies represents metamorphism in which the partial pressure of water is lower than total pressure. Granulites may be very common rocks in the base of the crust. The lowering of pressure is caused by solution and dilution of water in a silicate melt. Rocks whose appearance suggests that they were once a mixture of liquid and solid parts are often called migmatites.

The probable process of granulite formation

DISTRIBUTION OF METAMORPHIC ROCKS

A high-grade metamorphic rock is one that formed at a depth of tens of kilometres and subsequently returned to the surface. Hence, metamorphic regions are also regions of former or recent intense orogeny (mountain building). More stable regions of the Earth's crust tend to be covered with sediments, and only deep drilling will reveal the metamorphic rocks below.

The Earth's crust is made up of two basic units, the continents and ocean basins. Exploration of ocean floors has revealed that old, thick sedimentary piles are missing. Doubtless this is related to the processes of continental drift or sea-floor spreading (*qq.v.*); sediments are continuously swept up by continental motion and are added to the continents or returned to the upper mantle. Nearly all studies of metamorphic rocks have concentrated on the continents for this reason.

There are few large areas of the Earth's crust that are not affected by some type of igneous event from time to time. Although the intensity of volcanism may be focussed in certain geographic regions (*e.g.*, the Pacific margin), volcanism appears to be a rather random phenomenon, at times even occurring in the stable shield areas of the continents. In this sense, contact-metamorphic events may be found almost everywhere at almost any time on Earth. But these metamorphic events are of trivial volumetric significance compared to those of regional or burial metamorphism.

Belts of mountain building and shield areas

During the past 500,000,000 years or so of Earth history, major tectonic, seismic, igneous, and metamorphic events have been concentrated on continental margins (Figure 9). This has been a period of depression and uplift of the Earth's crust associated with the formation of the present continental distribution. The processes are still going on at dramatic rates in ocean trench environments. These modern regions of activity form immense linear belts. One such belt runs virtually around the entire Pacific margin and another through the Mediterranean and southern Asia to fuse with the circum-Pacific belt. It is in these belts that the spectacular development of zeolite facies, prehnite–pumpellyite facies, glaucophane-schist facies, and, occasionally, eclogite facies, as well as the more universal facies of regional metamorphism, have occurred. The granulite facies is almost missing.

The central and often dominant feature of most continents is their vast Precambrian-shield area; examples include the Canadian Shield, Brazilian Shield, African

Shield, and Australian Shield. In these rocks, dating reveals ages of 1,000,000,000 to 3,500,000,000 years, and they have been little affected by tectonic events postdating the Cambrian. But these shield areas are themselves complex. They consist of vast areas of granitic or granodioritic gneisses. Inside them, between them, and overlapping onto them are belts of sedimentary rocks quite like those in modern sedimentary belts of the Pacific margin or European Alps. These rocks are frequently metamorphosed in the greenschist, amphibolite, and granulite facies. Low-temperature facies and, in particular, low-temperature–high-pressure facies are missing—or have not yet been found. From marginal areas of these stable shield areas, a complex array of processes has been documented covering the past few hundred million years. The Caledonian orogeny (at the close of the Silurian Period) produced tectonic–metamorphic events along the east coast of North America, Greenland, the British Isles, Fennoscandia, Central Asia, and Australia. The Hercynian, or Variscan, orogeny followed about 300,000,000 years ago, affecting subparallel regions and the Urals and European Alps. In fact, the shield margins appear to have been subjected to a more or less constant battering by forces both destroying and rebuilding the margins of these protocontinents. As geologists study Precambrian areas in greater detail, the number of metamorphic and orogenic events recognized on a global scale increases.

It is the great task and problem of those who study metamorphic rocks to deduce the record of Earth dynamics and thermal history from metamorphic rocks. Among the questions to be answered are (1) whether the pattern of facies development through time—*e.g.*, the granulite facies in the Archean to glaucophane–lawsonite facies in the Tertiary—is a reflection of a cooling Earth and the decline of radioactivity in the crust and (2) whether the increase in size of global tectonic–metamorphic belts through time reflects changes in convective patterns in the mantle.

As understanding of the pressure–temperature regimes of metamorphism increases, and as knowledge of rock mechanics and fluid motion during metamorphism also increases through field and laboratory studies, it may become possible to understand the details of the motion of the chemical elements during such processes and hence much of the subject of economic geology or the search for man's essential raw materials.

BIBLIOGRAPHY. Fundamental texts treating all aspects of the mineralogy and petrology of metamorphic rocks include the following: T.F.W. BARTH, C.W. CORRENS, and P. ESKOLA, *Die Entstehung der Gesteine* (1939); W.A. DEER, R.A. HOWIE, and J. ZUSSMAN, *An Introduction to the Rock-Forming Minerals* (1966); W.S. FYFE, F.J. TURNER, and J. VERHOOGEN, *Metamorphic Reactions and Metamorphic Facies* (1958); A. HARKER, *Metamorphism*, 2nd ed. (1939); B. MASON, *Principles of Geochemistry*, 3rd ed. (1966); F.J. TURNER, *Metamorphic Petrology* (1968); J. VERHOOGEN *et al.*, *The Earth* (1970); H. WILLIAMS, F.J. TURNER, and C.M. GILBERT, *Petrography: An Introduction to the Study of Rocks in Thin Sections* (1954); and H.G.F. WINKLER, *Die Genese der metamorphen Gesteine* (1965; Eng. trans., *Petrogenesis of Metamorphic Rocks*, 1965). The structure and fabric of the metamorphic rocks are covered in books such as: J.G. RAMSAY, *Folding and Fracturing of Rocks* (1967); B. SANDER, *Gefügekunde der Gesteine* (1930); A. SPRY, *Metamorphic Textures* (1969); and F.J. TURNER and L.E. WEISS, *Structural Analysis of Metamorphic Tectonites* (1963).

(W.S.F.)

Metaphysics

Metaphysics is the philosophical study whose object is to determine the real nature of things—to determine the meaning, structure, and principles of whatever is insofar as it is. Although this study is popularly conceived as referring to anything excessively subtle and highly theoretical, and although it has been subjected to many criticisms, it is presented by metaphysicians as the most fundamental and most comprehensive of inquiries, inasmuch as it is concerned with reality as a whole.

This article is divided into the following sections:

I. Nature and scope of metaphysics
  Origin of the term
  Characterizations of metaphysics
  Metaphysics and other branches of philosophy
  Metaphysics and analysis
II. Problems in metaphysics
  The existence of Forms, categories, and particulars
  The existence of God
  The soul, mind, and body
  Nature and the external world
  Space and time
  The conception of spirit
III. Types of metaphysical theory
  Platonism
  Aristotelianism
  Thomism
  Cartesianism
  Idealism
  Materialism
IV. Argument, assertion, and method in metaphysics
  Metaphysics as a science
  Initial metaphysical insights
  Metaphysical arguments
V. Criticisms of metaphysics
  Metaphysics as knowledge of the supersensible
  Specific criticisms
VI. Tendencies in contemporary metaphysics
  Tendencies in the United States
  Tendencies in continental Europe
  The thought of Whitehead

I. Nature and scope of metaphysics

ORIGIN OF THE TERM

Etymologically, the term metaphysics is quite unenlightening. It means "what comes after physics"; it was the phrase used by early students of Aristotle to refer to the contents of Aristotle's treatise on what he himself called "first philosophy," and was used as the title of this treatise by Andronicus of Rhodes, one of the first of Aristotle's editors. Aristotle had distinguished two tasks for the theoretical philosopher: first, to investigate the nature and properties of what exists in the natural or sensible world, and second, to explore the characteristics of "Being as such" and to inquire into the character of "the substance that is free from movement," or the most real of all things, the intelligible reality on which everything in the world of nature was thought to be causally dependent. The first constituted "second philosophy" and was carried out primarily in the Aristotelian treatise now known as the *Physica;* the second, which Aristotle had also referred to as "theology" (because God was the unmoved mover in his system), is roughly the subject matter of his *Metaphysica.* Modern readers of Aristotle are inclined to take both the *Physica* and the *Metaphysica* as philosophical treatises; the distinction their titles suggest, between an empirical and a conceptual inquiry, has little real foundation. Aristotle was not indifferent to factual material either in natural or in metaphysical philosophy, but equally he was not concerned in either case to frame theories for empirical testing. It seems clear, nevertheless, that if the two works had to be distinguished, the *Physica* would have to be described as the more empirical, just because it deals with things that are objects of the senses, what Aristotle himself called "sensible substance"; the subject matter of the *Metaphysica,* "that which is eternal, free of movement, and separately existent," is on any account more remote. It is also evident that the connection marked in the original titles is a genuine one: the inquiries about nature carried out in the *Physica* lead on naturally to the more fundamental inquiries about Being as such that are taken up in the *Metaphysica* and indeed go along with the latter to make up a single philosophical discipline.

The background to Aristotle's divisions is to be found in the thought of Plato, with whom Aristotle had many disagreements but whose basic ideas provided a framework within which much of his own thinking was conducted. Plato, following the early Greek philosopher Parmenides, who is known as the father of metaphysics, had sought to distinguish opinion or belief from knowledge and to assign distinct objects to each. Opinion, for Plato, was a form of apprehension that was shifting and unclear, similar to seeing things in a dream or only through

their shadows; its objects were correspondingly unstable. Knowledge, by contrast, was wholly lucid; it carried its own guarantee against error, and the objects with which it was concerned were eternally what they were, and so were exempt from change and the deceptive power to appear to be what they were not. Plato called the objects of opinion phenomena or appearances; he referred to the objects of knowledge as noumena (objects of the intelligence) or quite simply as realities. Much of the burden of his philosophical message was to call men's attentions to these contrasts and to impress them with the necessity to turn away from concern with mere phenomena to the investigation of true reality. The education of the Platonic philosopher consisted precisely in effecting this transition: he was taught to recognize the contradictions involved in immediate appearances and to fix his gaze on the true realities that lay behind them, the realities that Plato himself called Forms or Ideas. Philosophy for Plato was thus a call to recognize the existence and overwhelming importance of a set of higher realities that ordinary men—even those, like the Sophists of the time, who professed to be enlightened—entirely ignored. That there were such realities, or at least that there was a serious case for thinking that there were, was a fundamental tenet in the discipline that later became known as metaphysics. Conversely, much of the subsequent controversy about the very possibility of metaphysics has turned on the acceptability of this tenet and on whether, if it is rejected, some alternative foundation can be discovered on which the metaphysician can take his stand.

Appearances and realities for Plato

CHARACTERIZATIONS OF METAPHYSICS

Before considering any such question, however, it is necessary to examine, without particular historical references, some ways in which actual metaphysicians have attempted to characterize their enterprise, noticing in each case the problems they have in drawing a clear line between their aims and those of the practitioners of the exact and empirical sciences. Four views will be briefly considered; they present metaphysics as: (1) an inquiry into what exists, or what really exists; (2) the science of reality, as opposed to appearance; (3) the study of the world as a whole; (4) a theory of first principles. Reflection on what is said under the different heads will quickly establish that they are not sharply separate from one another, and, indeed, individual metaphysical writers sometimes invoke more than one of these phrases when asked to say what metaphysics is—as, for example, the British Idealist F.H. Bradley does in the opening pages of his work *Appearance and Reality* (1893).

An inquiry into what exists. A common set of claims on behalf of metaphysics is that it is an inquiry into what exists; its business is to subject common opinion on this matter to critical scrutiny and in so doing to determine what is truly real.

Unreliability of common opinion

It can be asserted with some confidence that common opinion is certainly an unreliable guide about what exists, if indeed it can be induced to pronounce on this matter at all. Are dream objects real, in the way in which palpable realities such as chairs and trees are? Are numbers real, or should they be described as no more than abstractions? Is the height of a man a reality in the same sense in which he is a reality, or is it just an aspect of something more concrete, a mere quality that has derivative rather than substantial being and could not exist except as attributed to something else? It is easy enough to confuse the common man with questions like these and to show that any answers he gives to them tend to be ill thought out. It is equally difficult, however, for the metaphysician to come up with more satisfactory answers of his own. Many metaphysicians have relied, in this connection, on the internally related notions of substance, quality, and relation; they have argued that only what is substantial truly exists, although every substance has qualities and stands in relation to other substances. Thus, this tree is tall and deciduous and is precisely 50 yards north of that fence. Difficulties begin, however, as soon as examples like these are taken seriously. Assume for the moment that an individual tree—what might be called a concrete existent—qualifies for the title of substance; it is just the sort of thing that has qualities and stands in relations. Unless there were substances in this sense, no qualities could be real: the tallness of the tree would not exist unless the tree existed. The question can now be raised what the tree would be if it were deprived of all its qualities and stood in no relations. The notion of a substance in this type of metaphysics is that of a thing that exists by itself, apart from any attributes it may happen to possess; the difficulty with this notion is to know how to apply it. Any concrete thing one picks on to exemplify the notion of substance turns out in practice to answer a certain description; this means in effect that it cannot be spoken of apart from its attributes. It thus emerges that substances are no more primary beings than are qualities and relations; without the former one could not have the latter, but equally without the latter one could not have the former.

There are other difficulties about substance that cannot be explored now—*e.g.*, whether a fence is a substance or simply wood and metal shaped in a certain way. Enough has already been said, however, to indicate the problems involved in defining the tasks of metaphysics along these lines. Nevertheless, there is an alternative way of understanding the notion of substance: not as that which is the ultimate subject of predicates but as what persists through change. The question "What is ultimately real?" is, thus, a question about the ultimate stuff of which the universe is made up. Although this second conception of substance is both clearer and more readily applicable than its predecessor, the difficulty about it from the metaphysician's point of view is that it sets him in direct rivalry with the scientist. When the early Greek philosopher Thales inquired as to what is ultimately real and came up with the surprising news that all is water, he might be taken as advancing a scientific rather than a philosophical hypothesis. Although it is true that later writers, such as Gottfried Wilhelm Leibniz, a German Rationalist philosopher and mathematician, were fully aware of the force of scientific claims in this area and nevertheless rejected them as metaphysically unacceptable, the fact remains that the nonphilosopher finds it difficult to understand the basis on which a Leibniz rests his case. When Leibniz said that it is monads (*i.e.*, elementary, unextended, indivisible, spiritual substances that enter into composites) that are the true atoms of nature and not, for example, material particles, the objection can be raised as to what right he has to advance this opinion. Has he done any scientific work to justify him in setting scientific results aside with such confidence? And if he has not, why should he be taken seriously at all?

The ultimate stuff of the universe

The science of ultimate reality. To answer these questions, another description of metaphysics has been proposed: that it is the science that seeks to define what is ultimately real, as opposed to what is merely apparent.

The contrast between appearance and reality, however, is by no means peculiar to metaphysics. In everyday life, people distinguish between the real size of the sun and its apparent size, or again between the real colour of an object (when seen in standard conditions) and its apparent colour (nonstandard conditions). A cloud appears to consist of some white, fleecy substance, although in reality it is a concentration of drops of water. In general, men are often (though not invariably) inclined to allow that the scientist knows the real constitution of things as opposed to the surface aspects with which ordinary men are familiar. It will not suffice, therefore, to define metaphysics as knowledge of reality as opposed to appearance; scientists, too, claim to know reality as opposed to appearance, and there is a general tendency to concede their claim. It seems that there are at least three components in the metaphysical conception of reality. One characteristic, which has already been illustrated in Plato, is that reality is genuine as opposed to deceptive. The ultimate realities that the metaphysician seeks to know are precisely things as they are—simple and not variegated, exempt from change, and therefore stable objects of knowledge. Plato's own assumption of this position perhaps reflects certain confusions about the knowability of things that

Components in the metaphysical conception of reality

change; one should not, however, on that ground exclude this aspect of the concept from metaphysical thought in general. Ultimate reality, whatever else it is, is genuine as opposed to sham. Second, it is original in contrast to derivative, self-dependent rather than dependent on the existence of something else. When Aristotle sought to inquire into the most real of all things, or when medieval philosophers attempted to establish the characteristics of what they called the *ens realissimum* ("the most real being"), or the original and perfect being, they were looking for something that, in contrast to the everyday things of this world, was truly self-contained and could accordingly be looked upon as self-caused. Likewise, the 17th-century Rationalists defined substance as that which can be explained through itself alone. Writers like René Descartes and Benedict de Spinoza were convinced that it was the task of the metaphysician to seek for and characterize substance understood in this sense; the more mundane substances with which physical scientists were concerned were, in their opinion, only marginally relevant in this inquiry. Third, and perhaps most important, reality for the metaphysician is intelligible as opposed to opaque. Appearances are not only deceptive and derivative, they also make no sense when taken at their own level. To arrive at what is ultimately real is to produce an account of the facts that does them full justice. The assumption is, of course, that one cannot explain things satisfactorily if he remains within the world of common sense, or even if he advances from that world to embrace the concepts of science. One or the other of these levels of explanation may suffice to produce a sort of local sense that is enough for practical purposes or that forms an adequate basis on which to make predictions. Practical reliability of this kind, however, is very different from theoretical satisfaction; the task of the metaphysician is to challenge all assumptions and finally arrive at an account of the nature of things that is fully coherent and fully thought out.

The task of the metaphysician

It should be obvious that, to establish his right to pronounce on what is ultimately real in the sense analyzed, the metaphysican has a tremendous amount to do. He must begin by giving colour to his claim that everyday ways of thinking will not suffice for a full and coherent description of what falls within experience, thus arguing that appearances are unreal—although not therefore nonexistent—because they are unstable and unintelligible. This involves a challenge to the final acceptability of such well-worn ideas as time and space, thing and attribute, change and process—a challenge that metaphysicians have not hesitated to make, even though it has been treated with skepticism both by ordinary men and by some of their fellow philosophers (*e.g.*, G.E. Moore, a 20th-century British thinker who has greatly influenced modern Analytic philosophy). Second, granted that there are contradictions or incoherencies in the thought of common sense, the metaphysician must go on to maintain that they cannot be resolved by deserting common sense for science. He will not deny that the concepts of science are in many respects different from those of everyday thought; to take one aspect only, they are altogether more precise and sharply defined. They permit the scientist to introduce into his descriptions a theoretical content that is lacking at the everyday level, and in so doing to unify and render intelligible aspects of the world that seem opaque when considered singly. The metaphysician will argue, however, that this desirable result is purchased at a certain price: by ignoring certain appearances altogether. The scientist, in this way of thinking, does not so much offer a truer description of the phenomena of which ordinary thought could make no sense; he merely gives a connected description of a selected set of phenomena. The world of the scientist, restricted as it is to what can be dealt with in quantitative terms, is a poor thing in comparison with the rich if untidy world of everyday life. Alternatively, the metaphysician must try to show that scientific concepts are like the concepts of common sense in being ultimately incoherent. The premises or presuppositions that the scientist accepts contain unclarities that cannot be resolved, although they are not so serious as to prevent his achieving results that are practically dependable. Many ingenious arguments on these lines have been produced by philosophers, by no means all of whom could be said to be incapable of a true understanding of the theories they were criticizing. (Leibniz, for example, was a physicist of distinction as well as a mathematician of genius; G.W.F. Hegel, a 19th-century German Idealist, had an unusual knowledge of contemporary scientific work; and Alfred North Whitehead, a pioneer of 20th-century metaphysics in the Anglo-Saxon world, was a professor of applied mathematics, and his system developed from physics and contained a wealth of biological ideas.) The fact remains, nevertheless, that few if any practicing scientists have been seriously troubled by such arguments.

Even if the metaphysician were thus able to make good the negative side of his case, he would still face the formidable difficulty of establishing that there is something answering to his conception of what is ultimately real and of identifying it. The notion of an original being, totally self-contained and totally self-intelligible, may not itself be coherent, as the 18th-century British philosopher David Hume and others have argued; alternatively, there may be special difficulties in saying to what it applies. The fact that different metaphysicians have had widely different accounts of what is ultimately real is certainly suspicious. Some have wanted to say that there is a plurality of ultimately real things, others that there is only one; some have argued that what is truly real must be utterly transcendent of the things of this world and occupy a supersensible realm accessible only to the pure intellect, while others have thought of ultimate reality as immanent in experience (the Hegelian Absolute, for example, is not a special sort of existent, but the world as a whole understood in a certain way). That metaphysical inquiry should issue in definitive doctrine, as so many of those who engaged in it said that it would, is in these circumstances altogether too much to hope for.

The science of the world as a whole. Another way in which metaphysicians have sought to define their discipline is by saying that it has to do with the world as a whole.

The implications of this phrase are not immediately obvious. Clearly, a contrast is intended in the first place with the various departmental sciences, each of which selects a portion or aspect of reality for study and confines itself to that. No geologist or mathematician would claim that his study is absolutely comprehensive; each would concede that there are many aspects of the world that he leaves out, even though he covers everything that is relevant to his special point of view. By contrast, it might be supposed that the metaphysician is merely to coordinate the results of the special sciences. There is clearly a need for the coordination of scientific results in recent conditions in which scientific research has become increasingly specialized and departmentalized; individual scientific workers need to be made aware of what is going on in other fields, sometimes because these fields impinge on their own, sometimes because results obtained there have wider implications of which they need to take account. One can scarcely see metaphysicians, however, or indeed philosophers generally, performing this function of intellectual contact man in a satisfactory fashion. It might then be supposed that their concern with the world as a whole is to be interpreted as a summing up and synthesizing of the results of the particular sciences. Plato spoke of the philosopher as taking a "synoptic" view, and there is often talk about the need to see things in the round and avoid the narrowness of the average specialist, who, it is said, knows more and more about less and less. If, however, it is a question of looking at scientific results from a wider point of view and so of producing what might be called a scientific picture of the world, the person best qualified for the job is not any philosopher but rather a scientist of large mind and wide interests. Metaphysics cannot be satisfactorily understood as an account of the world as a whole if that description suggests that the metaphysician is a sort of superscientist, unlimited in his curiosity and gifted with a capacity for putting together other people's findings with a skill and imagina-

tion that none of them individually commands. Only a scientist could hope to become such a superscientist.

Metaphysics as an account of the world

More hope for the metaphysician can be found, perhaps, along the following lines. People want to know not only what the scientist makes of the world but also what significance to assign to his account. People experience the world at different levels and in different capacities: they are not only investigators but also agents; they have a moral and a legal, an aesthetic and a religious life on top of their scientific life. Man is a many-sided being; he needs to understand the universe in the light of his different activities and experiences. There are philosophers who appear to find no mystery here; they argue that there can be no question of the possibility of, say, a moral or a religious vision of the world that rivals the scientific vision. In this view, morals and religion are matters of practice, not of theory; they do not rival science, but only complement it. This neutralist attitude, however, finds little general favour; for most thinking people there is a question of choosing whether to go all the way with science, at the cost of abandoning religion and even morals, or to stick to a religious or moral world outlook even if it means treating scientific claims with some reserve. The practice of the moral life is often believed to proceed on assumptions that can hardly be accepted if science is taken to have the last word about what is true. Accordingly, it becomes necessary to produce some rational assessment of the truth claims of the different forms of experience, to try to think out a scheme in which justice is done to them all. Many familiar systems of metaphysics profess to do just that; among others there are Materialism, which favours the claims of science, Idealism, which sees deeper truth in religion and the moral life, and the peculiar dualism of the 18th-century German philosopher Immanuel Kant, which holds that science gives the truth about phenomena, while reserving a noumenal, or supersensible, sphere for moral agency.

This conception of metaphysics as offering an account of the world or, as is more often said, of experience as a whole accords more obviously with the practice of those who see ultimate reality as immanent, or inherent in what is immediately known, than of those who take it to be transcendent, or beyond the limits of ordinary experience. It is possible, in fact, to subscribe to the legitimacy of metaphysics as so understood without postulating the existence of any special entities known only to the metaphysician—a claim that plain men have often taken to connect metaphysics with the occult. This is not to say, of course, that metaphysical problems admit of easy solutions when understood along these lines. There is a variety of widely different ways of taking the world as a whole: each depends on which aspect or aspects of experience the individual metaphysician finds especially significant; each claims to be comprehensive and to confute the claims of its rivals, yet none has succeeded in establishing itself as the obviously correct account. Even systems that are widely condemned as impossible, such as Materialism, turn out in practice to command constantly renewed support, as new discoveries in the sciences suggest new ways of dealing with old difficulties. A cynic might take such facts as meaning that people subscribe to theories of this sort more as a matter of emotional than of rational conviction; metaphysics, as Bradley remarked with surprising frankness, consists in the finding of bad reasons for what one believes upon instinct.

The science of first principles. Another phrase used by Bradley in his preliminary discussion of metaphysics is "the study of first principles," or ultimate, irrefutable truths.

First principles as basic truths

Metaphysics could be said to provide a theory of first principles if it furnished men with a set of concepts in the light of which they could arrive at the connected account of experience as a whole just spoken of, and the two descriptions of the subject would thus be two sides of a single coin. The idea that metaphysics has to do with first principles, however, has wider implications.

The term first principles is a translation of the Greek word *archai*. An *archē* is something from which an argument proceeds—either a primary premise or an ultimate presupposition. Plato, in a famous passage in *The Republic*, contrasted two different attitudes to *archai:* that of the mathematician, who lays down or hypothesizes certain things as being true and then proceeds to deduce their consequences without examining them further; and that of the dialectician, who proceeds backward, not forward, from his primary premises and seeks to ground them in an *archē* that is not hypothesized at all. Unfortunately, no concrete details exist of the way in which Plato himself thought this program could be carried out; he spoke of it only in the most general terms. Nevertheless, the suggestion that metaphysics is superior to any other intellectual discipline in having a fully critical attitude toward its first principles is one that continues to be made, and it needs some examination.

As regards mathematics, for example, it might be said that mathematicians could be uncritical about the first principles of their science in the following ways: (1) They might take as self-evidently true or universally applicable some axiom or primary premise that turned out later not to possess this property. (2) They might assume among their first principles certain propositions about existence—to the effect that only certain kinds of things could be proper objects of mathematical inquiry (rational as opposed to irrational numbers, for example)—and time might show that the assumption was inappropriate. The remedy for both sorts of error, however, is to be found within mathematics itself; the development of the discipline has consisted precisely in eliminating mistakes of this kind. It is not clear even that the discovery and removal of antinomies in the foundations of mathematics is work for the metaphysician, although philosophically minded persons like Gottlob Frege, a German mathematician and logician, and Bertrand Russell, perhaps the best known English philosopher of the 20th century, have been much concerned with them. The situation is not fundamentally different when the empirical sciences are considered. Admittedly, the exponents of these sciences give more hostages to fortune insofar as they have to assume from the first the general correctness of the results of other disciplines; there can be no question of their checking on these for themselves. Mathematicians, too, assume the validity of common argument forms without making any serious attempt to validate them, and there is nothing seriously wrong with their doing so. If confidence in bad logic has sometimes been responsible for holding up mathematical advance, bolder mathematicians have always known in practice that the right thing to do is let the argument take them where it will on strictly mathematical lines, leaving it to logicians to recognize the fact and adjust their theory at leisure.

It thus seems that the assertion that a special science like mathematics is uncritical about its *archai* is false; there is a sense in which mathematicians are constantly strengthening their basic premises. As regards the corresponding claim about metaphysics, it has at one time or another been widely believed (1) that it is the business of metaphysics to justify the ultimate assumptions of the sciences, and (2) that in metaphysics alone there are no unjustified assumptions. Concerning (1), the question that needs to be asked is how the justification is supposed to take place. It has been argued that the metaphysician might, on one interpretation of his function, be said to offer some defense of science generally by placing it in relation to other forms of experience. To do this, however, is not to justify any particular scientific assumptions. In point of fact, particular scientific assumptions get their justification, if anywhere, when a move is made from a narrower to a more comprehensive science; what is assumed in, for example, geology may be proved in physics. But this, of course, has nothing to do with metaphysics. The difficulty with (2) is that of knowing how any intellectual activity, however carefully conducted, could be free of basic assumptions. Some metaphysicians (such as Bradley and his Scottish predecessor J.F. Ferrier) have claimed that there is a difference between their discipline and others insofar as metaphysical propositions alone are self-reinstating. For example, the Cartesian proposition *Cogito,*

Metaphysicians as self-critical

ergo sum ("I think, therefore I am") is self-reinstating: deny that you think, and in so doing you think; deny that you exist, and the very fact gives proof of your existence. Even if it could be made out that propositions of this kind are peculiar to metaphysics, however, it would not follow that everything in metaphysics has this character. The truth is, rather, that no paradox is involved in denying most fundamental metaphysical claims, such as the assertion of the Materialist that there is nothing that cannot be satisfactorily explained in material terms, or the corresponding principle of Aristotle that there is nothing that does not serve some purpose.

The view that metaphysics, or indeed philosophy generally, is uniquely self-critical is among the myths of modern thought. Philosophers rely on the results of other disciplines just as other people do; they do not pause to demonstrate the legitimacy of the principles of simple arithmetic before entering on calculations in the course of their work, nor do they refrain from employing the *reductio ad absurdum* type of refutation (*i.e.*, showing an absurdity to which a proposition leads when carried to its logical conclusion) until they have assured themselves that this is a valid way of confuting an opponent. Even in their own field they tend, like painters, to work within traditions set by great masters rather than to think everything out from scratch for themselves. That philosophy in practice is not the fully self-critical activity its exponents claim it to be is shown nowhere more clearly than in the reception that philosophers give to theories that are unfashionable; they more often subject them to conventional abuse than to patient critical examination. Nevertheless, it is from the conviction that philosophy, and especially metaphysical philosophy, operates without unjustified assumptions that current claims about the superiority of this branch of thinking derive their force. This conviction connects with the views already mentioned, that metaphysics is the science of first principles, and that the principles in question are ineluctable, in the sense that they are involved in their own denial.

METAPHYSICS AND OTHER BRANCHES OF PHILOSOPHY

The view of metaphysics as the queen of the sciences

It may be useful at this point to consider the question of the relations of metaphysics to other parts of philosophy. A strong tradition, derided by Kant, asserted that metaphysics was the queen of the sciences, including the philosophical sciences. The idea presumably was that those who worked within fields such as logic and ethics, as well as physicists and biologists, proceeded on assumptions that in the last resort had to be approved or corrected by the metaphysician. Logic could be conceived as a special study complete in itself only if the logician were allowed to postulate a correspondence between the neat and tidy world of propositions, which was the immediate object of his study, and the world existing in fact; metaphysics might and sometimes did challenge the propriety of this postulate. Similarly, ethics, like law, could get nowhere without the assumption that the individual agent is a self-contained unit answerable in general terms for what he does; metaphysics had the duty of subjecting this assumption to critical examination. As a result of such claims it was widely believed that any results obtained by logicians or ethicists must at best be treated as provisional; followers of Hegel, who advanced these claims with passionate conviction, were inclined in consequence to regard logic and ethics alike as minor branches of philosophy. It has been a feature of 20th-century philosophical thought, especially in Britain and the United States, to dispute these Hegelian contentions and argue for the autonomy of ethics and logic; that is, for their independence of metaphysics. Thus, formal logicians of the school of Frege and Russell were apt to claim that the principles of logic applied unequivocally to all thinking whatsoever; there could be no question of their having to await confirmation, still less correction, from the metaphysician. If metaphysical arguments suggested that fundamental laws of logic such as the principle of noncontradiction—that a statement and its contradictory cannot both be true—might not be in order, the only conclusion to draw was that such arguments must be confused: without observation of the laws of logic there could be no coherent thinking of any sort.

Similarly, G.E. Moore, in a celebrated section of his *Principia Ethica* (1903), tried to show that statements like "This is good" are *sui generis* and cannot be reduced to statements of either natural or metaphysical fact; the Idealist belief that ethics ultimately depends on metaphysics rested on a delusion. Moore perhaps failed to see the force of the Idealist challenge to the individualist assumptions on which much ethical thinking proceeds, and he did not note that, in one respect at least, ethical results can be dependent on those of metaphysics: if metaphysics shows that the world is other than it is initially taken to be, conclusions about what to do must be altered accordingly. Again, the reaction among logicians to Hegelian attempts to merge logic into metaphysics certainly went too far. There is a genuine philosophical problem about the relation between the world of logic and the world of fact, and it cannot be solved by simply repeating that logic is an autonomous discipline whose principles deserve respect in themselves. None of this, however, shows that metaphysics is the fundamental philosophical discipline, the branch of philosophy that has the last word about what goes on in all other parts of the subject.

METAPHYSICS AND ANALYSIS

Modern British and American philosophers commonly describe themselves as engaged in philosophical analysis, as opposed to metaphysics. The interests of a metaphysician, according to this view, are predominantly speculative; he wants to reveal hitherto unknown facts about the world and on that basis to construct a theory about the world as a whole. In so doing he is necessarily engaged in activities that rival those of the scientist, with the important difference that scientific theories can be brought to the test of experience, whereas metaphysical theories cannot. Eschewing this conception of philosophy as impossible, the critic of metaphysics believes that philosophy should confine itself to the analysis of concepts, which is a strictly second-order activity independent of science and which need involve no metaphysical commitment.

Analysis in philosophy

The notion of analysis in philosophy is far from clear. Analysis on any account is meant to result in clarification, but it is not evident how this result is to be achieved. For some, analysis involves the substitution for the concept under examination of some other concept that is recognizably like it (as Gilbert Ryle, an English Analyst, elucidated the concept of mind by replacing it with the notion of "a person behaving"); for others, analysis involves the substitution of synonym for synonym. If the latter understanding of analysis is required, as in Moore's classic example of the analysis of brother as male sibling, not much enlightenment is likely to ensue. If, however, the philosopher is permitted to engage in what is sometimes pejoratively described as "reductive analysis," he will produce interest at the cost of reintroducing speculation. Ryle's *Concept of Mind* (1949) is a challenging book just because it advances a thesis of real metaphysical importance—that one can say everything one needs to say about minds without postulating mental substance.

A further aspect of the situation that deserves mention is this. If it is the case, as is often claimed, that analysis can be practiced properly only when the analyst has no metaphysical presuppositions, by what means does he select concepts for analysis? Would it not be appropriate for him, in these circumstances, to take any concept of reasonable generality as a suitable subject on which to practice his art? It turns out, in fact, however, that the range of concepts commonly recognized as philosophical is more limited than that, and that those concepts to which Analytic philosophers give their attention are chosen because of their wider philosophical bearings. Thus, recent philosophers have paid particular attention to the concept of knowledge not just because it is a notion whose analysis has long proved difficult but also because on one account at least it involves an immediately experienced mental act—something that many Analysts would like to proscribe as mythical. Similarly, the celebrated analysis of the idea of causality put forward by David Hume was

not undertaken out of idle curiosity but with a wider purpose in mind: to undermine both the Aristotelian and the Cartesian views of the world and to substitute for them an atomism of immediate appearances in which all objects were "loose and separate"—that is, logically independent one of another. The insight into the constitution of nature promised in different ways by Aristotle and Descartes was an illusion, the truth being that scientific advance serves only to "stave off our ignorance a little." What Hume said about causation connects internally with his views about what exists. Despite his polemic against books of "divinity and school metaphysics," he had a metaphysics of his own to recommend.

Historical use of analysis by metaphysicians

The truth is that metaphysics and analysis are not separate in the way modern Analytic philosophers pretend. The speculative philosophers of the past were certainly not averse to analysis: witness the splendid discussion of the concept of knowledge in Plato's *Theaetetus*, or, for a more recent example, Bradley's account of the meanings of "self." The legend that a metaphysical philosopher has his eye so firmly set on higher things that he is entirely careless of the conceptual structure he seeks to recommend is absolutely without foundation. A metaphysical philosopher is a philosopher after all: argument and the passion for clarification are in his blood. Although some contemporary philosophers profess to undertake analysis entirely for its own sake and without explicit metaphysical motivation, it may be doubted if their claim is capable of being sustained. The "logical analysis" practiced by Russell in the early part of the 20th century was not metaphysically neutral, nor was the analysis of the Logical Positivists, who recommended a strongly scientific view of the world. Some current analytic work is motivated less by the desire to forward an overall theory than by a wish to destroy a prevailing or previously held theory that is considered objectionable. To seek to overthrow a metaphysical theory, however, is itself to engage in metaphysics—not very interesting metaphysics, perhaps, but metaphysics all the same.

It may be added, as a historical note, that the Rationalist philosophers of the 17th and 18th centuries, who emphasized the predominant role of reason in the construction of a system of knowledge, believed that the philosopher's task fell into two parts. He must first break down complex concepts into their simple parts; this was a matter of analysis. Then he must proceed to show how knowledge of these simples would serve to explain the detailed constitution of things; this would involve synthesis. That there are deep obscurities in this program—*e.g.*, whether it is a matter of analyzing concepts or getting down to the simplest elements of things—is less important in the present context than that analysis and synthesis were thus taken to be complementary. The classical statement of this point of view is to be found in Descartes's *Discours de la méthode* (1637; *Discourse on Method*, 1950), with the corresponding passages in the *Regulae ad Directionem Ingenii* (published posthumously 1701; *Rules for the Direction of the Mind*, 1961). That the idea persisted well into the 18th century is evidenced by the remarks made by Kant in his essay *Untersuchung über die Deutlichkeit der Grundsätze der natürlichen Theologie und der Moral* (1764; *Inquiry into the Distinctness of the Principles of Natural Theology and Morals*, 1949), in which he said that metaphysics was not yet in a position to pass beyond the stage of analysis to that of synthesis. He did not mean that for the time being philosophy must remain entirely nonmetaphysical, in the way some moderns suppose it can, but rather that it needs to go on elaborating a conceptual scheme, which, however, cannot be used constructively until it is complete. Actually, Kant belied his own professions at the time, insofar as he thought himself in possession of a definitive proof of God's existence, which he explained in his essay *Der einzig mögliche Beweisgrund zu einer Demonstration des Daseyns Gottes* (1763; "The Only Possible Ground for a Demonstration of the Existence of God"). This, however, only illustrates the not very surprising fact that philosophers are often less clear about the nature of their own activities than they think.

II. Problems in metaphysics

To give a comprehensive account of the main problems of metaphysics in the space of a few pages is clearly quite impossible. What follows is necessarily highly selective and to that extent misleading; nevertheless, it attempts to offer an introduction to metaphysical thinking itself, rather than reflection on the history of metaphysics.

THE EXISTENCE OF FORMS, CATEGORIES, AND PARTICULARS

Forms. The early Greek philosophers asked the question *ti to on*, "What is existent?" or "What is really there?" They originally interpreted this as a question about the stuff out of which things were ultimately made, but a new twist was given to the inquiry when Pythagoras, in the late 6th century BC, arrived at the answer that what was really there was number. Pythagoras conceived what is there in terms not of matter but of intelligible structure; it was the latter that gave each type of thing its distinctive character and made it what it was. The idea that structure could be understood in numerical terms was probably suggested to Pythagoras by his discovery that there are exact correlations between the lengths of the strings of a lyre and the notes they produce. By a bold extrapolation he seems to have surmised that what held in this case must hold in all cases.

Platonic theory

The Pythagorean theory that what is really there is number is the direct ancestor of the Platonic theory that what is really there is Forms or Ideas (*eidē*, or *ideai*). Plato's Forms were also intelligible structures and not material elements, but they differed from Pythagorean numbers by being conceived of as separately existent. There was, as Plato put it, a "place accessible to the intelligence," which was the place or realm of Forms. Each Form was a genuine existent, in the sense of being precisely what it pretended to be; the Form of Beauty, for example, was beautiful through and through. By contrast, the many particular things that partook of or resembled what was truly beautiful were one and all defective. However beautiful any one of them might be, it was also in another respect lacking in beauty. It turned out to possess contradictory characteristics, and as such could never be identified with true reality.

Plato had taken over from his predecessor Heracleitus, who flourished at about the beginning of the 5th century BC, the doctrine that the world of sensible things is a world of things in constant flux; as he put it in the *Theaetetus*, nothing *is* in this world because everything is in a state of becoming something else. Forms were needed to provide stable objects for knowledge as well as to answer the question of what is ultimately real. Although Plato played down the reality of sensible things, making them mere objects of opinion and describing them as falling between what is and what is not, he did not deny their existence. It was not his thesis that Forms alone exist. On the contrary, he appears to have held that God (who was certainly not a Form) had somehow fashioned the physical world on the model of the Forms, using space as his material. This is the description that is given in the *Timaeus*, in a passage that Plato perhaps meant his readers not to take quite literally but that stated his view as plainly as he thought it could be stated. In this passage God appears in the guise of the "Demiurge," although he is referred to freely in other Platonic dialogues. Souls were also distinct from Forms in Plato.

In the discussions that developed around the theory of Forms, many difficulties were revealed, most of them familiar to Plato himself. The question of how the one Form was supposed to relate to the many particulars that participated in or resembled it was nowhere satisfactorily answered. The difficulty turned on how the Form was to be thought of at once as an existent and as a structure. Plato seemed on occasion to think of it as a structure hypostatized, or given real existence. This thesis led to the antinomies exposed in the "third man" argument. According to this theory, particular men were alleged to be human because of their relationship to "Man himself"; *i.e.*, the Form of man. But whence did the latter derive its nature? Must there not be a second Form, to explain what the first Form and its particulars have in common,

and will not this generate an infinite regress? Again, the problem of the precise population of the world of Forms never got a definitive solution, perhaps because the theory of Forms was put to more than one purpose. Sometimes it was said that there is a Form answering to every general word, but elsewhere the theory was that what is merely negative (*e.g.*, being lifeless) has no need of a special Form, nor does what is manufactured. There is even a question as to whether trivial everyday things such as mud and hair and dirt have Forms, though it is agreed that there is a Form of man.

The problems just referred to were stated trenchantly in Plato's dialogue the *Parmenides*; the discussion there ends with the statement that the Forms must be retained if an account of intelligible discourse is to be given, but no indication is offered as to how the theory is to be refurbished. Some Platonic scholars have inferred that Plato virtually gave it up, but such evidence as there is suggests that he only transformed it into a theory of Form-numbers, more openly Pythagorean than the earlier version. There are many references in Aristotle to this theory of Form-numbers, but no writing of Plato's own on the subject has survived, and it is virtually impossible at this late stage to say what this theory really comprised.

One further feature of the theory of Forms must be mentioned here: the view that there is a supremely important Form, the Form of goodness, or of the Good, which somehow determines the contents of the world of Forms and brings order into it. In a celebrated but brief and tantalizing passage in *The Republic*, the Form of the Good is spoken of as being to the intelligible realm what the sun is to the visible realm; just as the sun makes living things grow and renders them visible, so the Good is responsible for the existence and intelligibility of Forms, though it is itself "on the other side of Being." This passage had a tremendous historical influence on the Neoplatonists, who saw it as anticipating the ultimate ineffable reality—the One, from which everything describable was in some way an emanation—in which they came to believe. It seems possible, however, that Plato had no such mystical thoughts in mind, but simply wanted to say that the world of Forms is ordered through and through, everything in it being there for a purpose. The Form of Good is, in fact, the counterpart of the *Nous* (Mind) of Anaxagoras, another of Plato's predecessors, which was supposed to arrange everything for the best.

Aristotelian doctrine

Categories and universals. The most famous critic of Plato's theory of Forms was Aristotle, who devised his doctrine of categories largely to counter it. According to this doctrine, "being is spoken of in many ways": one can say that there are such things as individual horses, but one can also say that there is such a thing as being a horse, or as being upside down. Expressions can be classified under various heads: predicates signify substances (*e.g.*, "man" or "horse"), qualities (*e.g.*, "white"), relations (*e.g.*, "greater"), quantities (*e.g.*, "three yards long"), time (*e.g.*, "last year"), and so on—sometimes Aristotle listed ten categories, sometimes only eight. The kind of being that any predicate possesses, however, is derivative in comparison with the being of an individual substance, a particular man or a particular horse. It is such things that exist in the primary sense, and it is upon their existence that the existence of other types of being depends. Or, to put the point in not quite Aristotelian terms, primary substances are the only concrete existents; Socrates, the bearer of a proper name, exists in a way in which humanity or whiteness or being greater do not. The latter are really no more than abstractions, and nothing but confusion can arise from neglecting that fact.

Mention has already been made of the difficulties into which this doctrine led when it came to describing primary substances; it appeared that these entities could not be characterized but only named or pointed to, a conclusion accepted much later by Ludwig Wittgenstein, a 20th-century philosopher, in his *Tractatus Logico-Philosophicus* and by Russell in his lectures on logical atomism. These difficulties, however, were not seen at the time the theory was promulgated, and it is more important now to emphasize the fact that it undermined any doctrine of the Platonic type. To argue that Forms, or numbers, alone are real is to argue for the reality of abstractions; to put the point succinctly, beauty exists only so long as something is beautiful, and that something must be a concrete individual. Or if this is not quite true (for, after all, it could be said that there is such a thing as having a million sides even if nothing in fact has a million sides), concrete existence must precede abstract existence in some cases at least: the "*X*" in "*X* is red" must sometimes be replaceable by an actual rather than a merely possible entity.

The problem of universals

A prominent subject of philosophical discussion in the Middle Ages was what came to be known as the problem of universals, which concerned the ontological status, or type of existence, to be assigned to the referents of general words. One of Plato's critics had said, "I see particular horses, but not horseness"; and Plato had answered, "That is because you have eyes but no intelligence." There can be no doubt that Plato thought that horseness, the Form of horse or Horse itself, to use his own expression, was something that existed separately; it could be discerned not by the bodily eyes but by the eye of the soul. The view that besides individual horses there also exists the Form of horse was known in the Middle Ages as Realism. Aristotle was also alleged to be a Realist, because he too thought that Forms were really there, although only as embodied in particular instances. More skeptical philosophers denied the reality of universals altogether, some identifying them with thoughts (conceptualists), others with mere names (nominalists).

The dispute about universals was in fact very confused. At least two quite separate issues were involved. First of all, there was the question about the status to be assigned to whatever it was that predicates referred to; this question seemed urgent just because, for example, geometricians were able to discuss the properties of the triangle or the circle. What and where were the triangle and the circle? In fact, the Aristotelian doctrine of categories had already indicated that the being of any predicate was necessarily different from that of primary substances; the circle did not and could not exist as this man or this horse did. When Aristotle is described as a Realist in the dispute about universals, the description is very misleading. In one sense he did not believe that universals are real at all; in another sense, however, he did, and this is where the second issue arose. Some people who denied the reality of universals wanted to say that all classification is artificial; the descriptions men give of things depend upon their interests as much as upon what is really there. Aristotle, by contrast, believed in a doctrine of natural kinds; he thought that every particular horse embodied the form or objective essence of horse, which was accordingly a genuine, if abstract, constituent of the world. The question of the extent to which classification is artificial is clearly quite different from that of the status of universals; it remains to be solved even if the latter problem is dismissed, as it is by modern philosophers who say that only proper names and individuating phrases have referents; general words do not. These differences, however, were not clearly seen either in the Middle Ages or during the 17th century, when the whole question was discussed at length by philosophers like Thomas Hobbes and John Locke. (For the problem of natural versus artificial classification, see CLASSIFICATION THEORY.)

Basic particulars. In discussions of the problem of universals it was often claimed, especially by nominalists, that only particulars exist. The notion of a particular is in many respects unclear. Strictly, the terms particular and universal are correlatives; a particular is an instance of universal (*e.g.*, this pain, that noise). It would seem from this that particulars and individuals should be the same, but there are writers who distinguish them. Bradley, in his *Principles of Logic* (1883), treated particulars as mere momentary instantiations of universals and contrasted them with individuals as continuants possessing internal diversity. An individual cannot merely be identified but re-identified; because it lasts through time,

it may possess incompatible attributes at different periods of its history. A particular, however, is nothing but an instantiation of an attribute and as such must possess that attribute if it is to be anything. Similarly, a particular can be met with once, but not again; as time moves on, it passes out of existence and is replaced by another particular that may resemble it but is not literally identical with it.

If particulars and individuals are thus distinguished, it is by no means clear that only particulars exist, or indeed that they exist at all; it could be that they are no more than abstract aspects of genuinely concrete entities such as persons or material things. But there are arguments on the other side, advanced in a variety of forms by David Hume and Bertrand Russell. Hume believed that the ultimate constituents of the world were either impressions or their fainter copies, ideas; both were species of perceptions. Impressions he defined as "internal and perishing existences"; they were of various kinds, embracing feelings as well as such things as experienced colours and smells, but all were at best extremely short-lived. Impressions arose in human consciousness from unknown causes; their existence could not, however, be denied. By contrast, the existence of continuing and independent material objects and of continuing minds was extremely precarious; analysis showed both to be no more than bundles of perceptions, united by certain relations, and Hume more than once referred to them as "fictions," although it turned out on examination that they were not fictions in the way ghosts are. Hume's reasons for advancing these views were primarily epistemological; he thought that statements about continuants were all open to doubt, although statements about the contents of immediate experience could not be challenged. When it was a question of what really existed, the only sure answer was items in consciousness—namely, impressions and ideas.

Hume's theory of particulars

Russell, who was generally sympathetic to this answer, added another argument derived from logic: proper names, he said, were names of particulars, which must accordingly exist. Ordinary proper names (such as "Socrates") had other functions than to denote, but logically proper names ("this" was Russell's example) served simply to pick out objects of immediate acquaintance. Russell was apparently unabashed by the consequence that such objects would be both private to the experience of particular persons and of very brief duration; he thought his doctrine of "logical constructions," which allowed for "inferred entities" on the basis of what is immediately certain, would provide the publicity and continuity necessary to do justice to actual experience. These assumptions, however, have met with serious criticism. P.F. Strawson, a British philosopher whose thought centres on the analysis of the structure of ordinary language, especially in his *Individuals: An Essay in Descriptive Metaphysics* (1959), not only attacked Russell's account of proper names but argued that experience demands a framework of basic particulars that are not Russell's momentary private objects but continuing public existents—in fact, individuals in the terminology explained above. If experience consisted of nothing but sounds, the minimum prerequisite of intelligibility would be that there should be a continuing master sound, an analogue in this medium of continuing material substance in the material order. Without such basic particulars as continuing material things, identification and reidentification would be impossible. Strawson conceded that persons as well as things were genuine continuants, but maintained all the same that the hypothesis that reality might consist of nothing but minds was quite untenable. Minds are no more than aspects of persons, and persons have bodies as well as minds. Strawson agreed that disembodied existence was logically possible, but added that such existence would make no sense except as a survival of embodied existence in a common public world.

Russell's theory of particulars

If this is correct, what exists cannot consist, as Hume supposed, of momentary items but must rather take the form of substances in the Aristotelian sense. These act as basic particulars in the actual intellectual scheme men adopt. Strawson, however, was not content merely to assert this fact; he wanted to argue that things must be like this if reference and description in their familiar form are to be possible at all. His main theory, which plainly owes a debt to Kant as well as to Wittgenstein, was worked out with primary reference to the physical world. It would be interesting to know if an examination of social reality would yield comparable results: whether individual persons or something larger—continuing societies or institutions—should be taken as basic particulars in that sphere. Many philosophers assert dogmatically that a society is nothing but an aggregate of its individual members. Nevertheless, men are members of society in virtue of their performance of a number of social roles, and role itself is a concept that makes sense only if the notion of society is presupposed. In one sense, a society is nothing apart from its members; remove them, and it would disappear. Equally, however, the members themselves are what they are because of their various roles; it is arguable that they would be nothing apart from their social relations. Hence, the force of Bradley's remark is evident, namely, that "the 'individual' apart from the community is not anything real."

It remains to add here that a number of philosophers have tried to argue that the basic items in reality should be described not as substances but in some other terms. Russell at one stage in his career spoke of the world as consisting of events; his former colleague A.N. Whitehead made the notion of process central in his metaphysics. Developments in modern physics undoubtedly lend a certain plausibility to these and similar views. Yet it remains difficult to understand what an event could be in which nothing was concerned, or how there could be a process in which nothing was in process. Event and process, in fact, are expressions that belong to derivative categories in the general Aristotelian scheme; like all other categories, they depend on the category of substance. If the latter is removed, as these metaphysicians propose to remove it, it is hard to know what is left.

THE EXISTENCE OF GOD

Perhaps the most celebrated issue in classical metaphysics concerned the existence of God. God in this connection is the name of "the perfect Being" or "the most real of all things"; the question is whether it is necessary to recognize the existence of such a being as well as of things that either are or might be objects of everyday experience. A number of famous arguments have been advanced from the time of the Greeks in favour of the thesis that such a recognition is necessary. The neatest and most ingenious was the a priori argument of St. Anselm in the 11th century, who said that "that than which nothing greater can be conceived" must exist in fact as well as in thought, for if it existed only in thought and not in fact, something greater than it could be conceived, namely the same thing existing in fact. God necessarily exists, because the idea of God is the idea of that than which nothing greater can be conceived. This is the argument later known as the ontological proof. Relatively few philosophical theologians, either in the Middle Ages or later, could bring themselves to accept this bold piece of reasoning (although Descartes, Spinoza, Leibniz, and Hegel all accepted it in principle); most preferred to ground their case for God's existence on premises that claimed to be empirical. Thus, St. Thomas Aquinas, perhaps the most influential Scholastic philosopher, in the 13th century argued that to explain the fact of motion in the world, the existence of a prime mover must be presupposed; that to account for contingent or dependent being the existence of something that is necessary or self-contained must be presumed; that to see why the world is orderly and why the different things in it fit together harmoniously, a situation that might not have obtained, a Creator who fashioned it on these lines must be postulated—adding in each case "and this all men call 'God'." These are versions of the first cause argument and the argument from design, which were to figure prominently in the thinking of later theistically inclined metaphysicians.

Proofs of the existence of God

The first cause argument should, perhaps, be examined

in somewhat greater detail, because it both has an immediate plausibility and lies at the basis of many different kinds of metaphysical systems (that of Hegel, for example, as well as that of Aquinas). The argument begins with the innocent-looking statement that something contingent exists; it may be some particular thing, such as oneself, or it may be the world in general (thus, the description of the proof as being *a contingentia mundi*, or "from the contingency of the world"). In describing oneself or the world as contingent, one means only that the thing in question does not exist through itself alone; it owes its being to the activity of some other thing, as a person owes his being to his parents. Contingent things are not self-complete; they each demand the existence of something else if they are to be explained. Thus, the move is made from contingent to necessary being; it is felt that contingent things, of whatever order, cannot be endlessly dependent on other contingent things but must presuppose a first cause that is self-complete and so exists necessarily. In Hegel the necessary being is not a separate existent but, as it were, an order of things; the loose facts of everyday life and even of science are said to point to a system that is all-embracing and in which everything is necessarily what it is. The principle of the argument, however, is unchanged despite the change in the conclusion.

Criticisms of proofs

Damaging criticism was brought against all the traditional arguments for God's existence by Hume and Kant in the 18th century. The ontological proof was undermined by the contention that "being is not a real predicate"; existence is not part of the concept of God in the way in which, for example, being all-powerful is. To say that something exists is not to specify a concept further but to claim that it has an instance; it cannot be discovered whether a concept has an instance by merely inspecting it. The first cause argument, it was contended, suffers from two fatal weaknesses. Even if it is correct in its assertion that contingent being presupposes necessary being, it cannot identify the necessary being in question with God (as happened in each of the Thomistic proofs) without resurrecting the ontological argument. If it is true, as supporters of the causal proof suppose, that God alone can answer the description of a necessary being, then whatever exists necessarily is God and whatever is God exists necessarily. Modern supporters of the causal proof have tried to meet this objection by saying that the equivalence is one of concepts, not of concept and existent; the existence of a necessary being is already established in the first part of the argument, and the equivalence in the second part of the argument is between the concept of necessary being and the concept of God. In other words, they distinguish between existence and essence. In the first part of the argument, the existence of a necessary being is proved; in the second part of the argument, the essence of that necessary being is identified with what men call God. Beyond this first contended weakness, however, there are grave difficulties in the move from contingent to necessary existence. Things in the experienced world are causally related, and some account of this relationship can be given in terms of the temporal relations of events; causal relations hold primarily between kinds of events, and a cause is, at least, a regular antecedent of a specific kind of effect. But when an attempt is made to extend the notion of causality from a relationship that holds within experience to one that connects the experienced world as a whole to something that falls wholly outside it, there is no longer anything firm on which to hold. The activities of God cannot precede happenings in the world because God is, by definition, not in time; and how the relationship is to be understood in these circumstances becomes highly problematic. Some metaphysicians, like some recent theologians, seek to evade the difficulty by saying that God is not the cause of the world but its ground, or again by distinguishing causes of becoming, which are temporal, from a cause of being, which is not. It is doubtful whether these moves do more than restate the problem in different terms.

The argument from design is itself a form of causal argument and accordingly suffers from all the difficulties mentioned above, together with some of its own, as Hume and Kant both point out. Even on its own terms, it is wrong to conclude the existence of a Creator rather than an architect. Furthermore, it infers that the being in question has unlimited powers, when all that the evidence seems to warrant is that its powers are very great. The argument retains a certain rhetorical appeal, although its force was much diminished by the publication of the English naturalist Charles Darwin's theory of natural selection. The unbroken reign of law throughout natural evolution is impressive, but as a line of reasoning it does not seem to bear close examination.

The metaphysical problem of God's existence is more of an issue today than the problem of universals; there are still thinkers who hope to restate the old proofs in more convincing ways. The ontological proof, in particular, has won renewed attention from thinkers as different as Norman Malcolm, a philosopher strongly influenced by Wittgenstein, and Charles Hartshorne, an American Realist whose form of theism is called panentheism (the doctrine of a God who has an unchanging essence but who completes himself in an advancing experience). Increasingly, however, philosophers of religion are preoccupied not with these metaphysical abstractions but with the status and force of actual religious claims. "The most real of all things" is no longer at the centre of their attention: what they seek to investigate is God as a suitable object for worship.

THE SOUL, MIND, AND BODY

Platonic and Aristotelian theories of the soul

The soul–body relationship. As well as believing in the reality of Forms, Plato believed in the immortality of the human soul. The soul was, he thought, an entity that was fundamentally distinct from the body, although it could be and often was affected by its association with the body, being dragged down by what he called in one passage "the leaden weights of becoming." The soul was simple, not composite, and thus not liable to dissolution as were material things; further, it had the power of self-movement, again in contrast to material things. Ideally, the soul should rule and guide the body, and it could ensure that this situation persisted by seeing that the bodily appetites were indulged to the minimum extent necessary for the continuance of life. The true philosopher, as Plato put it in the *Phaedo*, made his life a practice for death, because he knew that after death the soul would be free of bodily ties and would return to its native element. He also knew, incidentally, that the soul was "akin" to the Forms; it was through the intellect that the Forms were discovered.

Plato mentioned and attempted to refute alternative accounts of the relationship of soul and body, including a Pythagorean view that described the soul as an "attunement" of the body and thus tried to explicate it as a form or structure rather than an independently existing thing. A theory of this kind was worked out, but not taken to its logical conclusion, by Aristotle in his treatise *De anima* ("On the Soul"). Aristotle defined soul in terms of functions. The soul of a plant was concerned with nutrition and reproduction, that of an animal with these and with sensation and independent movement, that of a man with all these and with rational activity. The soul was, in each case, the form of some body, and the clear implication of this was that it would disappear as the body in question dissolved. To be more accurate, the soul was the principle of life in something material; it needed the material element to exist, although it was not itself either material or immaterial but, to put it crudely, an abstraction. Even though Aristotle was clearly committed by everything he said in the earlier parts of the *De anima* to the view that the soul is not anything substantial, he nevertheless distinguished toward the end of this work between what he called the active and the passive intellects and spoke of the former in Platonic terms. The active intellect was, it appears, separate from the rest of the soul; it came "from outside" and was in fact immortal. Moreover, it was essential to the soul considered as rational, for "without this nothing thinks." Aristotle thus showed the Platonic side of his thought in the very act of trying to emancipate himself from this aspect of Platonism.

The mind–body relationship. In more recent metaphysics less has been heard of the soul and more of the mind; the old problem of the relationship of soul and body is now that of the relationship of mind and body. Most, if not all, subsequent discussion of this subject has been affected by the thinking of Descartes. In his *Meditationes de Prima Philosophia* (1641), he argued that there was a total and absolute distinction between mental and material substance. The defining characteristic of matter was to occupy space; the defining characteristic of mind was to be conscious or, in a broad sense of the term, to think. Material substance was, so to speak, all one, although packets of it were more or less persistent; mental substance existed in the form of individual minds, with God as the supreme example. The mental and the material orders were each complete in themselves, under God; it was this fact that made it appropriate for him to use the technical term substance in this context: mental substance and material substance. The logical consequence of this view, drawn by some later Cartesians, was that there can be no interaction between mind and body; all causality is immanent, within one order or the other, and any appearance of mind affecting body or of body affecting mind must be explained as the result of a special intervention by God, who, on the occasion of changes in one substance, brings it about that there are corresponding changes in the other. There is nothing, however, to suggest that Descartes himself had any sympathy with this view, which was called occasionalism. On the contrary, he stated explicitly that he was not in his body as a pilot is in a ship but was "more intimately" bound up with it. Mind could affect body and vice versa because mind and body had a specially close relationship, which was particularly evident in the aspects of conscious life that have to do with sensation, imagination, and emotion as opposed to pure thought.

Cartesian theory

Descartes's conviction that, despite their intimate union in this life, mind is really distinct from body sprang from his confidence in the *cogito* argument. It was possible, he believed, to doubt the existence of his body (what was certain was only that he had the experience of having a body, and this might be illusory) but not the existence of his mind, for the very act of doubting was itself mental. That mind existed was evident from the immediate testimony of consciousness; that body existed was something that needed an elaborate proof, involving his doctrine of clear and distinct ideas and his attempt to establish the existence of a God who is no deceiver. Apart from this, Descartes appealed to arguments of a broadly Platonic type to bring out what was truly distinctive about mind. He admitted that sensation and imagination could be understood only if referred to the mind–body complex, but contended that acts of the pure intellect and of will (here his thought was influenced by that of St. Augustine, the great 5th-century Christian thinker) belonged to the mind as it was in itself. Descartes did not claim to have a philosophical proof of the immortality of the soul—that, in his view, required the assurance of revelation—but he did think that his theory prepared the way for that doctrine by establishing the separate existence of mind.

The Cartesian account of mind and body had many critics even in Descartes's own day. Hobbes argued that nothing existed but matter in motion; there was no such thing as mental substance, only material substance. Materialism of a sort was also supported by Descartes's correspondent Pierre Gassendi, a scientist and Epicurean philosopher. A generation later Spinoza was to refashion the whole Cartesian metaphysics on bold lines. In place of the two distinct substances, each complete in itself yet each liable to external interference should God will it, Spinoza posited a single substance, God or Nature, possessed of infinite attributes, of which the mental and the material alone are known to men. The "modes," or manifestations, of this substance were what they were as a result of the necessities of its nature; arbitrary will neither did nor could play any part in its activities. Whatever manifested itself under one attribute had its counterpart in all the others. It followed from this that to every mental event there was a precisely corresponding physical event, and vice versa. A man was thus not a mysterious union of two different elements but a part of the one substance that, like all other parts, manifested itself in different ways under different attributes. Spinoza did not explain why it was that physical events could be correlated with mental events in the case of a human being but not in that of, for example, a stone. His theory of psycho-physical parallelism, however, has persisted independently of his general metaphysics and has found supporters even in modern times.

Theory of Spinoza

One way in which Spinoza threw fresh light on the mind–body problem was in calling attention to the influence of the body on the mind and in taking seriously the suggestion that they be treated as a single unit. In this respect, his work on the subject was far in advance of the Empiricist philosophers of the next century. Hume notoriously dismissed Cartesian substance as a "chimera" and argued that minds and bodies alike were nothing but "bundles of perceptions," interaction between which was always possible in principle; in practice, however, he stuck to the old-fashioned view that mind is one thing and body another and did nothing to explore their actual relationships. Empiricist philosophy of mind, both in Hume and in his successors, such as James Mill, was generally crude; it consisted largely in an attempt to explain the entire life of the mind in terms of Hume's ontology of impressions and ideas. Nor did Kant make much, if any, advance in this particular direction, convinced as he was of the necessity of accepting an empirical dualism of mind and body. It was left to Hegel and the Idealists to look at the problem afresh and to bring out the way in which mental life and bodily life are intimately bound together. The accounts of action and cognition given by T.H. Green and Bradley, and more recently by R.G. Collingwood, are altogether more enlightening than those of Empiricist contemporaries just because they rest on a less dogmatic basis and a closer inspection of fact.

Modern theories

No metaphysical problem is discussed today more vigorously than that of mind and body. Three main positions are held. First, there are still writers (*e.g.*, H.D. Lewis in his work *The Elusive Mind* [1969]) who think that Descartes was substantially right: mind and body are distinct, and the "I" that thinks is a separate thing from the "I" that weighs 170 pounds. The testimony of consciousness is invoked as the main support of this conclusion; it is alleged that all men know themselves to be what they are, or at least who they are, apart from their bodily lives; it is alleged again that their bodily lives present themselves as experiences—*i.e.*, as something mental. The existence of mind, as Descartes claimed, is certain, that of body dubious and perhaps not strictly provable. Second, there are writers such as Gilbert Ryle who would like to take the Aristotelian theory to its logical conclusion and argue that mind is nothing but the form of the body. Mind is not, as Descartes supposed, something accessible only to its owner; it is rather something that is obvious in whatever a person does. To put it crudely, mind is simply behaviour. Finally, there are many philosophers who, although more generally sympathetic to the second solution than to the first, wish to provide for an "inner life" in a way in which Behaviourism does not; P.F. Strawson is a typical example. To this end they try to assert that the true unit is neither mind nor body but the person. A person is something that is capable of possessing physical and mental predicates alike. This is, of course, to say that the "I" that knows simple arithmetic and the "I" that has lost weight recently are the same. How they can be the same, however, has not so far been explained by supporters of this view.

Aside from these main positions, an interesting development is the stress laid by writers—such as Stuart Hampshire, an "ordinary language" philosopher—on self-activity as the distinguishing characteristic of mind. According to this view, a human being is a body among bodies but is, as Plato said, self-moving as material things are not. That this should be so—that human beings are possessed of wills and can in favourable circumstances act freely—is taken as an ultimate fact neither requiring nor capable of

explanation. It is often denied that any scientific discovery could give rational grounds for questioning this fact. It is also stressed that the causality of a human being is fundamentally different from that of a natural subject, intentional action being quite other than mere behaviour determined from without.

The problem of differentiating men and machines

Connected with these topics is the problem, much discussed in recent philosophy as a result of the rise of cybernetics, of what differentiates men from machines. Two answers used to be given: the power to think and consciousness. Now, however, there exist machines whose calculating abilities far surpass those of any human being; such machines may not literally think, but they certainly arrive at conclusions. Furthermore, it is not true that their operations are of a purely routine nature: there is a sense in which they can improve their performance in the light of their "experiences." They even have an analogue of consciousness in the sensitivity they show to external stimuli. These facts suggest that the gap between minds and machines is less wide than it has often been thought to be; they do not, however, destroy it altogether. Human beings possess powers of creative thought unlike anything found in machines; as Noam Chomsky, an American linguistics scholar, has stressed (and as Descartes urged in his *Discours de la méthode*), the ability of human beings to handle language in such a way that they comprehend any one of an infinite number of possible expressions is something that cannot be explained in mechanical terms. Again, as J.R. Lucas, a British philosopher, has argued, human beings have the ability to diagnose and correct their own limitations in a way to which there is no parallel in machines. As some older philosophers put it, man is a being with the power of self-transcendence; he can work within a system, but he can also move to another level and so see the shortcomings of the system. A machine can only work within a system; it operates according to rules but cannot change them of its own accord.

Finally, mention should be made of an extreme Materialist solution to the mind–body problem: this solution holds that states of mind are in fact states of the brain. Supporters of this theory agree that the two are separate in idea but argue that physiology shows that despite this they are contingently identical. What seems to be a state of mind, above all to its possessor, is really a state of the brain, and mind is thus reduced to matter after all. It is not clear, however, why physiologists should be granted the last word on a topic like this, and, even if it were agreed that they should be, the correlations so far established between mental occurrences and states of the brain are at best sketchy and incomplete. Central-state Materialism, as this theory is called, professes to have the weight of contemporary science behind it, but it turns out in fact to have drawn to a remarkable degree on what it thinks will be the science of tomorrow.

NATURE AND THE EXTERNAL WORLD

The problem of the existence of material things, first propounded by Descartes and repeatedly discussed by subsequent philosophers, particularly those working within the Empiricist tradition, belongs to epistemology, or the science of knowledge, rather than metaphysics; it concerns the question of how it can be known whether there is a reality independent of mind. There are, however, problems about nature and the external world that are genuinely metaphysical.

The reality of material things. There is first of all the question of the status or standing of material things, the kind of being they possess. It has been repeatedly suggested by metaphysical philosophers that the external world is in some way defective in reality, that it is a mere phenomenon, something that seems to be what it is not.

The external world as defective in reality

Plato, as has already been pointed out, held that objects of the senses generally answered this description; they each appeared to possess characteristics that they could not in fact have (water could not be at once hot and cold) and were to that extent delusive rather than real. There was no stability in the world of phenomena and therefore no true reality. In taking this view, Plato drew no contrast between the world of nature and the world of man, although he undoubtedly believed that souls had a superior status. Leibniz, a later philosopher who also followed this general line of thought, began by explicitly opposing souls to material things. To speak precisely, nothing truly existed except monads, and monads were souls, or spiritual beings: all had perceptions, although these varied enormously in degree of clarity (the perceptions of the monads constituting what is commonly called a stone were singularly faint). Although the final description of the world must thus be given in mental terms, it did not follow that nature as normally perceived is a total illusion. Men perceive as well as think, and although perception is in fact simply a confused form of thought it is not for that reason to be set aside altogether. The world of nature, the world of things in space and time, is, as Leibniz put it, a "well-founded phenomenon"; it is what all men must judge to be there, given that they are not pure intellects but necessarily remain to some extent prisoners of their senses.

A theory on somewhat similar lines was worked out by Kant in the *Kritik der reinen Vernunft* (*Critique of Pure Reason*), despite Kant's explicit dissent from Leibniz' account of perception as confused thinking. Kant contrasted a realm of things as they are in themselves, or noumena, with a realm of appearances, or phenomena. The former are unknown, and indeed unknowable, though it seems clear that Kant tended to think of them on lines like those of Leibniz; phenomena do not exist independently, but are dependent on consciousness, though not on any one person's consciousness. Kant expressed this position by saying that things phenomenal are empirically real but transcendentally ideal; he meant that they are undoubtedly there for the individual subject, though when examined from the point of view of critical philosophy they turn out to be conditioned by the mind through the forms of the understanding imposed upon them. Kant's most striking argument for this conclusion was that space and time are neither, as the English physicist Sir Isaac Newton supposed, vast containers inside which everything empirical is situated, nor, as Leibniz had suggested, relations between things confusedly apprehended, but are rather what he mysteriously called "pure intuitions," factors inherent in the sensibilities of observers. Without observers, space and time disappear, along with their contents; but once the human point of view is assumed, in the form of percipients who are directly aware of the world through their senses, space and time become as real as anything—indeed, more real because of their pervasive character. There is nothing that falls within experience that does not have temporal relations, and all the data of the external senses have spatial relations as well.

Kant's arguments in support of his revolutionary thesis about space and time unfortunately depend to a large extent on his mistaken philosophy of mathematics, and they have accordingly been discounted by later philosophers. In modern philosophy the issues raised in these discussions survive only in the form of an inquiry into the status of nature as investigated by the natural scientist.

Primary and secondary qualities of material things

Descartes already pointed out that material things in fact have properties different from those they seem to have; they appear to possess secondary qualities such as colour or smell, but turn out when thought about strictly to be colourless and odourless lumps of matter occupying and moving about in space. Locke endorsed this distinction between primary qualities (such as extension, motion, figure, and solidity) and secondary qualities; but George Berkeley, a major British Empiricist of the early 18th century, criticized it sharply as absurd: to imagine something that has primary but no secondary qualities is psychologically impossible. For Berkeley the world of the scientist was a fiction, and perhaps not even a necessary fiction at that. It seems clear, however, that Berkeley's arguments do not undermine the important distinction between primary and secondary qualities, where the former are treated as fundamental and the latter as derivative; they are valid only against Locke's mistaken claim that primary qualities are objective and secondary quali-

ties subjective. Whatever the explanation, the fact remains that the scientist often knows why the phenomena are as they are, in contrast to the plain man; to that extent nature, as he understands it, is truer, if not more real, than nature as it is taken to be in everyday experience. Why this should be is not satisfactorily explained by philosophers who follow Berkeley's lead on this question. Nor has either party to the controversy noted sufficiently the extent to which nature as commonly thought of is conceived as penetrated by mind, both when it is taken as intelligible and, still more interestingly, when poets ascribe to it moods or treat it as kindly or hostile. There is analytic work to be done here to which critical philosophers have still to address themselves.

The organizing principles of nature. Connected with the questions just discussed are problems about the organizing principles of nature; *i.e.*, about natural causality. It has been said that the Greeks thought of the world as a vast animal (indeed, the conceptual scheme that Aristotle devised for dealing with nature makes sense only if something like this is presupposed). Nature is the sphere in which different kinds of things are all striving to realize their characteristic form; purpose, though not perhaps explicit purpose, governs it throughout. Aristotle was not entirely insensitive to what are now known as the physical and chemical aspects of the universe, but he treated them as subordinate to the biological aspect in a way modern thinkers find surprising. Even the four elements —earth, air, fire, and water—were seen by him as each seeking its natural place in the cosmos. The contrast between this view and that favoured by Descartes could hardly be sharper. According to Descartes, nature is not an organism but a mechanism; everything in it, including animal and human bodies, although not including the human mind, must be understood on mechanical principles. In taking this line, Descartes was endorsing a way of thinking that was central in the new physical science developed by Galileo at the beginning of the 17th century and that was to remain central in the thought of Newton. Descartes himself was not a pure mechanist, because he believed that mind was governed by principles of its own; his work, however, undoubtedly encouraged the thought, frequently debated at the time of the Enlightenment, that mental life equally with the physical world must be explicable in mechanical terms. This was a position whose validity at the theoretical level Kant reluctantly admitted, only to try to turn its edge by his dichotomy of theory and practice. Everything in nature, including human behaviour, was subject to causal determination. The dignity and uniqueness of man, however, could be preserved because of the fact that in moral action man raised himself above the sphere of nature by thinking of himself as part of a world of free spirits.

Nature as mechanical or teleological

Kant also produced interesting thoughts on the subject of living phenomena. Reflection on the concept of an organism had convinced him that a being of this sort could never be accounted for satisfactorily in mechanical terms; it was futile to hope that someday in the future there would appear a Newton of biology capable of explaining mechanically the generation of even so apparently simple a thing as a blade of grass. To judge or speak of organic phenomena demanded a special principle that was teleological (*i.e.*, related to design or purpose) rather than mechanical. Kant, however, refused to allow that this principle had constitutive force. It belonged, he said, only to "reflective judgment" and thus did not rank alongside the principles of understanding that were so important in physical science. Men must have recourse to a principle of purposiveness in order to speak of living things, but they must not imagine that such recourse would enable them to explain their existence and behaviour in any strict sense of the term. They have insight only into what they can produce, and what they can produce are machines, not organisms. Many of Kant's detailed remarks on this subject seem outmoded in the light of subsequent scientific developments; nevertheless, the problem he raised is still the subject of vigorous debate among philosophically minded biologists. His emphasis on the uniqueness of the concept of an organism, which he says is only imperfectly explicated in the language of ends and purposes, is particularly valuable.

It remains to mention the seemingly eccentric view of nature taken by Hegel, who regarded it as at once the antithesis to and a prefiguration of the world of spirit. Nature had to exist to provide material for spirit to overcome, although it was a gross mistake to think of it as essentially a lifeless mechanism. Instead of reducing the organic to the inorganic, men should see the latter as pointing forward to the former, which in turn offered a foretaste of the rational structure exhibited by the world of mind. Hegel's disdain for scientists of proved ability, such as Newton and John Dalton, and his endorsement against them of amateur scientists such as the German writer Goethe, make it hard to take his philosophy of nature seriously. It contains, even so, some interesting points, not least the demonstration that in finding nature to be throughout subject to law the scientist is presupposing that it is thoroughly penetrated by mind. To understand these views properly, however, it is necessary to understand Hegel's system as a whole.

SPACE AND TIME

Many metaphysicians have argued that neither time nor space can be ultimately real. Temporal and spatial predicates apply only to appearances; reality, or what is real, does not endure through time, nor is it subject to the conditions of space. The roots of this view are to be found in Plato and beyond him in the thought of the Eleatic philosophers Parmenides and Zeno, the propounder of several paradoxes about motion. Plato conceived his Forms as eternal objects whose true location was nowhere. Similarly, Christian philosophers conceived of God as existing from everlasting to everlasting and as present in all parts of the universe. God was not so much in space and time as the source of space and time. Whatever falls within space and time is thereby limited, for one space excludes another and no two times can be simultaneous. God, however, is by definition an infinite being and so must exist timelessly and apart from space.

Theories of Kant and Bradley

Reference has already been made to the way in which Kant argued for an intimate connection between time and space and human sensibility: that human beings experience things as being temporally and spatially situated is to be connected with the nature of their minds, and particularly with their sensory equipment. Kant was entirely correct to describe space and time as "intuitions," by which he meant that they are peculiar sorts of particulars; he was right again to insist on the centrality in sensing of the notions of here and now, which can be indicated but not reduced to conceptual terms. It is highly doubtful, however, whether he had sufficient grounds for claiming a priori insight into the nature of space and still more that of time; his case for thinking that space and time are "pure" intuitions was palpably inadequate. The lesson to draw from his careful discussion of this subject might well be not that there must be a form of reality lying beyond space and time but rather that nothing can be real that does not conform to spatial and temporal requirements. Space and time are bound up with particularity, and only what is particular can be real.

It was only in a weak sense that Kant denied the reality of time and space. Other philosophers have certainly been bolder, though generally on the basis of a less solid grasp than Kant possessed of what it is to experience temporally and spatially. Thus, Bradley argued against the view that space and time are "principles of individuation" by alleging that no specification of spatial or temporal position, whether in terms of here and now or by the use of spatial coordinates or dating systems, could achieve uniqueness. Any descriptions such as "at 12 o'clock precisely on January 4, 1962" or "just 75 yards due north of this spot" might apply to infinitely many times or places in the universe, for there was nothing to prevent there being infinitely many temporal and spatial orders. Bradley forgot that the whole meaning of a spatial or temporal description is not exhausted when attention is given to the connotations of the terms used; what has to be considered is the words as used in their context, which

is that of a person who can indicate his position in space and time because of the fact that he is himself situated in space and time. One cannot express uniqueness in words as such, but he can use words to express uniqueness. Bradley's suggestion that it is possible to conceive of many temporal and spatial orders is by no means free from controversy. In general, men think of all events as happening before, simultaneously with, or after the moment that is called "now," all spatial positions as relating in some way or other to the point that is called "here." In circumstances where this cannot be done, as with events or places in a dream, men dismiss them as quite unreal. That there might be events or places with no relation to their own now and here is something they often refuse to take seriously, though there are theories in modern science that suggest that they are wrong to do so.

It was pointed out earlier that to say that something is unreal in a metaphysical context is often to say that it is unintelligible, and it is not surprising to find that arguments about the unreality of space and time have often turned on conceptual considerations. Thus, it is alleged that there is an incoherency in the notion of space because it claims to be a whole that is logically prior to its parts, and nevertheless turns out in practice to be merely an indefinitely extensible aggregate. Everything that occupies space falls within a wider spatial context; the thought of space as such is, as Kant saw, involved in any spatial description. Yet space as such is something that constantly eludes man's grasp; space, as man knows it, is just one spatial situation after another.

Theory of Bergson

The difficulties found in the notion of time turn on the combination in it of the idea that time is continuous and the idea that it is made up of discrete parts. Henri Bergson, a French philosopher who was concerned with the notions of duration and movement, said that time was experienced as continuous; it was only the "spatialized" time measured by clocks that was taken to have separable parts (minutes, hours, weeks, and so on), and this "public" time was merely conventional. This, however, seems altogether too easy a solution of the problem, for privately experienced time also goes by (one stretch of it follows another), and the thesis that public time is merely conventional is at best highly controversial. It must be allowed that time is commonly thought of as at once flowing and, as it were, subject to arrest. Whether this is, in fact, openly inconsistent may be doubted, but it is on points like this that the metaphysical case in question rests.

Few British or American philosophers discuss these questions now, largely because they have been persuaded by Moore that any attack on such central notions in men's thought as these must be mistaken in principle. As a result, little attention is given to a question that deserves investigation; namely, what is to take the place of space and time in metaphysical thought. Idealist writers constantly said that space and time qualified appearances, and that nothing that did so could fail to be taken up in the higher experience that was experience of reality. But how is this supposed to be done? Time is perhaps cancelled and yet preserved in the idea of eternity, space in the thought of something that is at once omnipresent yet not in any particular place. But what is there that is positive about these notions? The eternal, it is sometimes said, is not to be identified with what lasts through all time; it is, strictly, outside time altogether. But what does it mean to say this? When it is said, for example, that numbers or truths are eternal, the proper inference is that they have nothing to do with time; to inquire when they came into or will go out of existence is to ask a question that is ill posed. When God, however, is said to be eternal, the impression is often given that he has temporal characteristics, although in some higher form. What this higher form is deserves careful consideration, the result of which might be that it is not the conception of time that is incoherent but the conception of God.

THE CONCEPTION OF SPIRIT

As well as arguing for the separate existence of mental substance, metaphysicians have claimed that mind is, as it were, the key to the understanding of the universe. What exists is spirit, or at least is penetrated by spirit. This is the thesis of Idealism, a type of philosophy that is often derided but that, like its rival Materialism, has a constantly fresh appeal. This view is worth examining in more detail than has so far been possible.

It is best to begin by distinguishing the thesis of Idealism proper from some others with which it is readily confused. Leibniz said that the true atoms of nature were monads or souls; at bottom nothing existed except minds. Berkeley claimed that sensible things have no existence without the mind; there are spirits that experience, including an infinite spirit, and there are the contents of their experiences, but there is no independently existing world of matter. For the philosophers who followed Hegel, both Leibniz and Berkeley were "subjective" Idealists: they conceived of reality in terms of the experiences of individual minds. Hegel's view, by contrast, was that what exists is not so much pure mind as mind writ large; *i.e.*, the universe is penetrated by mind and exists for the sake of mind, and it cannot be understood unless this fact is grasped. Hegel was thus not committed to denying that there is an independent world of nature but, on the contrary, openly proclaimed it. Nature was there for mind to master it, and in so doing to discover itself.

Theory of Hegel

The field in which Hegel first worked out this theory was that of human affairs. The human world may be said to be mind made objective because it consists of a series of structures—examples would be a language, a set of moral or political procedures, a science, a practical art such as medicine—that constitute mental achievements. The mind involved in structures of this kind, however, is collective rather than personal. An art such as medicine or a science such as mathematics is not the invention of any particular individual; and although individuals have contributed and are contributing to the advancement of each structure, they do so not in their personal capacity but as embodying impersonal intelligence.

Because the human world thus embodies mind or spirit, it needs to be understood in a special way—in terms of what Hegel called "concrete universals." Concepts of this kind are in order when it is a question of grasping a particular sort of subject matter—one in which there are intimate connections between the data under consideration. Connections in nature are, on the surface at any rate, of a purely external character; striking a match, for example, has nothing internally to do with producing a flame. When, however, a historian considers the different stages of some movement or process, or when an anthropologist studies the various aspects of the life of a society, the material they confront is internally related just because it represents the work of mind—not, of course, of mind working in a vacuum, but of mind facing and reacting with greater or less intelligence to particular situations. It is not surprising in these circumstances to find that the conceptual structure employed by the student of human affairs is, in important respects, profoundly different from that employed by the student of nature. In the latter, what are in question are constant conjunctions, observed but not understood; in the former, men have insight into what happens or obtains, because they can re-enact in their own minds the thought that lay behind the material they have to study.

All this is, or should be, comparatively uncontroversial; it represents the truth behind the claim of Wilhelm Dilthey, a German philosopher and historian of ideas, that human affairs can be understood, as it were, from within, by means of what he called *Verstehen* ("understanding"). But of course it is one thing to say this and another altogether to argue that the universe at large should be construed as if it were mind writ large. What makes Hegelianism intriguing to some, and totally implausible to others, is precisely that it makes this extravagant claim. As has already been mentioned, the world of nature for Hegel is in one way independent of mind: its being is certainly not its being perceived. Nevertheless, it is relevant to mind in all sorts of important ways: in providing a setting in which mind can act, in constituting an obstacle that mind can overcome, in presenting mind

with something seemingly alien in which it can nevertheless find itself insofar as it discovers nature to be intelligible. If Hegel were asked why there was a world of nature at all, his answer would be "for the sake of mind." Just as man's social environment affords opportunities to the individual to come to full knowledge of himself by realizing his differences from and dependence upon others, so the world of nature affords similar opportunities. By transforming the natural scene, men make it their own. In so doing they come to know what they can do, and thus what they are.

All reality as spirit

There is, perhaps, more to this doctrine than appears at first sight. It is, however, easier to assent to it in general terms than to follow Hegel over it in detail. According to the Idealist account, there is in the end only one true description of the universe, that which is couched in terms of the concrete universal. Reality is a single self-differentiating system, all the parts of which are intimately connected; it is spirit expressing itself in the natural and human worlds and coming to consciousness of itself in so doing. Any other account of the matter—for example, that given by the scientist in terms of experienced uniformities—must be dismissed as inadequate. To Hume's objection that there is an absolute logical difference between propositions expressing matters of fact and existence and propositions expressing relations of ideas, Hegel replies brusquely that the distinction is untenable. At a certain level, perhaps, facts are taken as "brute." Even the scientist, however, never abandons his aspiration to understand them—it is only provisionally that he talks in terms of "ultimate inexplicabilities"—and the philosopher knows that the demand to incorporate all knowledge in a single system is not to be denied. It is a demand that, as Hegelians admit, can in practice never be met but that nonetheless ceaselessly makes itself felt. That such is the case is shown by the extraordinary fascination exercised by this strange but remarkable type of philosophy.

To try to understand the universe in terms of spirit is characteristic of philosophers whose main extra-philosophical interests are in the humanities, particularly in historical studies. Relatively few scientifically minded thinkers have followed this line of thought, and many Idealists of repute, including Bradley and Benedetto Croce (an Italian philosopher and literary critic whose major philosophical work was published in four volumes between 1902 and 1917 under the general title *Filosofia come scienza dello spirito*), have been least convincing when writing about science. Hegel himself, perhaps, had less sympathy with scientific than with historical aspirations; this is not to say, however, that he was ill-informed about contemporary science. He knew what was going on, but he saw it all from his own point of view, the point of view of one who was entirely convinced that science could not produce any ultimate answers. He valued science but rejected the scientific view of the world.

III. Types of metaphysical theory

To complement and, in a way, to correct this brief survey of the problems of metaphysics it will be useful at this point to insert a short summary of a number of overall metaphysical positions. Metaphysics, as already noted, professes to deal with "the world as a whole"; the thoughts of a metaphysician, if they are to make any impact at all, must be connected in a system. The object in what follows will be to present in outline metaphysical systems that have exercised and, indeed, continue to exercise a strong intellectual appeal. In all cases but one, these systems were given classical shape by particular philosophers of genius. Relatively little attention, however, will be paid to this fact here, because the present concern is with types of view rather than with views actually held. Thus, reference will be made to Platonism instead of to the philosophy of Plato, and so on in other cases.

PLATONISM

The essence of Platonism lies in a distinction between two worlds, the familiar world of everyday life, which is the object of the senses, and an unseen world of true realities, which can be the object of the intellect. The ordinary man recognizes the existence of the former and ignores that of the latter; he fails to appreciate the extent to which his beliefs both about fact and about values are arbitrarily assumed and involve internal contradictions. The philosopher, or rather the true philosopher, is in a position to show him how insubstantial is the foundation on which he takes his stand. The philosopher can demonstrate how little thought there is in popular conceptions of good and evil, and he can show that the very concept of sense knowledge involves difficulties, because knowledge presupposes a stable object, and the objects of sense are constantly changing. The claim, however, is that he can do more than this. Because of the presence in him of something like a divine spark, he can, after suitable preparation, fix his intellectual gaze on the realities of the unseen world and, in the light of them, know both what is true and how to behave. He will not attain this result easily—to get to it will involve not only immense intellectual effort, including the repeated challenging of assumptions, but also turning his back on everything in life that is merely sensual or animal. Yet, despite this, the end is attainable in principle, and the man who arrives at it will exercise the most important part of himself in the best way that is open to him.

Distinction between the world of appearances and the world of realities

That this type of view has an immediate appeal to persons of a certain kind goes without saying. There is ample evidence in poetry and elsewhere of the frequently experienced sense of the unreality of familiar things and the presence behind them of another order altogether. Platonism may be said to build on "intuitions" of this kind; as a metaphysics, its job is to give them intellectual expression, to transfer them from the level of sentiment to that of theory. It is important, however, to notice that Platonism is not just the intellectualizing of a mood; it is an attempt to solve specific problems in a specific way. In Plato's own case, the problems were set by loss of confidence in traditional morality and the emergence of the doctrine that "man is the measure of all things." Plato thought he could counter this doctrine by appeal to another contemporary fact, the rise of science as shown in the development of mathematical knowledge. Mathematics, as he saw it, offered certain truth, although not about the familiar world; the triangle whose properties were investigated by the geometrician was not any particular triangle but the prototype that all particular triangles presuppose. *The* triangle and *the* circle belonged not to the world of the senses but to the world of the intelligence; they were Forms. If this could be said of the objects of mathematical discourse, the same should also be true of the objects of morality. True justice and true goodness were not to be found in popular opinions or human institutions but should be seen as unchanging Forms, eternally existing in a world apart.

Modern philosophers have found much to criticize in this system: as indicated already, they have objected that Forms are not so much existents as abstractions, and they have found the argument from science to morality quite inconclusive because of what they allege to be an absolute dichotomy between fact and value. It may be that nobody today can subscribe to Platonism in precisely the form given it by Plato himself. The general idea, however, has certainly not lost its hold, nor have the moral perplexities to which Plato hoped to find an answer been dissipated by further thought.

ARISTOTELIANISM

For many people, Plato is the type of an other-worldly, Aristotle of a this-worldly philosopher. Plato found reality to lie in things wholly remote from sense; Aristotle took form to be typically embodied in matter and thought it his job as a philosopher to make sense of the here and now. The contrast is to some extent overdrawn, for Aristotle, too, believed in pure form (God and the astral intelligences—the intelligent movers of the planets—were supposed to satisfy this description), and Plato was sufficiently concerned with the here and now to want to change human society radically. It remains true, nevertheless, that Aristotelianism is in essentials a form of

Making sense of the here and now

immanent metaphysics, a theory that instructs men on how to take the world they know rather than one that gives them news of an altogether different world.

The key concepts in Aristotelianism are substance, form and matter, potentiality and actuality, and cause. Whatever happens involves some substance or substances; unless there were substances, in the sense of concrete existents, nothing could be real whatsoever. Substances, however, are not, as the name might suggest, mere parcels of matter; they are intelligible structures, or forms, embodied in matter. That a thing is of a certain kind means that it has a certain form or structure. But the structure as conceived in Aristotelianism is not merely static. Every substance, in this view, not only has a form but is, as it were, striving to attain its natural form; it is seeking to be in actuality what it is potentially, which is in effect to be a proper specimen of its kind. Because this is so, explanation in this system must be given in teleological rather than mechanical terms. For Aristotle, form is the determining element in the universe, but it operates by drawing things on, so that they become what they have it in themselves to be, rather than by acting as a constant efficient cause (*i.e.*, the agent that initiates the process of change). The notion of an efficient cause has a role in Aristotelianism—as Aristotle put it, it takes a man, a developed specimen of his kind, to beget a man; it is, however, a subordinate role and yields pride of place to a different idea, form considered as purpose.

For reasons connected with his astronomy, Aristotle postulated a God. His God, however, had nothing to do with the universe; it was not his creation, and he was, of necessity, indifferent to its vicissitudes (he could not otherwise have been an unmoved mover). It is a mistake to imagine that everything in the Aristotelian universe is trying to fulfill a purpose that God has ordained for it. On the contrary, the teleology of which use is here made is unconscious; although things all tend to an end, they do not in general consciously seek that end. They are like organs in a living body that fulfill a function and yet seemingly have not been put there for that purpose.

As this last remark will suggest, an important source of Aristotelian thought is reflection on natural growth and decay. Aristotle, who was the son of a doctor, was himself a pioneer in natural history, and it is not surprising that he thought in biological terms. What is surprising, and gives his system a continuing interest, is the extent to which he succeeded in applying ideas in fields that are remote from their origin. He was without doubt more successful in some fields than in others: in dealing with the phenomena of social life, for instance, as opposed to those of physical reality. His results overall, however, were impressive enough for his system not only to dominate men's minds for many centuries but to constitute a challenge even today. Men still, on occasions, think like Aristotle, and as long as that is so, Aristotelianism will remain a live metaphysical option.

THOMISM

Christian metaphysics

The advent of Christianity had important effects in philosophy as in other aspects of human life. Initially, Christians were opposed to philosophical claims of any kind; they saw philosophy as an essentially pagan phenomenon and refused to allow the propriety of subjecting Christian dogma to philosophical scrutiny. Christian truth rested on revelation and did not need any certificate of authenticity from mere reason. Later, however, attempts were made to produce a specifically Christian metaphysics, to think out a view of the universe and of man's place in it that did justice to the Christian revelation and nevertheless rested on arguments that might be expected to convince Christians and non-Christians alike. St. Thomas Aquinas was only one of a number of important thinkers in medieval times who produced Christian philosophies; others—such as the British philosophers John Duns Scotus in the late 13th century and William of Ockham in the first half of the 14th century—took significantly different views. In selecting the system of Aquinas for summary here, the factor that has weighed most has been its persistent influence, particularly in postmedieval times. Aquinas was not the only medieval philosopher of distinction, but Thomism is alive as other medieval systems are not.

The central claim of Thomism is that reflection on everyday things and the everyday world reveals it as pointing beyond itself to God as its sustaining cause. Ordinary existents, such as human beings, are in process of constant change. The change, however, is not normally the result of their own efforts, and even when it is, it does not depend on them exclusively. No object in the familiar world can fully account for its own *esse* (*i.e.*, its own act of existing), nor is it wholly self-sufficient; all are affected from without, or at least operate in an environment that is not of their own making. To say this is to say that they are one and all finite. Although finite things can be, and commonly are, stimulated to activity or kept in activity by other finite things, it does not follow that there might be finite things and nothing else. On the contrary, the finite necessarily points beyond itself to the infinite; the system of limited beings, each dependent for its activity on something else of the same kind, demands for its completion the existence of an unlimited being, one that is the source of change in other things but is not subject to change itself. Such a being would be not a cause like any other but a first or ultimate cause; it would be the unconditioned condition of the existence of all other things. Aquinas believed that human reason can produce definitive proofs of the existence of an infinite or perfect being, and he had no hesitation in identifying that being with the Christian God. Because, however, the movement of his thought was from finite to infinite, he claimed to possess only so much philosophical knowledge of the Creator as could be arrived at from study of his creation. Positive knowledge of the divine nature was not available; apart from revelation, man could only say what God is not, or conceive of his attributes by the imperfect method of analogy.

Aquinas worked out his ideas at a time when the philosophy of Aristotle was again becoming familiar in western Europe after a period of being largely forgotten, and many of his detailed theories show Aristotelian influence. He assumed the general truth of the Aristotelian picture of the natural world and the general correctness of Aristotle's way of interpreting natural phenomena. He also took over many of Aristotle's ideas in the fields of ethics and politics. He gave the latter, however, a distinctively different twist by making the final end of man not philosophical contemplation but the attainment of the beatific vision of God; it was Christian rather than Greek ideas that finally shaped his view of the *summum bonum* ("greatest good"). Similarly, his celebrated proofs of God's existence proceeded against a background that is obviously Aristotelian but that need not be presupposed for their central thought to have validity. Thomism can certainly be seen, and historically must be seen, as the system of Aristotle adapted to Christian purposes. It is important, however, to stress that the adaptation resulted in something new, a distinctive way of looking at the world that still has its adherents and still commands the respect of philosophers.

CARTESIANISM

Increased understanding of the physical world

René Descartes worked out his metaphysics at a time of rapid advance in human understanding of the physical world. He adopted from Galileo the view that physical things are not what they are commonly taken to be on the strength of sense experience—namely, possessors of "secondary" properties such as colour, smell, and feel—but are rather objects characterized only by the "primary" qualities of shape, size, mass, and mobility. To understand why a constituent of the physical world behaves as it does, what should be asked is where it is, how large it is, in what direction it is moving, and at what speed; once these questions are answered, its further properties will become intelligible. Descartes held further that all change and movement in the physical world is to be explained in purely mechanical terms. God was needed to give initial impetus to the physical system as a whole, but once it had got going it proceeded of its own accord. To pretend, as the Aristotelians had, to discern purposes in nature was to

make the impious claim to insight into God's mind. Descartes applied this theory to the movements of animals as much as to those of inanimate bodies; he thought of both as mere automatons, pushed and pulled about by forces over which they had no control.

Although Descartes thus acquiesced in, indeed emphasized, the mechanistic tendencies of contemporary science, he was far from being a Materialist. Besides material substance there was also thinking substance, and this was in fact wholly different from matter both in kind and in operation. Bodies had as their essence to occupy space; minds were not in space at all. Bodies, again, were determined in their movements; minds were in some sense free, because they possessed will as well as intelligence. Descartes was less explicit on this point than he might have been; the principles on which mental substance is supposed to operate are not made clear, with the result that critics have said that Descartes thought of mental activities in para-mechanical terms. Whether this is true or not, however, there was no reason for Descartes to be in any special difficulty over this point. All he needed to urge was that minds act in the strict sense of the term, which is to say that they take cognizance of their situation and respond more or less intelligently to it. That they can do this differentiates them fundamentally from material things, which are caused to do what they do and are entirely unaffected by rational considerations.

The main crux in Descartes's metaphysics was the difficulty of bringing together the two orders of being, once they were separated. Mention has already been made of the expedient to which later Cartesians were driven in trying to solve this difficulty: in effect, they made the unity of the universe a continuing miracle, dependent upon the grace of God. It is worth mentioning here another move in the same area that many have found instructive. Kant, who was in some respects both a latter-day Cartesian and a latter-day Platonist, argued that human activities could be looked at from two points of view. From the theoretical standpoint they were simply a set of happenings, brought about by antecedent events in precisely the same way as occurrences in the natural world. From the standpoint of the agent, however, they must be conceived as the product of rational decision, as proper acts for which the agent could be held responsible. The moment he began to act, a man transferred himself in thought from the phenomenal world of science to an intelligible world of pure spirit; he necessarily acted as if he were not determined by natural forces. The transference, however, was a transference in thought only (to claim any knowledge of the intelligible world was quite unjustified), and because of this the problem of the unity of the universe was dissolved. There was no contradiction in a man's thinking of himself both as a subject for science and as a free originator of action. Contradiction would appear only if he were present in both respects in an identical capacity. But appeal to the doctrine of the two standpoints was thought by Kant to rule this out.

Kantian development

It is only with some hesitation that one can speak of Kant as having put forward a metaphysics. He was in general highly suspicious of claims to metaphysical knowledge, and a principal aim of his philosophy was to expose the confusions into which professing metaphysicians had fallen. Nevertheless, it is clear that Kant had metaphysical convictions, for all his denial of the possibility of metaphysical knowledge; he was committed to the view that men can conceive a non-natural as well as a natural order and must necessarily take the former to be real when they act. The language he used—particularly his talk about man as phenomenon and man as noumenon—is not to the taste of present-day philosophers, but the thought behind it certainly survives. It is in this form, indeed, that Cartesianism may still be said to present a serious intellectual challenge.

IDEALISM

Descartes and Kant were both adherents of metaphysical dualism, though they worked out their dualisms in interestingly different ways. Many thinkers, however, find dualism unsatisfactory in itself; they look for a single principle by which to compass whatever exists. There are two broad steps that are open to the person who confronts a dualism of mind and matter and finds it unsatisfactory: he can either try to show that matter is in some sense reducible to mind, or conversely seek to reduce mind to matter. The first is the solution of Idealism, the second that of Materialism. Idealism has already been treated at length, and it will not be necessary to go into it again here. Only one point about it needs emphasis. As was pointed out, there are various forms of Idealism. In one version, this philosophy maintains that there literally is no such thing as matter; what the common man takes to be material things are, upon closer consideration, nothing but experiences in minds. Nothing exists but minds and their contents; an independently existing material world is strictly no more than an illusion. This was the view taken by Berkeley. In the more sophisticated Idealism of Hegel, however, it is not maintained that mind alone exists; material things are, in one way, taken to be as real as minds. The thesis advanced is rather that the universe must be seen as penetrated by mind, indeed as constituted by it. Spirit, to use Hegel's own word, is the fundamental reality, and everything that exists must accordingly be understood by reference to it, either as being directly explicable in spiritual terms or as prefiguring or pointing forward to spirit. Whatever the merits of this thesis, it is clear that it differs radically from that maintained by Berkeley. Idealism in the form espoused by Berkeley relies largely on arguments drawn from epistemology, though formally its conclusions are ontological, because they take the form of assertions or denials of existence. Hegel, however, had little or nothing to say about epistemology and was not even concerned to put forward an ontology. What he wanted to urge was a doctrine of first principles, a thesis about the terms in which to understand the world. The Hegelian "reduction" of matter to mind was thus reduction in a somewhat attenuated sense. It is important to get this point clear, if only because it has its parallel in the rival doctrine of Materialism.

Two solutions to the Cartesian dualism

MATERIALISM

The simplest form of Materialism is found in the claim that only matter exists. Stated thus baldly the claim is absurd, because it is clear that all sorts of things exist that are not of the nature of matter: thoughts and numbers and human institutions would be instances. In the light of these facts, the claim has to be revised to say that matter is the only substantial existent, with appeal being made to distinctions first worked out in Aristotle's doctrine of categories. According to this explanation, many things besides matter exist, but all of them are explicable (or so it is said) as modifications of matter. Thus, human institutions consist in patterns of movement among specific groups of human beings, and human beings in turn are nothing but highly complicated material bodies.

Matter as the only substantial existent

It is clear from these instances that Materialism is a controversial doctrine; it is also clear that its key word, modification, requires further explanation. When, for example, minds are said to be modifications of an underlying material substance, what is meant? A first and relatively easy point is that, like qualities and quantities, they could not exist separately. Unless there were material bodies, there could not be minds, because minds are—to put it crudely—states found in some material bodies. Minds are here equated with mentality, and mentality is clearly an abstraction. To say this, however, is not to remove the whole difficulty. When it is said that mentality is a state of some material body or bodies, is that meant literally or metaphorically? Bodies can often be described from the physical point of view as being in a certain state—for example, as being in a state of internal equilibrium. What is meant here is that the different particles of matter concerned stand in a certain relationship and as a consequence develop certain physical properties. But is mentality to be conceived as a physical property? It sounds extravagant to say so. Yet some such doctrine must be defended if Materialism is to be advanced as a form of ontology with a serious claim for attention. It is

interesting in this connection to notice the arguments advanced by scholars like J.J.C. Smart, which purport to identify states of mind with states of the brain. If the two are identical—literally the same thing described from two points of view—thoughts may really be modifications of matter, and Materialism may be tenable in a strong form. If, however, the identity cannot be made out—and very few philosophers are in fact ready to accept it—Materialism can be true at most in a modified form.

Naturalism

This modified form of Materialism is perhaps better described as naturalism. Naturalism holds not that all things consist of matter or its modifications but that whatever exists can be satisfactorily explained in natural terms. To explain something in natural terms is to explain it on scientific lines; naturalism is in fact a proclamation of the omnicompetence, or final competence, of science. It is not essential to this type of view to argue that phenomena can be spoken of in one way only; on this point, as on the point about ontological reducibility, the theory can afford to be liberal. It is, however, vital to make out that the scientific account of a set of happenings takes precedence over any other. Thus, the language in which men commonly speak of action and decision, which may be called for short the language of reasons, must be held to be secondary to the language in which scientists might speak of the same facts. Scientific language is basically causal, and the thesis of this form of Materialism is that causal explanations are fundamental. Naturalism is thus the obverse of Hegelianism; it is a theory of first principles, and it draws its principles from science.

If the question is raised why anyone should take this form of Materialism seriously, the answer lies in a number of significant facts. Physiologists have established correlations between general states of mind and general states of brain activity; their hope is to extend this to the point where particular thoughts and feelings can be shown to have their physiological counterpart. Cyberneticists have produced artifacts that exhibit mindlike behaviour to a remarkable degree; the inference that man is no more than a complicated machine is certainly strengthened by their achievements. Sociologists have shown that, whatever the explicit reasons men give for their beliefs, these are often intelligible in the light of factors of which they themselves take little or no account. The old assumption that human judgments are typically grounded in reason rather than merely caused, is called in question by the results of such investigations, which gain support from findings both in Freudian and in orthodox psychology. None of this evidence is decisive by itself; there are ways in every case of blocking the conclusions that Materialists tend to draw from it. Yet it remains true that, cumulatively, the evidence is impressive. It certainly has enough force to make it necessary to take this type of theory with the greatest seriousness. Metaphysical disputes in the modern world are fundamentally arguments for or against Materialism, and the other types of theory here explored are all seen as alternatives to this compelling, if often unwelcome, view.

IV. Argument, assertion, and method in metaphysics

Attention is now turned from description of the content of particular metaphysical views to more general treatment of the nature of metaphysical claims. The questions that will arise in this section concern such things as the nature and basis of metaphysical assertions, the character of metaphysical arguments and of what are taken to be metaphysical proofs, and the parts played in metaphysical thinking by insight and argument, respectively. They come together in the inquiry as to whether metaphysics can be said to be a science and, if so, what sort of a science it is.

METAPHYSICS AS A SCIENCE

Nature of an a priori science. Sciences are broadly of two kinds, a priori and empirical. In an a priori science such as geometry, a start is made from propositions that are generally taken to be true, and the procedure is to demonstrate with rigorous logic what follows if they are indeed true. It is not necessary that the primary premises of an a priori science should in fact be truths; for the purposes of the system they need only be taken as true, or postulated as such. The main interest is not so much in the premises as in their consequences, which the investigator has to set out in due order. The primary premises must, of course, be consistent one with another, and they may be chosen, as in fact happened with Euclidean geometry, because they are thought to have evident application in the real world. This second condition, however, need not be fulfilled; a science of this kind can be and commonly is entirely hypothetical. Its force consists in the demonstration that commitment to the premises necessitates commitment to the conclusions: the first cannot be true if the second are false.

Hypothetical nature of a priori sciences

This point about the hypothetical character of a priori sciences has not always been appreciated. In many classical discussions of the subject, the assumption was made that a system of this kind will start from as well as terminate in truths, and that necessity will attach to premises and conclusions alike. Aristotle and Descartes both spoke as if this must be the case. It is clear, however, that in this they were mistaken. The form of a typical argument in this field is as follows: (1) p is taken as true or given as true; (2) it is seen that if p, then q; (3) q is deduced as true, given the truth of p. There is no need here for p to be necessary or self-guaranteeing truth; p can be any proposition whatsoever, provided its truth is granted. The only necessity that needs to be present is that which characterizes the argument form, "If p is true, and p implies q, then q is true," that is $[p \cdot (p \supset q)] \supset q$, in which $\cdot$ symbolizes "and," and $\supset$ means "implies"; and this is a formula that belongs to logic. It is this fact that makes philosophers say, misleadingly, that a priori sciences are one and all analytic. They are not, because their premises need not answer this description. Nevertheless, they draw their lifeblood from analytic principles.

Metaphysics as an a priori science. It is clear that metaphysical philosophers have sometimes aspired to present their results in the form of a deductive system, to make metaphysics an a priori science. For this purpose they have taken a deductive system to require, not just that the premises entail the conclusions, but further that they themselves be necessarily true. Spinoza thus began the first book of his *Ethics* by laying down eight definitions and seven axioms whose truth he took to be self-evident, and then proceeded in the body of the text to deduce, as he thought with strict logic, 36 propositions that follow in order from them. He repeated the procedure in the rest of his work. That philosophical conclusions should thus be capable of being set out "in the geometrical manner" was something that Spinoza took as axiomatic; to be worthy of attention at all, philosophy must issue in knowledge as opposed to mere opinion, and knowledge proper had to be exempt from the possibility of doubt, which meant that it must either be intuitively evident or deducible from what was intuitively evident. Spinoza took this conception of knowledge from Descartes, who had himself toyed with the idea of presenting metaphysical arguments in the geometrical manner. Descartes, however, pointed out that, although there was no difficulty in getting agreement to the first principles of geometry, "nothing in metaphysics causes more trouble than the making the perception of its primary notions clear and distinct"; the whole trouble with this discipline is that its students fail to see that they must start from what are in fact the basic truths. Descartes himself spoke as if the problem were no more than pedagogical; it was a question of making people see as self-evident what is in itself self-evident. His own "analytic" approach in the *Meditationes* was chosen to overcome these difficulties; it was, he said, "the best and truest method of teaching." But it may well be that this account is too optimistic. The difficulty with a system such as those of Descartes and Spinoza is that there are persons who cannot be brought to see that the primary propositions of the system are self-evidently true, and this not because they are lacking in attention or insight but because they see the world in a different way. This suggests that in any such system there will necessari-

Deductive systems of Spinoza and Descartes

ly be an element that is arbitrary, or at least noncompulsive. However cogent the links that bind premises to conclusions, the premises themselves will lack a firm foundation. If they do, the interest of the system as a whole must be greatly diminished; it can be admired as an exercise in logic but not valued for more than that.

To avoid this unpalatable conclusion two expedients are possible. The first is to say that the first premises of a metaphysical system must be not merely self-evident but also self-guaranteeing; they must be such that any attempt to deny them can only result in their reaffirmation. Descartes believed that he could satisfy this requirement by grounding his system in the *cogito,* though strictly this was the primary truth only from the point of view of subjective exposition, and not according to the objective order of things. Aristotle somewhat similarly had argued that the logical principle of noncontradiction, which he took to express a highly general truth about the world, must be accepted as axiomatic on the ground that its correctness is presupposed in any argument directed against it.

Even the Idealists Bradley and Bernard Bosanquet at times spoke as if the first principles of their system were in some way logically compulsive; as Bosanquet put it, one had either to accept them or recognize that he could know nothing. Whatever the position may be about particular metaphysical propositions, however, it seems clear that not all truths that are taken as basic in metaphysics have the characteristic of being self-guaranteeing. A Materialist takes it as fundamental that whatever occurs happens as a result of the operation of natural causes; a theist sees things in the world as finite and thus as pointing beyond themselves to the infinite being who is their ground. No contradiction is involved in denying these positions, though of course for those who accept them the denial necessarily involves commitment to falsehood. It is, however, one thing for a proposition or set of propositions to be false, another altogether for it to be necessarily false. If the first principles of metaphysics were really self-guaranteeing, only one system of metaphysics could be coherent, and it would be true just because it was coherent. The very fact that there is an apparent choice between competing metaphysical systems, which may differ in plausibility but agree in being each internally self-consistent, rules this possibility out.

The alternative is to argue that fundamental metaphysical propositions, though not self-guaranteeing, are nevertheless not arbitrary; they have or, to be more cautious, can have a firm foundation in fact. Metaphysical speculation is not, as some opponents of metaphysics have suggested, essentially idle—*i.e.*, the mere working out of the logical consequences of premises that the metaphysician chooses to take as true. Or, rather, it does not necessarily answer this description, because a metaphysician can have insight into the true nature of things and can ground his system on that. This second position in fact involves arguing that metaphysics is not an a priori but an empirical science.

Reference to experience

Metaphysics as an empirical science. If metaphysics is an empirical science, the question of whether or not to accept a metaphysical theory must be answerable, in part at any rate, by reference to experience. It will not depend on experience alone, any more than does the acceptability of a scientific theory, because here, as in the scientific case, thinking comes into the reckoning too. A metaphysician can be mistaken in his deductions, just as a scientist can. But even if these are impeccable he will not necessarily succeed, on this view of his undertaking. It may be that he argues correctly from premises that are unacceptable—unacceptable because they lack the necessary foundation in fact. He will then be like a scientist who puts forward a hypothesis and deduces its consequences without mistake, only to find that experience fails to confirm the supposition on which he is working.

Scientific hypotheses are refuted, or at least called seriously into question, when predictions based on them fail to come true. As Karl Popper—who has emphasized that there is a unity of method in all generalizing or theoretical sciences—has insisted, every scientific hypothesis must be testable, and the way to test it is to look for circumstances in which it does not hold. To content oneself with favourable evidence is not enough; one must be searching all the time for unfavourable evidence. Further, it must be possible, if the hypothesis is genuinely scientific, to specify in advance what would count as unfavourable evidence; the circumstances in which the hypothesis needs to be abandoned, or at least modified, must be indicated precisely. In ideal conditions it is possible to devise a crucial experiment that will test a hypothesis definitively; the Michelson–Morley experiment, which disposed of the theory of the luminiferous ether, was such an experiment.

It can be asked, however, what parallels there are to this in metaphysics. The difficulty with testing a metaphysical thesis is twofold. First, metaphysical theories tend to be extremely general and as such highly unspecific. They announce, for example, that every event has some cause or other, or that every change is part of a process that serves some purpose. To find counterexamples to theses of such generality is on any account exceedingly difficult: how can one be sure that all the possibilities have been explored? There is, however, another and still more serious difficulty. The scientist, once he has laid down the conditions that would have to obtain for his hypothesis to prove false, makes no bones about their occurrence; it is, typically, a matter of whether or not a certain pointer reading is registered, and this is a simple question of ascertainable fact. Fact for the metaphysician, however, is altogether more slippery. Different metaphysicians see the world each in his separate way; what they take to be the case is coloured by their metaphysical conceptions. There is no neutral body of facts to which appeal can be made to show that a metaphysical theory falls down, and this being so, the attempt to assimilate metaphysics to science must fail.

That this should be the case is perhaps not surprising. Scientific thinking proceeds within a framework of presuppositions that it is the business of the scientist to use, not to argue for and still less to challenge—presuppositions to the effect, for example, that every change has a natural explanation. No doubt scientists can change their presuppositions, but they seldom do so consciously; their usual practice is to take them for granted. Metaphysicians, however, necessarily take a very different attitude toward presuppositions. It is their business to tell men how to understand the world, and this means that they must, among other things, put forward and argue for a set of interpretative principles. Metaphysicians differ radically in the interpretative principles they accept, and it is this that explains their failure to agree upon what to take as fact. It is naïve to suppose that the points at issue between, for example, a Thomist and a Materialist can be settled by observation or even by experiment; the facts to which one might appeal in support of his theory may be seen in a very different light by the other, or perhaps be dismissed as simple illusion. Reflection on the phenomenon of religious experience will illustrate what is meant here. That men undergoing this experience are affected mentally and physically in certain specific ways is perhaps common to both Thomist and Materialist. But the further description of their state is entirely controversial and owes its controversial character to the varying preconceptions that the disputants bring to their task.

INITIAL METAPHYSICAL INSIGHTS

Origin. If metaphysics is far from being a simple empirical discipline, however, it does not follow that it is wholly without foundation in fact. The true situation can perhaps be put as follows. Every metaphysic consists in an imaginative view of the world elaborated into a conceptual system. Metaphysics, like poetry, begins by being a matter of vision; a metaphysician sees the scheme of all things in a certain light; for example, as nothing more than a vast mechanism or as God's creation. As a metaphysician, however, he cannot be content to rest in a vision of this sort, as for example the Romantic poet William Wordsworth does in his "Intimations of Immortality." He needs to think out terms in which whatever

Metaphysics as a matter of vision

exists can be described so as to accord with his primary insight; he needs to produce and apply a conceptual system and to argue against possible alternatives. Whatever its origins, metaphysics is strictly intellectual in its development. It should not, however, be inferred from this that it is essentially idle speculation. For when the question is raised of the source from which metaphysicians gain their initial insights, the answer that occurs most readily is that they are derived from reflection on certain evident facts. Thus, the source of the Materialist view of the world is undoubtedly the practice of science; the Materialist proposes to give unrestricted validity to ways of thinking that scientists have found effective in a certain restricted sphere. The source of Idealist thought is to be found in the practice of history, or more generally in the interpersonal relations of beings who are at once rational and sensitive; the Idealist philosopher takes concepts that are appropriate in these limited areas to apply to the whole of reality. Every system of metaphysics is grounded in some real experience and owes its initial appeal to that fact. This is not to say, however, that the metaphysician builds on experience as does his scientific colleague. To think that is to take altogether too simple a view of the whole question.

Tests of validity. A question of immense importance is whether there are any means of comparing the validity of initial metaphysical insights. If it has to be answered negatively—if it has to be allowed that, as it were, all candidates in this field start and finish on an equal footing—the argument that each of them has a foundation in fact will be entirely discounted. Whatever respectability their concepts possess in their original homes will be lost once they fall into the hands of the metaphysician, because the procedure of the latter in taking them up and extending them is essentially arbitrary. For example, that one sees the sum of things as a vast machine may be suggested by what goes on in science, but this view can neither claim scientific warrant itself nor draw on scientific prestige, because it seems to spring from nothing better than mere whim. There are, however, two reasons for thinking that initial metaphysical insights are based not on mere whim but on valid grounds.

Insights as limited in number and subject to experience

First, the number of what may be called viable metaphysical insights is in practice limited: there are varying ways of taking the world as a whole, but not an infinite variety. In the outline account of metaphysical theories given above, six different kinds of view were distinguished, each of which may be said to be grounded in one or more areas of experience. It would be possible to extend the list, but probably not very far; further candidates might well turn out to be no more than variations on themes already considered. Thus, Leibniz might be seen as a latter-day Platonist, and Spinoza as offering a different version of the dualism of Descartes, one that is more sympathetic to Materialism than was Descartes himself. If these claims are true, they are certainly important; for the facts here adduced suggest that the experiences or visions on which different metaphysicians build are not peculiar to individual minds but occur commonly and regularly. They are not the product of passing moods, seized on and exploited for no good reason, but connect with thoughts that recur repeatedly in sensitive and intelligent reflection.

Second, there is a sense in which, despite everything said above, metaphysical theories are subject to the test of experience. That metaphysics aspires to give an account of the world as a whole means that each metaphysician claims that his fundamental insight illuminates every department of life. It may be that there are no neutral facts to which a metaphysician can appeal to show the shortcomings of his opponents; metaphysicians pronounce on what is to count as fact, and this puts them in the happy position of being judges in their own case. It remains true, however, that everyone who engages in that type of philosophy has the formal task of accounting for all the facts that he recognizes, and this is something that can be done more or less well. The value of different metaphysical insights is sometimes shown in the success with which they are applied. Furthermore, it is not quite true that the metaphysician need consult no opinion but his own when it comes to working out his views. What might be called public opinion has a part to play as well, though it has no absolute right to a hearing. A metaphysician who chooses to dismiss areas of experience or ways of thinking that are commonly accepted as being in order does so at his peril; he reduces the initial plausibility of his own theories the oftener he finds himself in this position. He could, of course, be right and common opinion wrong; no genuine metaphysician is put off by the thought of such a conflict. Though he is not put off, however, he has to be wary all the same. He may be able to say what in the end is to count as fact, but if this involves him in dismissing as illusory what instructed opinion generally takes to be real, his triumph may be hollow. Whether he likes it or not, he has to frame a theory that will carry conviction with experts in the different fields concerned, or, if that is going too far, one that will strike them as not wholly implausible. A metaphysician who exercises his veto past that point is simply failing to do his job.

The test of self-consistency of theory

It must be admitted that the tests one can apply to determine the value of a metaphysical theory are at best unsatisfactory. Often one is driven back onto the expedient of asking if the theory is internally self-consistent; a surprisingly large number of philosophical theories are not. To confute a philosopher out of his own mouth is, perhaps, the most effective form of confutation. If this expedient will not apply, however, the questioner is not quite helpless. Whatever the explanation, it is a well-known fact that a philosopher can purchase consistency at the expense of plausibility; he can put forward theories that evade difficulties by simply declaring them nonexistent. In so doing, he turns his back on what instructed opinion generally takes to be fact. His hope is, of course, to persuade others to see the situation as he does, and there is always the possibility that he will succeed. If, however, after a suitable interval he has not, that must surely count against him. It is by this test that one decides, for example, that the metaphysics of Hobbes is not worth prolonged study, despite the enormous ingenuity of its author; there is too much in this system that seems to be sheerly arbitrary. The same comment could be made of certain forms of Idealism, which are so intent on the omnipresence of spirit that they neglect the materiality of the material order. Admittedly, the test is harder to apply when attention is transferred to the major theories in their most persuasive form, because here the question concerns views that have stood the test of time. It is not, however, entirely inapplicable even there. An individual, at least, may feel that this or that view will not do precisely because it achieves comprehensiveness by turning its back on fact; and, though it is unsatisfactory to fall back on personal judgment in this way, there is perhaps no other alternative in this difficult area.

Role of personal or social factors. Some writers on the philosophy of philosophy, such as Dilthey, have suggested that the persistence of a plurality of metaphysical systems is to be explained in terms of personal or social factors. Certain kinds of metaphysical outlook appeal to certain types of human being, or gain currency in social circumstances of this kind or that; to understand why they are accepted, recourse must be had to psychology or sociology or both. In the above account, stress has been laid on the historical background against which a number of famous metaphysical theories got their classical formulations; it is idle to deny that each was originally designed to solve a problem deemed to be urgent at the time. Nevertheless, the problem was, of course, an intellectual problem, and the solution offered claimed to be true, not simply comforting. No doubt wishful thinking is as rife in the field of metaphysics as anywhere; it is all too easy here to confuse what men ought to believe with what they want to believe. Philosophies reveal something about their authors and even about their historical age, as works of literature do; they constitute historical evidence as books on mathematics, perhaps, do not. Yet all this can be admitted without agreeing that metaphysics is merely of psychological or historical importance. Science does not cease to be true because it is shown to be useful.

Nor is it true that metaphysical theories always in fact give comfort; there are cases in which men find themselves returning over and over again to possibilities that they would very much like to believe were not realized. A philosopher can commit himself to a view of the world that is not at all to his taste, simply because it seems to him on due consideration that this is how things are. That philosophers are godlike beings able to rise entirely above the limitations of their age seems unlikely. It is equally unlikely, however, that their opinions are determined throughout by nonrational factors, and thus that their thinking can lay no claim to truth.

METAPHYSICAL ARGUMENTS

Classes of metaphysical statements

Logical character of metaphysical statements. Metaphysical statements fall into two main classes: statements about what exists and prescriptions about how to take or understand what exists. It might seem obvious that the first is the more important; the metaphysician first lays down what he takes to exist, and then tells how to interpret it. This would be correct if metaphysics were a departmental inquiry like, for example, botany; but, of course, it is not. Metaphysicians possess no special resources for the detection of unfamiliar entities, and in consequence the realities they accept must all be argued for. The fundamental items that fill the metaphysical world are one and all theoretical; they are not so much palpable realities as artificial constructs. That being so, there is less of a gulf between the two types of metaphysical pronouncement than might at first appear. It could indeed be argued that the two go closely together to constitute what may be called a metaphysical point of view, a standpoint whose primary purpose is to provide understanding. In a metaphysical context, to say what exists is itself a step on the way to understanding; it is not something that antedates theory, but part of a theory itself.

It may be asked whether metaphysical pronouncements are empirical or a priori and, if the latter, whether they are analytic or synthetic. They are certainly not straightforwardly empirical, for reasons just set out, and cannot be merely analytic (*i.e.*, true in virtue of the definitions of their terms and of the laws of logic) if metaphysics is to retain any significance. The conclusion that they must be synthetic a priori (*i.e.*, such that, unlike analytic propositions, they convey new knowledge and yet claim complete universality and necessity) seems to follow, and it is just what the opponent of metaphysics wants the metaphysician to adopt. Metaphysics, as he sees it, is a wholly unwarranted attempt to say what the world must be like on the strength of pure thinking, an attempt that is doomed to failure from the start. Before this condemnation is accepted, however, the function that the metaphysician assigns to his principles should be considered. When this is done, it becomes plain that the charge that he claims factual knowledge of a nonempirical sort is false; in one way he recognizes exactly the same facts as anyone else. Where he claims superiority is in knowing how to take facts, and the burden of his message consists in the advocacy of principles that, he alleges, will provide overall understanding. One can describe these principles as synthetic a priori if one chooses. It is probably best, however, to avoid this misleading term and simply say that they are thought of by the metaphysician as applying unequivocally to whatever falls within experience. These metaphysical principles are instructive at least in the sense of having alternatives, and they are certainly treated as being necessary. It is not true, however, that they take the form of statements of fact, even highly general statements of fact; nor is their necessity the same as that which characterizes logical truths. The principles are prescriptions rather than statements, and their necessity arises from the role they play in the constitution of experiential knowledge. It is a necessity that is in one way absolute: nothing that can claim to be real can escape their jurisdiction, because they tell how to take whatever occurs. Nevertheless, in another way the necessity of the principles is merely conditional, for other ways of interpreting the same data can be conceived, and it is admitted that there are circumstances, however hard to specify exactly, in which it would have to be agreed that they do not apply.

Deductive and inductive arguments

Logical form of metaphysical arguments. There is also the question whether metaphysical arguments are inductive or deductive or whether they have some logical form peculiar to themselves. It is obvious that much metaphysical reasoning is, or purports to be, reasoning in the strict sense, which is to say that its form is deductive. Arguments like the first cause argument for God's existence claim to be demonstrations; their exponents believe that anyone who commits himself to the truth of the premises stands logically committed to the truth of the conclusions. This claim can stand, even if it turns out that the project to set out metaphysical results in the geometrical manner is a mistake. It may be impossible to model metaphysics on mathematics, but that does not make particular metaphysical arguments any less deductive.

As regards inductive arguments, it would be odd to find a metaphysician contending, as, for example, historians regularly do, that p is true and q is true and therefore it is reasonable to conclude that r is true. To assess probabilities in the light of established facts is too cautious for the average metaphysical mind. Yet it would be wrong to deny that metaphysicians are preoccupied with facts. Their objective is to give a reasoned account of what exists or obtains, and for this purpose attention to fact is of course indispensable. It figures in metaphysical thinking at two stages. First, at the beginning, when the metaphysician is concerned to formulate his main thesis; here there is a move from what holds in a restricted sphere (the sphere of physics, for example) to what is supposed to hold generally, a move that is possible only if the theorist concerned has an interest in the sphere in question. To arrive at his own position the metaphysician must extrapolate from what goes on outside metaphysics, and this means that he must be sensitive to significant developments in at least some of the main fields of learning and areas of practical activity. But he needs this extra-philosophical knowledge for a second purpose too: in estimating the success of his own theories. In principle he must show that his interpretation of experience covers the facts in an adequate way, and for this purpose what experts in the different spheres take to be established is of crucial importance. Metaphysics is not an empirical science—the element of speculation it includes is too strong for that—but the metaphysician can no more ride roughshod over facts than the scientist can. At the least he must explain away phenomena that seem to count against his thesis, or indicate how they might be explained away. Whether he explains or explains away, he needs to know what the main phenomena are.

Transcendental argument

It is sometimes said that metaphysics can make use of a form of argument that is neither deductive nor inductive but transcendental; a transcendental argument is supposed to proceed from a fact to its sole possible condition. The interest of these arguments as they appear in philosophy—in the philosophy of Kant, for example—lies in the kinds of things that are taken as starting points; in Kant's case it was such things as the possibility of pure mathematical knowledge, the possibility of making objectively true statements, the fact that there is a unitary system of time. Kant purported to prove a number of surprising propositions by the use of transcendental arguments; what he had to say under this head has attracted particular interest in recent years. It seems clear, however, that from the logical point of view no special significance attaches to this form of argument. A transcendental argument is simply a form of deduction, with the typical pattern: only if p then q; q is true; therefore, p is true. The interest of the argument, and its difficulty, reside not in the movement from premises to conclusion, which is absolutely routine, but in the setting up of the major premise. Kant tried to commend his major premises (for example, in his arguments about causality and substance) by showing what would result if the protasis (*i.e.*, p) did not hold. Though in this way he showed that a sufficient is also a necessary condition, he did not make clear why it should be taken as the

sole such condition. There is an important gap in his reasoning here, as there is in other metaphysical writers.

V. Criticisms of metaphysics

Metaphysics has many detractors. The man who aspires "to know reality as against mere appearance," to use Bradley's description, is commonly taken to be a dreamer, a dupe, or a charlatan. Reality in this context is, by the metaphysician's own admission, something that is inaccessible to sense; as Plato explained, it can be discovered only by the pure intelligence, and only if the latter can shake itself free of bodily encumbrances. The inference that the metaphysical world is secret and mysterious is natural enough. Metaphysics in this view unlocks the mysteries and lets the ordinary man into the secrets. It is, not to put too fine a point on it, a study of the occult.

Metaphysics as a study of the occult

METAPHYSICS AS KNOWLEDGE OF THE SUPERSENSIBLE

That there are aspects of metaphysics that lend colour to this caricature can scarcely be denied. The language of Plato, in particular, suggests an absolute distinction between the deceitful world of appearances, which can never be an object of knowledge, and the unseen world of Forms, each of which is precisely what it appears to be. Plato urged his readers not to take seriously the things of sense; he told them that everything having to do with the senses, including the natural appetites and the life of the body, is unreal and unimportant. The philosopher, in his view, needs to live an ascetic life, the chief object of which is to cultivate his soul. Only if he does this, and follows a rigorous intellectual training, has he any hope of getting the eye of his soul fixed on true reality and so of understanding why things are what they are.

Yet even this program admits of an innocuous, or relatively innocuous, interpretation. The "dialectician," as Plato called his metaphysical philosopher, is said in one place to be concerned to "give an account," and the only things of which he can give an account are phenomena. Plato's interest, despite first appearances, was not in the unseen for its own sake; he proposed to go behind things visible in order to explain them. He was not so much disdainful of facts as critical of accepted opinions; his attack on the acquiescence in "appearances" was an attack on conventional wisdom. That this was so comes out nowhere more clearly than in the fact that his targets included not just beliefs about what there is but also beliefs about what is good. It is the opinions of the many that need correction, and that can get it only if men penetrate behind appearances and lay hold on reality.

Plato is often presented as an enemy of science, on the ground that he was bitterly opposed to Empiricism and because he said that, if there was ever to be progress in astronomy, the actual appearances of the starry heavens must be disregarded. He understood by Empiricism, however, the uncritical acceptance of apparent facts, with the attempt to trace regularities in them; it is an attitude that, in his view, is marked by the absence of thought. As for the starry heavens, it is certainly difficult to take Plato quite literally when he compares their function in astronomy to that of a well-drawn diagram in geometry. Yet he was not wrong to suggest that no progress could be made in astronomical inquiries until appearances were seen to be what they were and not taken for absolute realities. The subsequent progress of astronomy has shown this view to be entirely correct.

There are respects in which Plato's attitude to phenomena was precisely the same as that of the modern scientist. Nevertheless, the fact remains that he believed in a realm of unseen realities and is of course far from being the only metaphysician to do so. Many, if not quite all, metaphysicians are committed to claiming knowledge of the supersensible, in some degree at least; even Materialists are alleged to make this claim, when they say that behind the familiar world of everyday experience there lies material substance that is not accessible to the senses. It has been a commonplace among critics of metaphysics since the early 18th century that no such claims can be justified; the supersensible cannot be known about, or even known of, whether directly or by inference.

SPECIFIC CRITICISMS

Hume. An early but powerful statement of these criticisms is to be found in the writings of David Hume, *A Treatise of Human Nature* (1739–40) and *An Enquiry Concerning Human Understanding* (1748). Hume argued first that every simple idea was derived from some simple impression, and that every complex idea was made up of simple ideas; innate ideas, supposed to be native to the mind, were nonexistent. There were eccentricities in Hume's conception of idea (and for that matter in his conception of impression), but these did not destroy the force of his argument that the senses provide the materials from which basic concepts are abstracted. A being that lacked sense experience could not have concepts in the normal sense of the term. Next, Hume proceeded to make a sharp distinction between two types of proposition, one knowable by the pure intellect, the other dependent on the occurrence of sense experiences. Propositions concerning matters of fact and existence answer the latter description; they either record what is immediately experienced through the senses or state what is taken to be the case on the basis of such immediate experiences. Such statements about matters of fact and existence are one and all contingent; their contradictories might have been true, though as a matter of fact they are not. By contrast, propositions of Hume's other type, which concern relations of ideas, are one and all necessary; reflection on the concepts they contain is enough to show that they must, in logic, be true. Though, in a sense, knowledge of these propositions is arrived at by the exercise of pure reason, no real significance attaches to this fact. It is not the case of some special insight into the nature of things; the truth is rather that these propositions simply make explicit what is implicit in the definitions of the terms they contain. They are thus what Kant was to call analytic propositions, and it is an important part of Hume's case that the only truths to which pure reason can attain are truths of this nature.

Necessity of sense experience

Finally, Hume sought to block the argument that, even if the supersensible could not be known directly, or through pure intellectual concepts, its characteristics could nevertheless be inferred. His analysis of causality had this as one of its aims. According to Hume, the only means by which men can go beyond the impressions of the memory and the senses and know what lies outside their immediate experience is by employing causal reasoning. Examination of the causal relation shows that it is, however, among other things, always a relation of types of events in time, one of which invariably precedes the other. Causality is not, as Descartes and others supposed, an intelligible relation involving an internal tie between cause and effect; it is a matter of purely factual connection and reduces on its objective side to nothing more than regular precedence and succession. The importance of this for the present inquiry lies in the consequence that causal relations can hold only between items, or possible items, of experience. According to Hume, if the temporal element is removed from causality, nothing concrete is left; if it is kept, it becomes impossible to argue that one can proceed by causal reasoning from the sensible to the supersensible. Yet it was precisely this that Aristotle, Aquinas, and Locke had all attempted.

Hume's own explicit pronouncements about metaphysics are ambivalent. There is a famous passage in which he urged men to consign volumes of divinity and "school metaphysics" to the flames, "as containing nothing but sophistry and illusion," but in at least one other place he spoke of the need to "cultivate true metaphysics with some care, in order to destroy the false and adulterate." "True metaphysics," in this connection, meant critical philosophical reflection.

Critical philosophical reflection

Kant. In Hume's successor Kant, there was a sharper distinction between metaphysics and critical philosophy. Much of Kant's philosophical effort was devoted to arguing that metaphysics, understood as knowledge of things supersensible, is an impossibility. Yet metaphysics, as a study of the presuppositions of experience, could be put on "the sure path of science"; it was also possible, and indeed necessary, to hold certain beliefs about God, free-

dom, and immortality. But however well-founded these beliefs might be, they in no sense amounted to knowledge: to know about the intelligible world was entirely beyond human capacity. Kant employed substantially the same arguments as had Hume in seeking to demonstrate this conclusion, but introduced interesting variations of his own. One point in his case that is especially important is his distinction between sensibility as a faculty of "intuitions" and understanding as a faculty of concepts. According to Kant, knowledge demanded both that there be acquaintance with particulars and that these be brought under general descriptions. Acquaintance with particulars was always a matter of the exercise of the senses; only the senses could supply intuitions. Nevertheless, intuitions without concepts were blind; one could make nothing of particulars unless he could say what they were, and this involved the exercise of a very different faculty, the understanding. Equally, however, the concepts of the understanding were empty when considered in themselves; they were mere forms waiting to be brought to bear on particulars. Kant emphasized that this result held even for what he called "pure" concepts such as cause and substance; the fact that these had a different role in the search for knowledge from the concepts discovered in experience did not give them any intuitive content. In their case, as in that of all other concepts, there could be no valid inference from universal to particulars; to know what particulars there were in the world, it was necessary to do something other than think. Thus is revealed the futility of trying to say what there is on the basis of pure reason alone.

Differences between Hume's and Kant's criticisms

Kant's distinction between analytic and synthetic propositions has peculiarities of its own, but for present purposes it may be treated as substantially identical with Hume's distinction set out above. Similarly, the important differences between Kant and Hume about causality may be ignored, seeing that they agreed on the central point that the concept can be properly applied only within possible experience. If it is asked whether there are substantial differences between the two as critics of metaphysics, the answer must be that there are, but that these turn more on temperament and attitude than on explicit doctrine. Hume was more of a genuine iconoclast; he was ready to set aside old beliefs without regret. For Kant, however, the siren song of metaphysics had not lost its charm, despite the harsh words he sometimes permitted himself on the subject. Kant approached philosophy as a strong believer in the powers of reason; he never abandoned his conviction that some of man's concepts are a priori, and he argued at length that the idea of the unconditioned, though lacking constitutive force, had an all-important part to play in regulating the operations of the understanding. His distinction between phenomena and noumena, objects of the senses and objects of the intelligence, is in theory a matter of conceptual possibilities only; he said that, just as one comes to think of things sensible as phenomena, so one can form the idea of a world that is not the object of any kind of sense experience. It seems clear, however, that he went beyond this in his private thinking; the noumenal realm, so far from being a bare possibility invoked as a contrast with the realm that is actually known, was there thought of as a genuine reality that had its effects in the sense world, in the shape of moral scruples and feelings. A comparison of what was said in Kant's early essay *Träume eines Geistersehers erläutert durch Träume der Metaphysik* (1766; *Dreams of a Spirit-Seer*), with the arguments developed in the last part of his *Grundlegung zur Metaphysik der Sitten* (1785; *Fundamental Principles of the Metaphysic of Morals*, 1949), would seem to put this judgment beyond serious doubt.

Though Kant remained convinced of the existence of things supersensible, he nonetheless maintained throughout his critical writings that there can be no knowledge of them. There can be no science of metaphysics because, to be true to fact, thinking must be grounded in acquaintance with particulars, and the only particulars with which human beings are acquainted are those given in sense. Nor was this all. Attempts to construct metaphysical systems were constantly being made; philosophers repeatedly offered arguments to show that there must be a first cause, that the world must consist of simple parts, that it must have a limit in space, and so on. Kant thought that all such attempts could be ruled out of court once and for all by the simple expedient of showing that for every such proof there was an equally plausible counterproof; each metaphysical thesis, at least in the sphere of cosmology —*i.e.*, the branch of metaphysics that deals with the universe as an orderly system—could be matched with a precise antithesis whose grounds seemed just as secure, thus giving rise to a condition that he called "the antinomy of pure reason." Kant said of this antinomy that "nature itself seems to have arranged it to make reason stop short in its bold pretensions and to compel it to self-examination." Admittedly, the self-examination led to more than one result: it showed on the one hand that there could be no knowledge of the unconditioned, and demonstrated on the other that the familiar world of things in space and time is a mere phenomenon, thus opening up the way to a doctrine of moral belief. Though this doctrine could not be expunged from Kant's philosophy without destroying it altogether, it is quite wrong to present it, as some modern German writers do, as amounting to the advocacy of an alternative metaphysics. What Kant was concerned with here is what must be thought, not what can be known.

Logical Positivists. Despite what has just been said, it must be admitted that Kant's constant talk about the supersensible makes many critics of metaphysics regard him as a dubious ally. This was certainly true in the case of the Logical Positivists, the philosophical school that has attacked metaphysical speculation most sharply in the 20th century. The Positivists derived their name from the "positive" philosophy of Auguste Comte, a 19th-century Frenchman who had represented metaphysical thought as a necessary, but now superseded, stage in the progression of the human mind from primitive superstition to modern science. Like Comte, the Logical Positivists thought of themselves as advocates of the cause of science; unlike Comte, they took up an attitude toward metaphysics that was uniformly hostile. The external reason for this was to be found in the philosophical atmosphere in the German-speaking world in the years following World War I, an atmosphere that seemed to a group of thinkers known as the Vienna Circle to favour obscurantism and impede rational thought. But there were, of course, internal reasons as well.

According to the Positivists, meaningful statements can be divided into two kinds, those that are analytically true or false, and those that express or purport to express matters of material fact. The propositions of logic and mathematics exemplify the first class, those of history and the natural and social sciences the second. To decide whether a sentence that purports to state a fact is meaningful, one must ask what would count for or against its truth; if the answer is "nothing," it cannot have meaning, or at least not in that way. Thus, they adopted the slogan that the meaning of a (nonanalytic) statement is the method of its verification. It was this verification principle that the Positivists used as their main weapon in their attacks on metaphysics. Taking as their examples statements from actual metaphysical texts—statements such as "The Absolute has no history" and "God exists"—they asked first if they were supposed to be analytically or synthetically true, and then, after dismissing the first alternative, asked what could be adduced as evidence in their favour or against them. Many metaphysicians, of course, claimed that there was empirical support for their speculative conclusions; thus, as even Hume said, "the order of the universe proves an omnipotent mind." The very same writers, however, proved strangely reluctant to withdraw their claims in the face of unfavourable evidence; they behaved as if no fact of any kind could count against their contentions. It followed, said the Positivists, that the theses in which they were interested were compatible with any facts whatsoever and thus were entirely lacking in significance. An analytic proposition, such as "It either will or will not rain tomorrow," tells nothing,

The verification principle

though there may be a point in giving voice to it. A metaphysical proposition claims to be very different; it purports to reveal an all-important truth about the world. But it is no more informative than a bare tautology, and if there is a point in putting it forward, it has to do with the emotions rather than the understanding.

In point of fact, the Positivists experienced great difficulty in devising a satisfactory formulation of their verification principle, to say nothing of a satisfactory account of the principle's own status. In the early days of the movement the demand for verifiability was interpreted strictly: only what could be conclusively verified could be significant. This had the effect of showing that statements about the past and propositions of unrestricted generality, to take only two instances, must be without meaning. Later a move was made toward understanding verifiability in a weak sense: a statement was meaningful if any observations bore on its truth. According to A.J. Ayer, an English disciple of the Vienna Circle, writing in 1936,

> It is the mark of a genuine factual proposition, not that it should be equivalent to an experiential proposition, or any finite number of experiential propositions, but simply that some experiential propositions can be deduced from it in conjunction with certain other premises without being deducible from those other premises alone.

As Ayer admitted in his second edition, however, this formulation lets in too much, including the propositions of metaphysics. From "The Absolute has no history" and "If the Absolute has no history, this is red," it follows that "This is red," which is certainly an experiential proposition. Nor were subsequent attempts, by Ayer and others, to tighten up the formulation generally accepted as successful, for in every case it was possible to produce objections of a more or less persuasive kind.

This result may seem paradoxical, for at first glance the Positivist case is extremely impressive. It certainly sounds odd to say that metaphysical sentences are literally without meaning, seeing that, for example, they can be replaced by equivalent sentences in the same or another language. But if the term meaning is taken here in a broad sense and understood to cover significance generally, the contention is by no means implausible. What is now being said is that metaphysical systems have internal meaning only; the terms of which they consist may be interdefinable, but perhaps do not relate to anything outside the system. If that were so, metaphysics would in a way make sense, but for all that would be essentially idle; it would be a game that might amuse but could hardly instruct. The Positivists confront the metaphysician with the task of showing that this criticism is not correct. Whatever difficulties are involved in formulating a principle of verifiability, the challenge can hardly be ignored.

Moore and Wittgenstein. The Positivists were not the only modern critics of metaphysics. G.E. Moore never argued against metaphysics as such, but nevertheless he produced criticisms of particular metaphysical theses that, if accepted, would make metaphysical speculation difficult, if not impossible. It was characteristic of a certain type of philosopher, according to Moore, to advance claims of a highly paradoxical nature—to say, for instance, that "Time is not real" or that "There are no such things as physical objects." Moore's case for rejecting such claims was that they go against the most central convictions of common sense, convictions that people accept unhesitatingly when they are not doing philosophy. Men constantly say that they did this before that, that things are better or worse than they were; from time to time they put off things until later or remark that tomorrow will be another day. Moore took these facts as definitive proof of the reality of time and definitive disproof of any metaphysical theory that denied it. Supporters of Bradley, the philosopher here criticized, replied that Moore had missed the point. Bradley never denied the truth of temporal propositions as used in the description of appearances; what he questioned was the coherence and ultimate tenability of the whole temporal way of thinking. As Rudolf Carnap, a Logical Positivist, was to put it, he raised an external question and was given an internal answer by Moore. It was an answer, however, that carried considerable conviction. The simple denial of what seem to be obvious facts had always been part of the stock in trade of metaphysicians; they make much of the distinction between appearance and reality. Moore may not have demonstrated the impropriety of this insistence, but at least he made it necessary for the metaphysician to be more circumspect, to explain explicitly what he was denying and what he was ready to accept, and so to make his own case sharper and thus easier to confirm or reject.

Common-sense criticisms

Moore's implied criticisms of metaphysics lead on naturally to those of Wittgenstein. Moore took his stand on common sense, whereas Wittgenstein based his on living language. Arguing that men are each involved in a multitude of language games or autonomous linguistic activities, insofar as they are scientific investigators, moral agents, litigants, religious worshippers, and so on, Wittgenstein asked in what language game the claims and questionings of philosophers arose. He replied that there was no genuine linguistic context to which they belonged; philosophical puzzlement was essentially idle. Philosophers were preoccupied with highly general questions; they aspired to solve *the* problem of meaning or *the* problem of reality. Against that, Wittgenstein argued that words and sentences have meaning as used in particular contexts; there is no single set of conditions that have to be fulfilled if they are to be thought meaningful. Equally, there is no single set of criteria that have to be satisfied by everything one takes to be real. Sticks and stones and men are taken as real in everyday discourse, but so are numbers in the discourse of mathematicians, and so is God in the discourse of religious men. There is simply no warrant for preferring one of these above the others—for saying, for example, with persons of an Empiricist turn of mind, that nothing can be real that does not have existence in space and time.

Language criticisms

Wittgenstein's antipathy to metaphysical philosophy was in part based on self-criticism; in his early work the *Tractatus Logico-Philosophicus,* published in 1922, he had himself tried to give a general account of meaning. At least one doctrine of that enigmatic book survived in his later thought: the distinction between saying and showing. Wittgenstein in the *Tractatus* sought to pronounce on "what can be said" and came to the conclusion that only "propositions of natural science" can be. Though at this stage he spoke as if metaphysical statements were senseless, his motives for doing so were very different from those of the Positivists. The latter saw metaphysics as an enemy of science; in their view, there was only one way to understand the world, and that was in scientific terms. But Wittgenstein, though agreeing that science alone can be clear, held that scientific thought has its limitations. There are things that cannot be said but can nonetheless be shown; the sphere of the mystical is perhaps a case in point. Unlike his Viennese contemporaries, Wittgenstein had no wish to rule out of court the thought that there are more things in heaven and earth than can be compassed in the language of science; writers whom he admired—such as Blaise Pascal, a 17th-century French scientist and writer on religious subjects, and Søren Kierkegaard, a Danish philosopher and theologian who is regarded as the founder of modern Existentialism —had discoursed of such matters in a way that was highly illuminating. They had made clear, however, that, just as one here went beyond the province of science, so also one went beyond that of philosophy. For them the idea that the metaphysician is privy to the most important of all things is absurd. There may be a sense in which men transcend everyday experience in moments of religious feeling or artistic insight, but there is no justification for thinking that when they do they arrive at the metaphysician's Absolute. As Kierkegaard said, the man who looks for speculative proofs in the sphere of religion shows that he does not understand that sphere at all.

Religious philosophers. It is important, in considering current criticisms of metaphysics, to appreciate that this discipline is now under double attack. In the first place, it must face the assault of those who regard it as a rival to science; it is against this assault that sympathizers like

R.G. Collingwood, a British philosopher, historian, and archaeologist, seek to defend it. But metaphysics is also in disfavour among many religious philosophers. In earlier days, partisans of religion, and more generally believers in a spiritual order, looked to metaphysics to vindicate their claims against skeptical attack; now they are altogether more reluctant to do so. The continuing controversy about metaphysics has no doubt influenced this result; it scarcely seems sensible to take refuge in a fortress whose walls are so frequently breached. There is, however, another motive that operates here: the feeling that metaphysics is not only dubious but, worse, unnecessary. In an age whose tendencies are antiphilosophical rather than philosophical, there is widespread acceptance of the view that religion and morals, and for that matter science and history, are their own justification; none of them stands in need of a certificate of respectability from philosophy, and any pretense by metaphysicians to supply or refuse such a certificate must be without foundation. Though this view is widespread, it is even so not unchallenged; there are persons who find the fragmentation it involves—belief in God on Sundays, belief in science for the rest of the week—intolerable. For such persons, at least, the search for metaphysical truth and metaphysical answers must retain its fascination. (W.H.W.)

VI. Tendencies in contemporary metaphysics

Background: the thought of Kant and Hegel

Kant's efforts to limit metaphysics opened new lines for its development. He had thought that reason is established by being limited and that some truths are certain independent of anything that can happen in experience because experience is structured by the interpretive categories reflected in the truths. Thus, it is possible to be certain of the world in its general structure, but only insofar as it is an experienced or phenomenal world—*i.e.*, a world known by man, not a world as it is in itself. Hegel, however, argued persistently that knowledge of a thing unknowable in itself is a contradiction and that reason can know all that is real if the mind first accepts the given thing as "always already within experience as other." The mutual implication of knowing mind and reality known is accepted, and a science of self-consciousness that relates all categories and all reality to the knowing subject is envisaged. Thus, Kant's mutual implication of knowing subject and phenomenal thing was given ultimate metaphysical validity by Hegel, and Kant's reformulations of traditional dualisms—*e.g.*, subject–object, appearance–reality, perceptual–categorial, immanent–transcendent, regulative–constitutive—became momentous for metaphysics.

TENDENCIES IN THE UNITED STATES

John Dewey. In this milieu, John Dewey, an American educational reformer and pragmatic philosopher, published his "Kant and Philosophic Method" in 1884 in the journal of a group known as the St. Louis Hegelians. Although Dewey later rejected the full-scale Hegelianism expressed in the article, he did so only after gathering up in a partial synthesis the thought of both Kant and Hegel. In this he sounded the thematic notes of much contemporary American and continental metaphysics. Whether or not this metaphysics is explicitly termed transcendental (*i.e.*, concerned with experience as determined by the mind's conceptual and categorial makeup), it does two things: (1) it affirms Kant's insight that physical particulars cannot first be identified and later interrelated by means of the categories, but to be identified at all they must be assumed to be already categorized, and reasoning must proceed to expose those categorial structures that make the actuality of knowledge possible; (2) it agrees with Hegel's critique at least to the extent that Kant's idea that the source of sensations is external to the mind in a noumenon is regarded as a transgression of Kant's own doctrine that the categories, particularly that of causation, can be applied only within phenomenal experience. Dewey thought that Kant confused the empirical and transcendental standpoints by mixing analysis of the organism as sensationally responsive with analysis of mind. Kant forgot that it is only because the knowing subject already grasps the world through its categories that it can self-deceivingly regard its sensations as subjective and as caused by something not known. Thus, for Dewey, "The relation between subject and object is not an external one; it is one in a higher unity that is itself constituted by this relation."

In Dewey's extended later thought, metaphysics became the study of "the generic traits of existence." Concern with God and immortality slips nearly from view, and this is typical of much contemporary philosophy. Even so, Dewey's rethinking of the subject–object relation engenders a concept of a democratic and scientific community of persons, bound to each other through common ideals, which has religious overtones. Vague and ambivalent as this concept may be, it helps undermine the whole contrast between immanent and transcendent, and leads metaphysics on new paths.

William James. The work of William James, a leader of the Pragmatic movement, was typical of many contemporary tendencies, one of which is the attempt to locate the role of science in knowledge and culture. Trained in medicine, James hoped to protect the autonomy of psychology as a science by adopting a dualistic view of mind and matter: He "supposes two elements, mind knowing and thing known, and treats them as irreducible. Neither gets out of itself or into the other, neither in any way is the other." He presumed that mental states could be identified independent of a commitment to the metaphysical status of the things known by them and that they could then be correlated to the brain. Ironically, his attempts to identify mental states involved him in commitments to the nature of the world as presented to mind. The only meaning that can be given things is in terms of the anticipated consequences of one's actions upon these things in the world; this anticipation also supplies the meaningfulness of thoughts. This is the basis of the "instrumental" view of thoughts—*i.e.*, reflecting upon thoughts as "tools," or as "plans of action," tells one something about the things known by them, the "tooled"; the converse also occurs.

Location of the role of science in knowledge

Each realm of the world is experienced in terms of temporal standards of thought natural to that realm; *e.g.*, standards of mathematics are peculiar because of their ideal, changeless objects. These criteria are not derived from mind alone or from things alone but from their relationship in what is termed experience. This is a "double-barrelled" term—*i.e.*, an experiencing of experienced things. The mind cannot be specified independent of things that appear to the mind, and things cannot be specified independent of their modes of appearing to the mind. Phenomena regarded abstractly as singular, or "pure," are neutral between mind and matter, which are different contexts of the very same pure experiences—contexts that comprise a single world.

James would not claim that his method is transcendental. Yet the fact remains that for him subject and object cannot be specified independent of each other, and James undercuts dualism and moves toward a transcendental explanation of the conditions of knowledge.

James tried to avoid what can be called logicism, physicalism, and psychologism. The last claimed that, because knowing is a psychical act, all that is known about must be subject to psychological laws. James replied that the known-about, the experienced, has its own autonomy, either as pure experience, a "specific nature" studied by philosophy, as a physical context studied by physics, or, finally, as a psychical context, a human history, studied by psychology. The latter two are both dependent, at least for their ultimate meaningfulness, upon the first. Physicalism attempts to infer the nature of the psychical directly from the physical, thus reducing it to the physical. Most logicisms claimed that pure reason can grasp the real in itself. James agreed that reason entertains ideal objects, the relations between which are fixed independent of the sequence of sensory experience, but asserted that this experience must decide which necessary truths apply to the world. Although some always do apply, the ascertainment of what is categorial for the world is always incomplete. Just when the world "plays into the hands of

logic" is decided in that endless interaction of "worlds" or "orders of experience"—such as the perceptual, the imaginary, the mathematical—occasioned by a thing experienced sifting through the orders trying to find one that can contain it without contradiction; Pegasus, for example, is a mythical creature just because it cannot find a place in the world of real horses. The world of perceptual things, experienced as experienceable by all and as existing simultaneously, serves as a paradigm of reality, even though other orders of experience are not reducible to it. Existence is an unusual predicate for James; it means that practical relationship of doing and concern within which things must be able to stand to men if they are to be counted as fundamentally real. James was not giving a subjectivistic account of reality, however, because he included in the fundamentally real all that can be related spatially and temporally to what can stand over against men's bodily selves. This was commonly forgotten by critics of James' popularized theory of truth, Pragmatism, which was thus systematically misunderstood.

James' contemporaries Charles Sanders Peirce and Josiah Royce stood in close dialectical exchange with him on these themes. Differences between them concerned the scope and conditions to be assigned experience. In general, Peirce argued that experience is to be construed more narrowly, in terms of mathematical logic and physics, whereas Royce argued that the understanding of truth, error, and meaning requires the assumption of an absolute knower or experiencer. Peirce was an authentic genius whose thoughts were often seeds in the more systematically developed philosophies of the other Americans.

TENDENCIES IN CONTINENTAL EUROPE

Edmund Husserl and Phenomenology. Edmund Husserl, the German philosopher, used the term Phenomenology to name a whole philosophy. In order to rid his transcendental investigation of empirical prejudgments and to discover connections of meaning that are necessary truths underlying both physical and psychological science, Husserl bracketed and suspended all judgments of existence and empirical causation. He did not deny them; rather, he no longer simply asserted them. He reflected upon their intended meaning. In reflection he claimed to see that things have meaning in terms of how they appear to men in their pre-reflective life and that awareness is in terms of this "how." In pre-reflective life, however, men are not aware of the "how" as such. By exposing this basic meaning through which men refer to things, he can free their eyes of the "cataracts" of the stereotyped and the obvious and can summon them "back to the things themselves."

Reflection on pre-reflective life

Husserl took traditional metaphysics to be infested with pre-critical commitments to existence, either physicalistic, psychologistic, or logistic. He used the term ontology, however, to apply to his study of objects of consciousness and even appropriated the Aristotelian term first philosophy. The world appears within the reflective bracket as existentially neutral (*i.e.*, as regards whether things have existence in themselves or exist for men) but ontologically ordered, because, if various orders of beings exist, then what they are can be nothing but what they are intended to be. And what they are cannot be known until all they are intended to be is known.

Husserl distinguished two types of ontologies: formal ontologies, which are the domain of meanings, or essences, such as "one," "many," "whole," or "part," that are articulated by formal logic and which Husserl referred to as empty; and material ontologies, which discover and map the meaning and structure of sensory experience through transcendental investigation. In material ontology, for example, the essence of any physical thing is discovered by varying in the imagination the object that is given within its strictly correlative mode of perceptual consciousness; the essence is that identical something that continuously maintains itself during the process of variation. It is intuited that the perceived thing cannot vary in the imagination beyond the point of something given perspectively and incompletely to any given perceiving glance; hence, this is the essence of any physical thing. This is a truth of eidetic necessity and comprises a first principle in Husserl's projected philosophical science; *e.g.*, numbers are what they are because of the ways in which they are not like things.

The Existentialists. Husserl had early distinguished the primary task of description of "morphological essences" (those with "floating" spheres of application in the sensory life) from description of essences like those in geometry, which described closed, or definite, manifolds; but the question of the theoretical status of the ordinary perceptual world, or lived world (*Lebenswelt*), became increasingly disputed among Existentialists. They asked whether there can be a philosophical science that has made all its presuppositions transparent to itself. If transcendental elucidation of the *Lebenswelt*, with its historically established sediments of meaning, is really essential to show how theoretical sciences are grounded, then one may reasonably ask how Phenomenology can be sure it has accomplished the elucidation completely, because it is itself a theory. The question was intensified by Husserl's nearly imperceptible slide into what appeared to be an Idealist position regarding the source of all meaning, a commitment to an absolute ego. If this ego is regarded as individual in any way, the problem arises of how any other individual can be as other, because it is constituted in this primal ego.

Husserl's theory of the ego was rejected by French Existentialists such as Jean-Paul Sartre and Maurice Merleau-Ponty. For the latter, the bracketing of meanings can never be completed, for consciousness is not an enclosed individual that could grasp through reflection all its possible motivations to experience and give meaning to a world. Knowers are subjects with bodies, whose perceptual life is articulated only incompletely and discloses the world in progressively surprising ways. More meaning is found in existence than can at any moment be predicted, and even the meaning of existence is not reducible to any definable set of meanings.

Husserl's approach was not nearly radical enough for Martin Heidegger, a German thinker sometimes called an Existentialist. In thinking that he could prescind so neatly from fact and retain the essence of facts, Husserl was still involved to some extent in the prejudgments—the psychologistic, physicalistic, and logistic dualisms—that he inveighed against. For Heidegger there is no realm of consciousness that constitutes meaning, and he does not think that some sharp but harmless line could be drawn between essence and fact. The ambiguity in Husserl between "object" as sense of the particular and as the encountered particular in its bodily presence is not harmless. It is unjustifiable to think that consciousness can once for all demarcate the essential sense of a thing. Thus, Heidegger discarded the very concept of consciousness, and proposed a "fundamental ontology" of human being (*Dasein*). Man as a subject in the world cannot be made the object of sophisticated theoretical conceptions such as "substance" or "cause"; furthermore, man finds himself already involved in an ongoing world that cannot as a whole be made the object of such conceptions; yet the structure of this involvement is the transcendental condition of any science of objects. For example, a man can band with other men in philosophical groups and can think about the metaphysical status of other men only because he is already essentially with others. He cannot hope to so purify his own thinking that it becomes that of an impersonal thinker, an absolute ego.

Ontology of *Dasein*

According to Heidegger, to rethink the problem of reality at its roots, it is necessary to rethink the fundamentally temporal, already-given structures of human involvement. Prejudice in the West, which construes reality, or being, on the basis of beings (*i.e.*, being as the most general feature of beings), must be overturned, and the problem of the real, the "transcendent," must be rethought on a ground on which distinctions between immanent and transcendent and between perceptual and categorial have been reconstructed. The being of the world transcends any constitution of the meaning of the

world and is a condition of experience. Thus, a sense is required of being not as object but as the underlying condition for the reality of the being of all objects.

Heidegger wanted to propose a genuine phenomenology, a study that would presuppose nothing of the traditionally formulated distinctions such as subjective–objective or phenomenal–real. The transcendence of the world can be understood only as it appears; *i.e.*, when encountered openly, things appear as appearing in part, as revealing and concealing themselves. If to the uneducated eye the sun appears to be smaller than it is, this can be corrected only by educating the person to interpret appearances—to calculate, for example, the speed and direction of light. The real is given in and through its appearances.

THE THOUGHT OF WHITEHEAD

The thought of Alfred North Whitehead is a distinctive variation on these contemporary themes. Dualisms are undermined by a phenomenology that does not bracket factual assertions. Logical and mathematical deductive schemes must be able to be interpreted in relationships crudely observable in experience, and abstractions of physics and common sense parading as realism (*e.g.*, that things exist separately within their own surfaces) must be revealed for what they are, namely, abstractions. The basic units of reality are organismic unities, "actual occasions," which are spatial and temporal extensions that cannot be exhaustively expressed in terms of distributions of matter at an instant. Their unity is constituted in a perception-like responsiveness to the universe that, though usually lacking consciousness or apprehension, is an appropriation to and for itself of the whole. This appropriation cannot be exhaustively expressed by point-instant mechanics (mechanics that is worked out in connection with the physics of relativity and thus measures not only the distance but also the time intervals between points), but is minimally a "prehension" (a term proper to Whitehead indicating the point-transcending function of perception and consciousness).

Organismic unities

Each enduring object of ordinary perception—tables, chairs, animals—is, for Whitehead, a "society" of actual occasions inheriting, through a process of appropriation and re-enactment in a predictable way, characteristics of its predecessors. Human perception is understood as a special case of prehension, in which qualities of the environment are mediated and projected on the basis of organic and affective experience of the perceiver's body, but in such a way that some of this process can be acknowledged by the percipient upon reflection. Because human consciousness is regarded as only a special case of prehensive relations, and because vacuous realisms and notions of transcendence are regarded as "fallacies of misplaced concreteness and simple location," mind–body dualisms are rejected.

Whitehead thought of "the primordial nature of God" as a general ordering of the process of the world, the ultimate basis of all induction and assertion of law, a "conceptual prehension" that functions in the selection of those "eternal objects," or repeatable patterns that are enacted in the world. God, however, does not create actual entities. He provides them with initial impetus, in the form of their subjective aim, to self-creation. Even God is the outcome of creativity, the process by which the events of the world are synthesized into new unities. It is the creative, not fully predictable, advance into novelty of a pluralistic process. The freedom of man and the determinism of nature were regarded by Whitehead as another artificial dualism.

The future of metaphysics is uncertain, not mainly because of 20th-century critics, the Logical Positivists, but because of its own not fully predictable nor controllable dynamisms. (B.W.W.)

BIBLIOGRAPHY. These works deal mainly with the nature and possibility of metaphysics.

PLATO'S *Republic*, Books V–VII, is the *locus classicus* for the idea that, over and above the particular sciences, there is a "synoptic" study concerned with "first principles"; the idea of "the Good" as the supreme object of knowledge is also briefly expounded. For the latter, see also PLATO'S *Symposium* and the *Seventh Letter*. ARISTOTLE, in his *Metaphysics*, book Λ, argued that "first philosophy" has to do with "nonsensible substance" and identified it with "theology"; in books Γ and E of the same work, however, he presented metaphysics as the study of "Being as such." For discussion of these apparently conflicting views, see W.D. ROSS (ed.), *Aristotle's Metaphysics*, 2 vol. (1924); and WERNER JAEGER, *Aristoteles, Grundlegung einer Geschichte seiner Entwicklung* (1923; Eng. trans., *Aristotle: Fundamentals of the History of His Development*, 2nd ed., 1948). Modern discussions of metaphysics spring from the preoccupation of 17th-century philosophers with method and with the origin of ideas. On the former, see especially RENE DESCARTES, *Philosophical Works*, trans. by E.S. HALDANE and G.R.T. ROSS, 2 vol. (1911–12; corrected ed., 1924, reprinted 1967), and *Philosophical Letters*, trans. and ed. by A. KENNY (1970). Descartes is particularly interesting in his discussion of whether metaphysical arguments can be set out in geometrical form (see Haldane and Ross, vol. 2), a view explicitly adopted by BENEDICT SPINOZA in his *Ethica* (1677; Eng. trans. by W.W. WHITE and A.H. STIRLING, 4th ed. rev., 1927). LEIBNIZ was more interested in doing metaphysics than in theorizing about it, whereas his academic follower CHRISTIAN WOLFF combined both in his voluminous metaphysical writings (*Vernünftige Gedanken von Gott, der Welt und der Seele der Menschen*, 1720, in German; *Philosophia Prima Sive Ontologia, Cosmologia Generalis, Psychologia Rationalis*, and *Theologia Naturalis*, various dates between 1729 and 1737, in Latin). ALEXANDER BAUMGARTEN'S compendium *Metaphysica* (1739), on which Kant lectured, was in effect a digest of these last four works. The problem of the origin of ideas was first posed in JOHN LOCKE, *An Essay Concerning Human Understanding* (1690), on which Leibniz wrote a critical commentary (not published until 1765; Eng. trans., *New Essays Concerning Human Understanding*, 1896). GEORGE BERKELEY, in criticizing Materialism in his *Treatise Concerning the Principles of Human Knowledge* (1710), pressed Locke's demand that every simple idea be derived from experience, and DAVID HUME applied this Empiricist principle with complete generality in *A Treatise of Human Nature*, 3 books (1739–40), and *Enquiry Concerning Human Understanding* (1748). IMMANUEL KANT first discussed metaphysical method in his *Untersuchung über die Deutlichkeit der Grundsätze der natürlichen Theologie und der Moral* (1764; Eng. trans., *Inquiry into the Distinctness of the Principles of Natural Theology and Morals*, 1949); he examined the whole question of the possibility of metaphysical knowledge in *Kritik der reinen Vernunft* (1781; *Critique of Pure Reason*, 1929), and *Prolegomena zu einer jeden künftigen Metaphysik die als Wissenschaft auftreten können* (1783; *Prolegomena to Any Future Metaphysics*, 1953). For a sustained criticism of Kant's critical point of view, see the writings of HEGEL, especially *Phänomenologie des Geistes* (1807; Eng. trans., *Phenomenology of Mind*, 2 vol., 1910; 2nd ed., 1 vol., 1949); and *Encyklopädie*, 3rd ed. (1830; Eng. trans.: *The Logic of Hegel*, 1892; *Hegel's Philosophy of Nature*, 1970; *Hegel's Philosophy of Mind*, 1894, enl. ed. 1971). Nineteenth-century philosophers added little to the fundamental criticisms of metaphysics developed by earlier writers. AUGUSTE COMTE in his *Cours de philosophie positive*, 6 vol. (1830–42), treated the metaphysical attitude as the precursor of the scientific. J.S. MILL in his influential *System of Logic*, 2 vol. (1843), produced a comprehensive restatement of Empiricism that was unremittingly hostile to a priori speculation. His views were sharply criticized by T.H. GREEN in *Prolegomena to Ethics*, and by F.H. BRADLEY in *Principles of Logic* (both 1883). In America, C.S. PEIRCE, though sharing Mill's interest in natural science, adopted an altogether more sympathetic attitude toward metaphysics—see his *Collected Papers*, 8 vol. (1931–35); and even WILLIAM JAMES, who presented Pragmatism as a form of "radical Empiricism," was less restrictive in his outlook than his orthodox British counterparts—compare, *e.g.*, *A Pluralistic Universe* (1909). JOHN DEWEY'S lengthy creative life rose to a number of climaxes, two of which are *Experience and Nature*, 2nd ed. (1929), in which he constructed a distant reinterpretation of Hegel's dialectic, and *Art as Experience* (1934), in which he maintained that a philosophy's aesthetics is the ultimate test of its ability to grasp the integrity of experience. There are interesting remarks on the philosophy of philosophy in the works of WILHELM DILTHEY; see his *Gesammelte Schriften*, 2nd ed., 12 vol. (1957–60), especially vol. 5. Twentieth-century criticisms of metaphysics derive mainly from the work of the Vienna Circle—for which see VIKTOR KRAFT, *Der Wiener Kreis der Ursprung des Neopositivismus* (1950; Eng. trans., *The Vienna Circle: The Ori-*

gin of Neo-Positivism, 1953); and from LUDWIG WITTGENSTEIN, *Tractatus Logico-philosophicus* (1921), which was read as an improved version of Empiricism. Among the authors who influenced these Logical Positivists were ERNST MACH, *Die Mechanik in ihrer Entwicklung historisch-kritisch dargestellt* (1883; Eng. trans., *The Science of Mechanics*, 1883); and BERTRAND RUSSELL, with A.N. WHITEHEAD, *Principia Mathematica*, 3 vol. (1910–13), and *Our Knowledge of the External World* (1914). A.J. AYER, in his anthology, *Logical Positivism* (1959), reprints in translation some of the most famous papers from the Circle's periodical *Erkenntnis*. Ayer's own book, *Language, Truth and Logic* (1936), was extremely successful in spreading Positivist ideas in America and Britain, where the work of G.E. MOORE—especially in his "Defence of Common Sense" (1925)—had created an atmosphere in which metaphysical claims were viewed with suspicion. Another influential book along the same lines as Ayer's was H. REICHENBACH, *The Rise of Scientific Philosophy* (1951); see also MORRIS LAZEROWITZ, *The Structure of Metaphysics* (1955), which attempts to explain the activities of metaphysicians in terms of psychoanalysis. For criticism of Positivist ideas, see W.H.F. BARNES, *The Philosophical Predicament* (1950); D.F. PEARS (ed.), *The Nature of Metaphysics* (1957); and K.R. POPPER, *Conjectures and Refutations*, 2nd ed. rev. (1965). R.G. COLLINGWOOD, *An Essay on Metaphysics* (1940), though it purports to answer Ayer, produces an unconventional account of metaphysics as historical analysis, of considerable interest. For further discussions, see W.H. WALSH, *Metaphysics* (1963); and Ayer's later thoughts in his *Metaphysics and Common Sense* (1969). Chief works of EDMUND HUSSERL's three phenomenological periods, each of which can be regarded as an introduction to Phenomenology, are *Logische Untersuchungen*, 1st ed., 2 vol. (1899–1901; Eng. trans., *Logical Investigations*, 2 vol., 1970); *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*, vol. 1 (1913; Eng. trans., *Ideas: General Introduction to Pure Phenomenology*, 1931, reprinted 1969); and *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie* (1954; Eng. trans., *The Crisis of European Sciences and Transcendental Phenomenology*, 1970). Among the Existentialists, important works include MARTIN HEIDEGGER, *Sein und Zeit* (1927; Eng. trans., *Being and Time*, 1962) and his later thought in *Einführung in die Metaphysik* (1935; Eng. trans., *An Introduction to Metaphysics*, 1959); MAURICE MERLEAU-PONTY, *Phénoménologie de la perception* (1945; Eng. trans., *Phenomenology of Perception*, 1962); and JEAN-PAUL SARTRE, *L'Être et le néant* (1943; Eng. trans., *Being and Nothingness*, 1956). A chief work of A.N. WHITEHEAD is *Process and Reality* (1929, reprinted 1960). For a discussion of metaphysics on an introductory level, see BRUCE W. WILSHIRE, *Metaphysics: An Introduction to Philosophy* (1969).

(W.H.W./B.W.W.)

Meteor

The word meteor, derived from the Greek *meteōros* for raised or lofty, originally meant any phenomenon or object in the air. It is now used in a more restricted sense to refer to the streak of light that results from the entry of a small interplanetary solid particle into the Earth's atmosphere.

Synonymous terms are shooting star or falling star. In a broader sense the word meteor is often used loosely for the solid particle itself, though this is more correctly termed a meteoroid. When a meteor is bright enough to attract the attention of the most casual observer, it is called a fireball.

The fireball

One of the most spectacular sights in the night sky is the fall of a very large fireball. Here the brilliance may rival that of the full moon, or, rarely, even that of the Sun. The dazzling head, often shaped like a teardrop, is followed by a scintillating trail that may exhibit all the colours of the rainbow, while it sheds sparks that fade from bright orange to red as they disappear in the darkness. In daylight or twilight conditions the observer may also see a dark cloud of dust and vapour left along the path of the fireball.

Frequently, after some minutes, detonations or rumblings may be heard, giving evidence of the arrival of shock waves and the sound of air collapsing into the semivacuum created along the trajectory. The object is then termed a detonating fireball or bolide, and a portion of the original meteoroid may survive its atmospheric passage to fall to earth as a meteorite. Such events, however, are rare and localized.

Interaction with the atmosphere. When a meteoroid first enters the upper atmosphere at high velocity, its surface collides directly with the atoms and the molecules of air, primarily nitrogen and oxygen. The energy of motion (kinetic energy) is very high. To illustrate: a 30-calibre rifle bullet weighing just under 10 grams may leave the muzzle of the gun with a velocity of 840 metres per second. A meteoroid of the same weight, entering the atmosphere in the low-speed range at 30 kilometres (about 19 miles) per second, has 1,250 times as much energy; and if it enters in the high-speed range at 60 kilometres per second, its energy is 5,000 times that of the rifle bullet. Collisions with the atmospheric gas particles dislodge material from the meteoroid surface and break up atoms and molecules of both meteoroid and atmosphere into positively and negatively charged particles. The latter, mainly electrons, form a dense column immediately behind the meteor head and act as a good radar target. Both neutral and ionized atoms are excited to produce visible luminosity. As the meteoroid penetrates lower in the atmosphere a dense, high-temperature gas cap builds up in front. Almost anywhere along the trajectory the meteoroid may start to crumble and fragment. A sudden breakdown of the solid structure may produce one or more luminous bursts, and a bright terminal burst frequently appears at the end of the visible path.

Physical structure. It is of interest to speculate on the nature of the particle that produces a visible meteor. Some 2,000 separate meteorite falls have been recognized on the basis of recovered specimens; a simple extrapolation from laboratory knowledge of these objects could be used to draw the conclusion that meteoroids are just more of the same. That this may not be so is shown by the fact that, although the majority of meteors seen are probably members of current or previously existing meteor showers (see below *Meteor showers*), not a single meteorite has been clearly identified with any particular shower. Also, physical theory, applied to the observed heights, velocities, decelerations, and luminosities, indicates that in most cases the mean densities of the meteoroids may be below that of water and that they have a fragile structure with a tendency to crumble or fragment. A small fraction, probably 1 or 2 percent and not greater than 10 percent, consists of denser, compact particles corresponding more closely to meteorites. These latter are either nickel-iron, with densities about eight times that of water, or heavy stone, with densities between three and four times that of water. It is likely that the easily fragmented types of meteoroids originated with the comets, while the denser objects are fragments of small planets that disintegrated hundreds of millions of years ago.

Densities of meteoroids

Historical appearances. From antiquity, meteors have been noted as a relatively common sight in the night sky, though a physical understanding of them has only been arrived at in the last hundred years or so. References in literature, though not numerous, are fairly widespread. In Shakespeare's play, *Richard II*, Act II, the Captain says:

> The bay trees in our country all are wither'd,
> And meteors fright the fixed stars of heaven.

Tennyson, in *The Princess*, Part VII, writes:

> Now slides the silent meteor on, and leaves
> A shining furrow, . . .

Great meteor showers, when thousands of meteors appear in a few minutes, are fairly uncommon, but given clear skies, they may be seen over a wide area and they have always made a marked impression on the watchers. One of the earliest records, from the Chinese classic *Ch'un Ch'iu* (Spring and Autumn Annals), describes an event in 687 BC: "Stars fell like a shower." The same type of occurrence is described in the Bible, Rev. 6:13: "the stars of heaven fell unto the earth." An early belief that a meteor shower foretold high winds and bad weather is indicated by Virgil, in 30 BC, in the *Georgics*, Book I:

> Oft you shall see the stars, when wind is near,
> Shoot headlong from the sky, and through the night
> Leave in their wake long whitening seas of flame.

In the early morning hours of Nov. 13, 1833, a startling shower of meteors was visible over most of eastern North

America. An eye-witness account from South Carolina bears testimony to the effect produced.

> I then opened the door, and it is difficult to say which excited me most—the awfulness of the scene, or the distressed cries of the negroes. Upward of one hundred lay prostrate on the ground—some speechless, and some with the bitterest cries, but with their hands raised, imploring God to save the world and them. The scene was truly awful; for never did rain fall much thicker than the meteors fell towards the earth.

Meteor showers. It was first noticed on the occasion just described that practically all the meteors seemed to be radiating outward from a point in the constellation Leo. As Leo rose in the east so did this position, called the radiant of the shower. A shower is customarily named after the constellation in which the radiant is located, hence the November meteors are known as the Leonid shower. Other examples are the Perseid shower in the middle of August, radiating from Perseus, and the Geminid shower, which comes in the middle of December with a radiant in Gemini. The meteors of any one shower are, in reality, all moving along parallel trajectories. The effect of radiation is a result of the viewer's perspective, and the radiant shows the direction in space from which the meteors move down through the atmosphere.

Some meteor showers, including the Perseids and the Geminids, are quite regular and return with about the same strength every year but never become very spectacular. Others, like the Leonids, are remarkable at their peak but in some years few or none are seen. A search of old records, particularly those from China, Korea, and Japan, showed that the Leonid meteors had been observed since AD 902. It was found that there was a strong tendency for the great Leonid showers to occur at intervals of 33 or 34 years. On the basis of these calculations a good return was predicted for either 1866 or 1867, and large numbers of meteors appeared on the expected date in both of these years. Also in 1866, the Comet Tempel-Tuttle was discovered, and calculation revealed the remarkable fact that the meteoroids forming the Leonid stream were moving along almost the same orbit as the comet. Other associations of this type soon became known, and now over ten examples of comets and meteor streams sharing the same orbit are recognized. There is good reason to believe that shower meteoroids are the fragments of comets that are steadily disintegrating, scattering their substance along their orbits. When the Earth crosses such an orbit it collides with large quantities of these comet fragments.

Association of meteors with comets

One of the strongest Leonid returns of the 20th century came early on the morning of November 17, 1966. Unfortunately, most of eastern United States and Canada were cloud bound so that only radar records were secured in these areas. In the American Southwest, where clear skies prevailed, 40 Leonid meteors per second was estimated as the rate for a single visual observer at the shower maximum. Notable meteor showers were provided by the Draconid (Giacobinid) meteor shower, on Oct. 9, 1933, and on Oct. 10, 1946. A number of the easily recognized meteor showers are listed in the table.

Annual Meteor Showers

shower	average date of maximum	normal duration in days	visual strength, northern hemisphere	associated comet
Quadrantid	January 3	1	medium	—
Lyrid	April 22	1	weak	1861 I
Eta Aquarid	May 5	2	weak	Halley
Delta Aquarid	July 29	15	weak	—
Perseid	August 12	5	strong	1862 III
Draconid	October 10	¼	irregular	Giacobini-Zinner
Orionid	October 21	2	medium	Halley
Taurid	November 5	30	weak	Encke
Leonid	November 17	¼–2	irregular	Temple-Tuttle
Geminid	December 14	4	strong	—

Although scientific study of meteors began early in the 19th century, their rapid motion, combined with the unpredictable nature of the appearance of an individual meteor, made it difficult until the 1940s to use instruments for securing accurate information. Most of the work before that decade was carried out visually and was subject to large errors. The best observational data have been secured since 1945 with fast new meteor cameras of radical design, and with radio techniques. Radar records can be obtained in daylight and during periods of cloudy weather, and have greatly increased the information available. One of the first results of using radar to study meteors was the discovery of a new group of meteor showers. These have radiants that are only above the horizon during daylight hours and so could never be observed optically. More recently electronic image intensifiers have been used to record meteors, thus extending the optical range to fainter limits.

Techniques of observation

By courtesy of the National Research Council of Canada

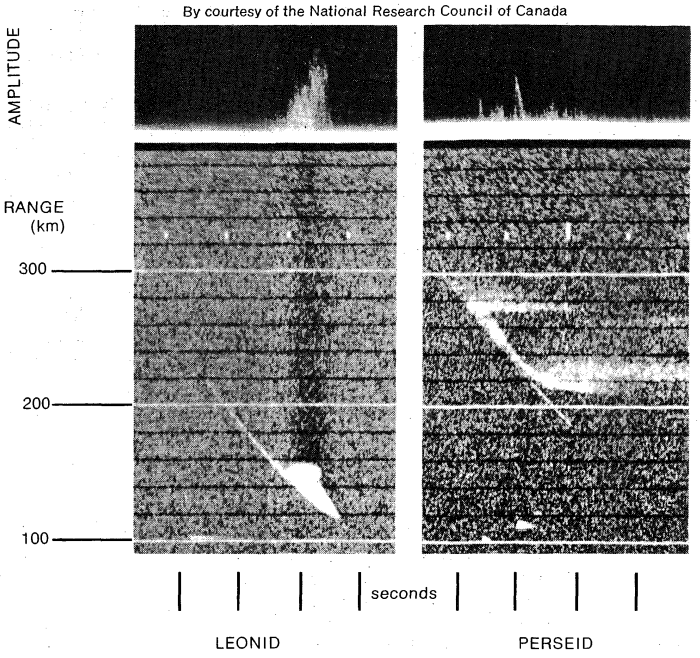


Figure 1: Radar records of two bright shower meteors, made in 1966 at Ottawa, Ontario.

Rates of occurrence. The frequency with which meteors enter the atmosphere has been the subject of much study but, until 1950, most of the data had been collected by visual observers, who recorded the number of meteors observed per hour. Such records, qualitatively useful, are difficult to reduce to quantitative values. A bright meteor will be seen if it appears anywhere in the general area of the sky the observer is facing. On the other hand, the faintest meteor visible will be recorded only if it is in a small area near the centre of the field of view. Slight changes in sky transparency alter the counts of faint meteors producing large random errors in the hourly rates. Even the best cameras do not record meteors down to the faint limit seen by the unaided eye. Radar observations have provided the best relative hourly rates throughout the year but these have not yet been accurately calibrated on the visual and photographic scales of brightness.

On an average clear dark night with no meteor shower, a single visual observer will generally see from five to ten meteors an hour, but unusual atmospheric transparency may double the number. A team of six observers will record between three and four times the number of meteors seen by an individual, and will make possible a better estimate of the total visible from one location. Meteors are relatively local in occurrence. In general, an observer at one location will not see the same meteors as a second observer at a location 500 or more kilometres (about 300 miles) away. More meteors are seen during the second half of the night, since in the early morning hours the rotation of the Earth carries the observer to the forward side of the planet in its orbit about the Sun. In the first half of the night, only those meteoroids fast enough to overtake the speeding Earth are seen as they plunge into the atmo-

sphere. The hourly rate during a meteor shower may increase by almost any amount; by factors of two to ten for the regular annual showers, up to factors as high as 10,000 to 100,000 for the really great showers.

Extrapolating from the counts for the area covered by a single observing station, the numbers of meteors encountered daily by the Earth during the entire 24 hours can be calculated. The count works out to be about 1,000,000 for objects as bright or brighter than the stars Vega, Arcturus, or Capella (zero magnitude in astronomical terms) and over 200,000,000 down to the limit of meteors visible to the naked eye. Still fainter objects can be seen in telescopes, and the number increases until meteors are included that are more than a million times fainter than the standard zero magnitude defined above. Somewhere in this size range the meteoroids no longer radiate visible light as they are slowed down in the upper air. These tiny particles, much less than a millimetre in diameter, are called micrometeoroids.

Heights. If a meteor is photographed from two stations separated by a reasonable distance—*e.g.*, 30 to 60 kilometres—its height above the Earth can be determined by triangulation. Average heights range from 70 to 110 kilometres, roughly 45–70 miles. As relatively few meteors are seen close to the zenith, the most common elevations above the horizon being from 20 to 70 degrees, it is obvious that the great majority of meteors are more than 90 kilometres from the observer when seen. In exceptional cases, usually slow, bright fireballs, the object when seen may be as close as 50 kilometres or as far away as 500 kilometres. The observation that a bright meteor has fallen to Earth "just beyond some trees" almost always may be discounted as incorrect. Normally, when a meteoroid does reach the ground as a meteorite it travels the final portion of its trajectory as a dark body; it is rarely luminous below heights of 10 to 20 kilometres.

Velocities. Even casual observers have noted the speed with which most meteors move across the sky. When this rapid angular motion is considered along with the distances noted above, it is clear that the real speed of meteors in the atmosphere is very high. The best determinations of meteor velocity are made by photography, using cameras equipped with shutters that block the lenses at a fixed frequency. This produces a series of breaks in the photographic trail, the number of breaks indicating the number of lens occultations and thus the duration of the meteor. The use of two such records from opposite ends of a base line makes possible the calculation of accurate velocities. Slightly less accurate values can be obtained by employing radio equipment. Most meteors slow down very little during the time they are visible, average decelerations amounting to not more than 1 or 2 percent.

Determining meteor velocity

The velocity at which meteoroids enter the atmosphere results primarily from two motions, that of the meteoroid in its orbit and that of the Earth (about 30 kilometres, or about 19 miles, per second) in its annual path about the Sun. The gravitational pull of the Earth on the particle increases the speed of the meteor by from 1 to 3 or more kilometres per second. The observed speeds of meteors range from a low near 15 kilometres per second, for the objects overtaking the Earth, to a high of over 70 kilometres per second for the particles that the Earth meets head on. A frequency plot of observed speeds shows a tendency to clump around two ranges, low-velocity meteors with speeds from 20 to 40 kilometres per second, and high-velocity meteors with speeds from 60 to 70 kilometres per second. The faster meteors are visible at greater heights. Those in the high-velocity group appear near 110 kilometres altitude and disappear near 95 kilometres. The corresponding values for the low-velocity group are 95 and 80 kilometres, respectively. Maximum luminosity normally occurs two-thirds of the way down the visible trajectory. The above heights are general averages only and there is considerable variation for individual meteors. Very bright objects, for example, will usually penetrate much lower in the atmosphere, by as much as 20 kilometres or more; the size of the original meteoroid and the inclination of its trajectory to the horizontal will also affect the height of the visible path.

Analysis of luminosity. A brilliant meteor is often described as looking like a magnesium flare, and this subjective impression happens to be quite close to the truth. Meteors, like flares and fireworks, have a large percentage of their luminosity produced by individual atoms radiating with their characteristic colours in an energized cloud of gas. Practically all of the visible light originates in a compact, gaseous cloud of atoms and molecules surrounding or closely trailing the meteoroid as it moves through the atmosphere. These atoms and molecules are a mixture of those evaporated or ablated from the meteoroid, together with others trapped from the upper atmosphere.

Meteor spectra

The nature of the light of meteors may be studied if a dispersing unit, such as a glass prism, or a diffraction grating, is used in conjunction with a camera. This produces a record of the meteor's spectrum. Low-energy radiations of the elements iron, oxygen, magnesium, sodium, nitrogen, and calcium are the most prominent in the spectra of meteors. This does not mean, necessarily, that these are the most abundant elements in the radiating source. Other atoms that are recorded frequently include hydrogen, silicon, manganese, and chromium, while the molecule of nitrogen is observed as a general background in the radiation. Whether a given element appears in a meteor spectrum depends not only on its presence in the meteoroid or the atmosphere but also on whether the atom or molecule has low-energy radiations in the wavelengths that can be photographed. The meteors that move through the atmosphere at higher speeds radiate light that requires higher energies for its production.

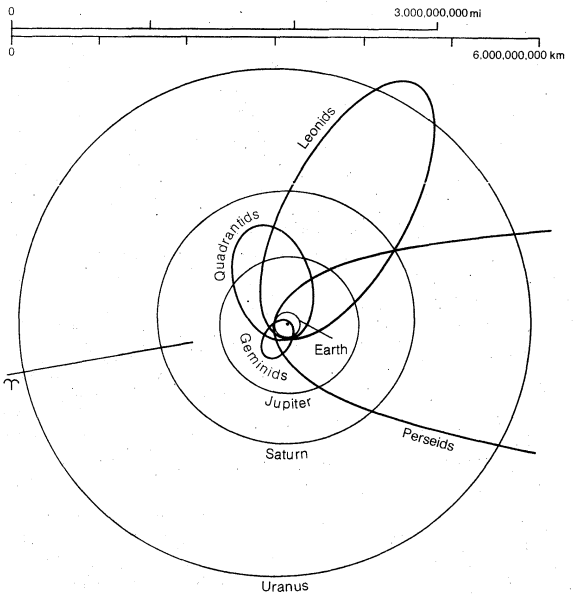


Figure 2: Orbits of four meteor showers plotted in relation to planetary orbits. Planes of meteor orbits deviate from those of planets. To illustrate their true shape, all meteor orbits have been plotted as if they were in the plane of the paper which represents the ecliptic, the plane of the Earth's orbit.

Meteor trains. The major portion of the light from a meteor is produced near the moving solid particle and is of very short duration, as is shown by the sharpness of the breaks produced in the photographed trail by an occulting shutter on the camera. In the case of bright meteors, particularly if they are fast, a residual luminosity called the train may be left behind the meteor head. That part of the train with a duration measured in small fractions of a second is called the wake, and the light is produced by some of the same types of atoms appearing in the head, but at a lower excitation. Another kind of train luminosity appears in the upper part of the meteor trajectory at heights from 100 to 115 kilometres, about 60–70 miles. It has a duration of one to two seconds and consists of the green radiation of the neutral oxygen atom, the strongest and commonest feature in most auroral displays. A third type of train luminosity may last up to 10 or 15 minutes as a faint

glow. This type has not yet been fully analyzed, but the light of the sodium and magnesium atoms and of the nitrogen molecule probably contributes to it.

The long-enduring trains of bright, fast meteors were the first phenomena to reveal the strong wind shears and jet streams common at heights between 80 and 100 kilometres. The train, initially a thin, straight line, rapidly diffuses and becomes twisted into complicated snakelike forms. This circumstance may well explain some rather bizarre accounts of things seen in the sky during great meteor showers of the past.

Orbits. Once the direction of motion of a meteor in the atmosphere and the corresponding speed along the trajectory have been accurately determined, it is possible to compute the orbit of the original meteoroid. Since the meteors that become visible either vaporize or fall to earth, only the orbits of meteoroids that have ceased to exist, as such, are known. It can be inferred that the observed orbits are typical of the others that are still active, but the Earth encounters only a very small fraction of the meteoroids in interplanetary space. These little objects are far too small to be detected until they become luminous in the upper air. Hence, only those meteoroid orbits can be studied that intersect the orbit of the Earth. Assumptions about meteoroids in other orbits depend on uncertain extrapolations.

The correspondence between shower meteoroid orbits and certain comet orbits has been noted already. As a class these orbits are elongated—that is, of high eccentricity—and objects moving in such orbits vary greatly in their distance from the Sun. The planes of the orbits tend to be fairly random and may deviate widely from the plane of the Earth's orbit, the ecliptic. This means that in some cases the meteoroid streams move about the Sun in the opposite direction to the planets, and the Earth meets the meteoroids almost head on. The Leonid shower is of this type; the Leonid meteors enter the Earth's atmosphere at a speed of over 70 kilometres (43 miles) per second, the fastest of any shower.

Orbits of sporadic meteors

The large numbers of meteors that do not belong to a recognized shower, often called the sporadic meteors, show a preference for direct motion; that is, they revolve around the Sun in the same direction as the planets. The planes of the orbits of sporadic meteors tend to be fairly near the ecliptic plane. The orbits for very bright, slow fireballs have planes that cluster still more closely near the ecliptic, and the shapes of these fireball orbits are more nearly circular, corresponding to those of the minor planets. Analysis of large numbers of faint-meteor orbits indicates that the density of small meteoroids in interplanetary space is high between the Earth's orbit and the Sun and decreases out toward the orbits of Mars and Jupiter.

In the early decades of this century good instrumentally determined orbits for meteors were nonexistent. It was concluded from indirect evidence that a significant percentage of the meteoroids were interstellar; that is, they came from outside the solar system. Such objects would have open-ended, hyperbolic orbits about the Sun and would move faster than objects in elliptical orbits. The use of radar techniques and of fast cameras has now proved that practically all observed meteoroids move in closed elliptical orbits, and it is difficult to find even a single example of a particle with an accurately determined hyperbolic orbit. Apparent exceptions to this rule generally prove to be particles recently accelerated into a hyperbolic orbit by a close approach to a large planet. Therefore the complex of meteoroids and meteoritic dust is part of the Sun's family, along with comets, planets, and satellites.

Masses. One of the major problems in the study of meteors is to determine the mass of the original meteoroid corresponding to any given meteor. Since the majority of meteoroids are vaporized high in the upper air, and since even the meteorites that land have lost a significant fraction of their mass in their passage through the atmosphere, all calculations of original mass must be made by indirect methods. Unfortunately, a complete physical theory to facilitate such calculations has not yet been worked out, though with the help of artificial meteors fired from rockets good progress is being made in determining the necessary constants for the theoretical equations.

It may be assumed that on the average a meteoroid that produces a zero-magnitude meteor weighs close to one gram. This can be thought of as a standard meteoroid and in one day the earth encounters about a million objects of this mass or heavier. The best recent knowledge concerning the relative numbers of meteoroids of various masses is based on radar studies. If a count is taken of particles down to a mass 0.01 gram, which will produce a meteor near the limit of visibility with the naked eye, the number will be 400 times the count down to the standard meteoroid of 1 gram. Following the same rule, a count down to masses of 0.0001 gram will give 160,000 times the number for the standard. For still smaller particles the ratio is not so well known, but other sources of evidence, such as the detection of nickel-iron particles in arctic snows and deep-sea drill cores, suggest that the Earth sweeps up several hundred tons of mass a day in the form of micrometeoroids or meteoritic dust. A simple calculation will show that, if the above figures for the numbers of visible meteors are of the right order of magnitude, the Earth collects far more mass in the form of dust and micrometeoroids than it does in the range of the larger particles.

Although the frequency of occurrence of specific masses in space is not yet known with high accuracy, the Pegasus satellite program of the U.S. has produced good values of penetration frequency. A sheet of aluminum one metre square and 0.01 centimetre thick, remaining in space in the vicinity of the Earth and Moon, will be penetrated once every 10 days on the average. Increasing the aluminum thickness 10 times decreases the frequency of penetration by a factor of nearly 200.

BIBLIOGRAPHY. L. KRESAK and P.M. MILLMAN (eds.), *Physics and Dynamics of Meteors* (1968), a detailed symposium discussing current research, with particular reference to meteor orbits and the physical theory of meteor ionization and luminosity; G.S. HAWKINS, *Meteors, Comets, and Meteorites* (1964), a brief popular treatment of the subject; B.M. MIDDLEHURST and G.P. KUIPER (eds.), *The Moon, Meteorites, and Comets* (1963), general review articles on meteors, comets, the Moon, meteorites and impact craters; D.W.R. MCKINLEY, *Meteor Science and Engineering* (1961), a technical monograph with particular emphasis on the engineering and observational data in the field of radio meteor research; G.S. HAWKINS (ed.), *Meteor Orbits and Dust*, Smithson. Contr. Astrophys., vol. 11, NASA SP-135 (1967), a symposium dealing with the determination of meteor orbits and the study of interplanetary dust through the use of satellites and rockets; J.L. WEINBERG, (ed.), *The Zodiacal Light and the Interplanetary Medium*, NASA SP-150 (1967), a symposium covering current research on interplanetary dust and faint meteor orbits, through zodiacal-light observations, radio meteor observations, and the use of satellites and spacecraft.

(P.M.M.)

Meteorite

Though the term meteor, until the end of the 17th century, meant almost any atmospheric or supposedly atmospheric phenomenon—hail, rain, lightning, rainbows, etc. —the modern use of the term is restricted to what the 18th century called fiery meteors: shooting stars and those larger, more brilliant, but essentially similar objects usually called fireballs, or bolides. Meteorites may be defined as solid objects reaching the Earth's surface from interplanetary space. Strictly speaking, although any solid mass that manages to survive passage through the atmosphere and any dust arising from material melted or vaporized during passage would both fall under this definition, the term is usually restricted to the former.

The fall of a meteorite in this narrower sense is a striking phenomenon, involving a brilliant fireball and startling, thunderous noises. A very large meteorite—one weighing from about 200 tons up—explodes on impact with the ground (see METEORITE CRATERS), but, fortunately, such events are very rare. The solid mass itself is of great scientific interest, not only as the only extraterrestrial material available before the lunar landings took place but also because the analyses of the chemical and rock characteristics of meteorites carry clues to the early history of the solar system. The mean abundance of the

elements in a certain class of meteorites, the chondrites, is believed to approximate the abundance of elements in the solar system, excluding a few volatile ones. From study of the textures and internal structures of meteorites and of their content of various radioactive atoms (radionuclides) and the breakdown products of radionuclides, conclusions may be drawn about the early history of both meteorites and the solar system as a whole. It is widely accepted that meteorites originate from collisions between asteroids, probably 100 to 300 kilometres (roughly 60 to 190 miles) in diameter, which aggregated from fragmentary material about the same time as the Earth, some 4,500,000,000 years ago. A few astronomers, however, still champion a theory of an origin from the Moon or from Mars for at least some classes of meteorite.

Meteorites are named according to the place of discovery or fall, usually identified by the nearest named location, which may be a small farm, village, or other geographic place name. They are described as stones, irons, or stony irons, according to their composition (see below *Classification*).

Historical. Men have been aware at least since the dawn of civilization that stones and, more rarely, masses of iron sometimes fall from the sky. Nickeliferous iron beads found in predynastic graves at Gerzeh, Egypt, and fragments of iron found at Ur, in Mesopotamia, predate knowledge of iron smelting and are certainly of meteoritic origin. Such celestial objects were believed to have been sent by the gods. It was suggested as long ago as 1796 that cult objects found in classical times in a number of cities of Egypt and the Levant were meteorites, though the evidence is not conclusive. A black stone built into the sanctuary of the Ka'bah, in Mecca, has also been regarded as probably meteoritic.

Meteorite falls in ancient records

There are numerous records in ancient Chinese annals, in Greek and Latin literature, and in medieval annals of the fall of stones, but in many cases these reports may really refer to giant hailstones, and in others they are either insufficiently detailed or contain too many contradictions to be acceptable as conclusive evidence. Even reports of the fall of masses of iron (*e.g.*, Lucania, Italy, 56 BC; Meissen, Saxony, 1164; Naunhof, Saxony, about 1540) cannot be accepted as well established; and the tendency to embroider the details was and still is so strong that scientists are rarely prepared to accept any individual fall as proved unless material has been preserved and studied.

Although up to the early 18th century it was generally believed that stones do fall from the sky, there followed a period of almost a hundred years when this phenomenon was widely regarded as an impossibility. This view probably resulted from the mixture of inaccurate observation, misremembered detail, secondhand, misunderstood reporting, and preconceived notions of what was observed in the centuries preceding. The general philosophy of the mid-18th century was doubtless also an influence—it was a notion, largely unformulated, that Newtonian physics had explained the cosmos, leaving no room in the solar system for anything but the Sun, the major planets with their satellites, and a few wandering comets. An unconscious carry-over from the Ptolemaic cosmology, in which everything outside the orbit of the Moon was immutable, may also have been involved. The idea that the fall of stones from the sky is an absurdity was particularly strongly held in the Académie Française, and, under this influence, meteorites preserved from earlier times in several museums and collections of curiosities were thrown away. The U.S. president Thomas Jefferson refused to believe in the extraterrestrial origin of a meteorite that fell in Weston, Connecticut, in 1807. The oldest meteorite seen to fall and still preserved is the stone that fell at Ensisheim, Alsace, on November 16, 1492. Originally weighing about 127 kilograms (280 pounds), it was kept in the parish church for a long time, and the main mass is still treasured in the town hall of Ensisheim, though fragments have been distributed to museums in many parts of the world. Few other meteorites now preserved fell or were found earlier than 1790.

In 1766 an Italian scientist published a careful description of the Albareto (Italy) fall of that year; in 1769 the Lucé (France) fall of 1768 was described; and in 1794 a description of the Siena (Italy) fall of that year was published. In 1794 reports of iron masses for which an artificial origin seemed unlikely were collected and in 1796 reports of the fall of stones and irons were collected from older literature, chiefly through the efforts of the German physicist Ernst Chladni. To these were added notes on the Siena fall and a preliminary account of the Wold Cottage (Yorkshire) fall of 1795. This accumulated evidence, with reports of further falls at a rate of nearly one a year, shook the skeptics. When a shower of over 2,000 stones fell at L'Aigle, France, a commission was appointed to investigate, and its report, in 1804, firmly established the reality of falls of stones from outer space.

The sounds heard during the fall of a meteorite are very like thunder, and, before the nature of lightning was understood, it was reasonable to attribute damage from it to the fall of some solid object—a thunderbolt. The phrase a bolt from the blue indicates that a meteorite may fall from a cloudless sky. With this confusion of the two phenomena, it is not surprising that during the period of skepticism the Lucé meteorite that fell in 1768 was explained as a stone struck by lightning. In cases in which the actual impact of the meteorite had been witnessed, the scientific explanation was that these were stones thrown out by terrestrial volcanoes or, more reasonably, dust from volcanic eruptions carried by the winds for long distances and then fused and coagulated by lightning. The very light coloured and friable Pettiswood (Ireland) stone was attributed to sand from a nearby lake carried aloft by winds and there "coagulated and compacted."

Classification. The term meteorite, in the wide sense of any natural object arriving on the Earth's surface from outside the atmosphere, includes ablation material and micrometeorites (meteorites of microscopic size) and, possibly, some things not usually classed as meteoritic.

The kinetic energy (energy of motion) of any body entering the atmosphere is largely converted into heat, and much of the body's substance is melted or even vaporized to fall to the surface later as ablation dust; it is rarely possible to distinguish such material from volcanic dusts or certain industrial dusts (perhaps the best authentic example is that of the spherules of metallic iron and iron oxides found in association with the fall of iron meteorites in 1947 at Sikhote-Alin, northeast of Vladivostok, Russian S.F.S.R.), but the total amount of such material reaching the Earth's surface must greatly surpass the total of macroscopic, or large, meteorites. Calculations indicate that a small (less than about 0.01 millimetre) dust particle entering the Earth's atmosphere at cosmic velocity (9 to 40 kilometres [about 6 to 25 miles] per second) may lose heat by radiation as fast as it gains it by deceleration and so escape melting and slowly drift down to the surface. There can be little doubt that much material collected as cosmic dust was indeed of meteoritic origin, but for the most part it is almost impossible to distinguish such fine meteoric material from terrestrial dusts and, especially, from some industrial dusts.

Three main groups of meteorites

Excluding this dust, meteorites fall into three major groups: irons, or siderites; stones, or aerolites; and stony irons, or siderolites; each of which is further subdivided. An example of each major group is shown in Figure 1. For many purposes it is also useful to distinguish between falls (that is, meteorites whose falls were observed) and finds (meteorites discovered on or embedded in the ground but whose date of arrival on the earth is unknown); this distinction is important for detailed study of the chemical composition of meteorites, because a specimen that has lain on or in the Earth for an unknown time may well have undergone appreciable weathering or other alteration; but it must not be forgotten that individual specimens of an observed fall often were not picked up for months or even years and may also be appreciably altered.

Rates of meteorite weathering vary widely; a porous stony meteorite exposed to conditions of alternate wetting and drying and, especially, to frost action could be disintegrated in a few months, whereas estimates of age based on the rate of decay of the radioactive isotope

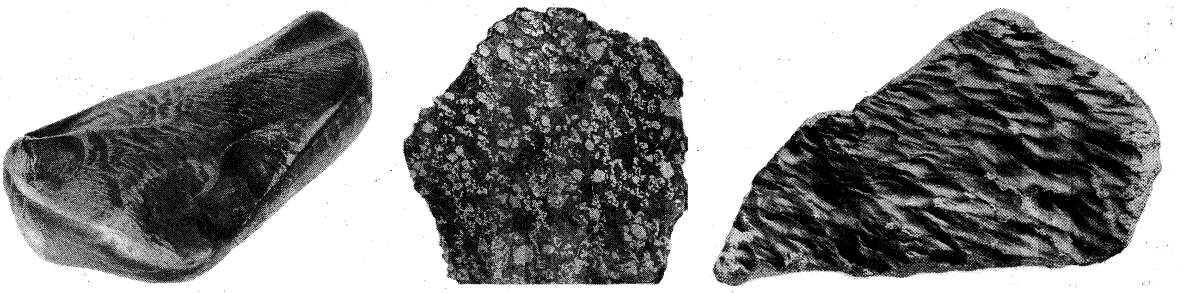


Figure 1: (Left) Iron meteorite: Nedagolla, an observed oriented fall showing the thin crust with bright metal in places and good oriented flow lines (approximately 30 percent of actual size). (Centre) Stony iron meteorite: Estherville (approximately 11 percent of actual size). (Right) Stone meteorite: Barwell, a fragment showing crust and regmaglyphs (approximately 24 percent actual size).
By courtesy of the trustees of the British Museum (Natural History)

carbon-14 suggest that Potter, a compact stone that penetrated to the subsoil in southwest Nebraska, fell to the Earth some 20,000 years ago; and there is some evidence that under favourable conditions meteoritic irons may survive for 1,000,000 years. Probably an average life for a stony meteorite in all but arid climates is of the order of 100 years.

The course of weathering is fairly simple for iron meteorites; for most there is simply a gradually thickening crust of iron oxides—a mixture of goethite, hematite, and nickelian magnetite. Nickel is usually, but not always, partially leached out; *i.e.*, washed out by percolating water. Very rarely there is preferential oxidation of the nickel-poor component kamacite.

Weathering of stony irons and those stones that contain metal (the great majority) commences with rusting of the iron; this may be almost wholly oxidized before there is appreciable alteration of the sulfides and silicates, but ultimately the silicates undergo hydration (combination with water) and oxidation, and the final product will no longer be recognizable as of meteoritic origin.

Proof of origin

Strictly speaking, the final proof that an object is a meteorite is actual observation of its fall, but mistakes are common—any unfamiliar object may be picked up. A body of plausible evidence for extraterrestrial origin has gradually been built up. A freshly fallen stony meteorite will almost invariably be covered with a thin, glassy, black fusion crust produced during its passage through the atmosphere (in a few rare types of meteorite the crust is cream coloured owing to the absence of iron); most stony meteorites contain more or less metal, which is always a nickel–iron alloy with 5 to 30 percent nickel, and careful examination of most discloses rounded structures called chondri or chondrules unknown in any terrestrial rock; such meteorites are called chondrites. In all but a few rare classes, the chemical composition is different from that of any terrestrial rock; if there are types as yet unrecognized, the presence of a fusion crust will support direct evidence of fall, as it does for a few unusual types, such as calcium-rich achondrites (lacking chondrules) and carbonaceous meteorites.

For irons, the principal evidence is chemical and metallographic: no iron meteorite is known to contain less than 4 percent nickel, and only a few contain more than 20 percent; nearly all are extremely coarse grained, most medium-sized irons being basically single crystals; most irons with nickel contents in the range of 7 to 11 percent show a large-scale characteristic pattern of separation between crystals with different nickel contents, called the Widmanstätten structure, that cannot be developed in artificial alloys in any practicable cooling time. The presence of the Widmanstätten structure can therefore be used as conclusive evidence that a piece of iron is meteoritic in origin. In any case of doubt, a study of the amounts of the minor and trace elements present, especially gallium, germanium, iridium, copper, chromium, and vanadium, will usually confirm or disprove meteoritic origin.

Terrestrial native iron, which occurs in some basalts, usually has less than 3 percent nickel, though one analysis with 6.5 percent has been reported. There appeared to be no data available in the early 1970s for the trace elements in native iron. The alloying elements mainly used in artificial irons, apart from nickel, are very minor constituents of meteoritic iron.

Each of the three major classes of meteorites is subdivided, as shown in Table 1. The significance of these classes will become apparent when the composition and structure of meteorites are considered in more detail, but it is generally interesting and significant to note their numerical distribution as falls and finds among the various classes. It will be noted that, whereas, among finds, stones are far fewer than the more resistant irons, when observed falls are compared, it is clear that far more stones than irons arrive at the Earth's surface. The proportion of stones must be even greater in interplanetary space, for their greater mechanical weakness as compared to the irons would make for greater destruction in the atmosphere.

Within each of the three major groups, the unequal distribution among the several subdivisions is very noticeable. About a third of all irons are medium octahedrites; over half the stony irons are pallasites (see Table 1); over 80 percent of all stones are chondrites; and, of the chondrites, about 40 percent fall into each of the olivine–bronzite and olivine–hypersthene classes, which are often referred to jointly as common chondrites.

Phenomena of fall. Studies of meteor trajectories indicate that velocities up to 40 kilometres (about 25 miles) per second occur; the speed on entry for meteorites ranges from 13 to 21 kilometres per second. An object entering the upper atmosphere at such speeds rapidly collects a nose cap of ionized air, which can attain a large size and become brilliantly luminous; the forward surface of the object is intensely heated, melts, and is swept off as fine droplets, which condense to form a smoke trail that may persist for a considerable time. At night there may be a luminous trail, caused by slow recombination of ionized air. The fireball commonly becomes visible at a height of 150 to 120 kilometres (about 90 to 75 miles).

Fireballs and other flight phenomena

As the mass proceeds, the pressure in its gaseous nose cap builds up; most meteorites tumble over and over during flight, and, whenever there is a concave surface to the front, there will be a splitting moment tending to break up the mass. Most stony meteorites appear to break up in flight, and often the fragments break again. Some stones from the Richardton (North Dakota) fall preserve evidence of at least three fragmentations, in the shape of successive generations of fusion crust. Thus, well-documented observed falls are commonly multiple, and the number of individuals may run into many thousands, of all sizes down to fractions of a gram. Complete crusted stones of the Holbrook (Arizona) fall weighing no more than 18 milligrams (0.0063 ounce, avoirdupois) have been collected. Irons, because of their greater tensile strength, rarely fragment—the Sikhote-Alin (Siberia) fall of 1947 is a notable exception, but there the mass, a very coarse octahedrite, fractured along the taenite plates (taenite is an alloy of iron and nickel). Rarely, a meteorite has sufficient stability in flight, through symmetry or perhaps through sufficient rotational energy, to maintain one orientation; then, ablation tends to shape it into a

Table 1: Classification of Meteorites
(numbers as of December 1, 1965)

	falls	finds	"paired" and doubtful
Irons (Siderites)	43*	551*	67*
Nickel-poor ataxites†, showing no structure on a polished and etched surface; Ni 4.3% to 6%	1	23	4
Hexahedrites†, showing slipbands (Neumann lines) on etching; Ni 5.5% to 6.7%	6	39	10
Octahedrites, showing a Widmanstätten pattern on etching; Ni ~6% to ~13%	30‡	412‡	24‡
further subdivided into:			
Coarsest octahedrites, kamacite lamellae over 2.5 mm wide	2	22	
Coarse octahedrites, kamacite lamellae 1.5 to 2.5 mm	4	87	7
Medium octahedrites, kamacite lamellae 0.5 to 1.5 mm	14	189	13
Fine octahedrites, kamacite lamellae 0.2 to 0.5 mm	4	62	1
Finest octahedrites, kamacite lamellae less than 0.2 mm		21	1
Nickel-rich ataxites, showing no structure on etching; Ni 9% to 60%	1	41	
Brecciated octahedrites, in which the Widmanstätten pattern only retains a given orientation over regions less than 10 cm^2; often with silicate inclusions	2	5	
Metabolites, in which the Widmanstätten pattern appears to have been destroyed by natural or terrestrial reheating		3	
Stony irons (Siderolites)	12	58	3‖
Pallasites, consisting of grains of olivine (Fe_2SiO_4 10–20 mole %) embedded in a matrix of nickel–iron (10%–20% Ni)	4	39	2
Siderophyres, bronzite (20 mole % $FeSiO_3$) and tridymite in nickel–iron		1	
Lodranite, olivine (13 mole % Fe_2SiO_4), bronzite (17 mole % $FeSiO_3$) with nickel–iron (9% Ni)	1		
Mesosiderites, a very miscellaneous group; the metal amounts to 40% or more but does not form a continuous network	7	18	
Stones (Aerolites)	725¶	404¶	72¶
Chondrites	624♀	380♀	25♀
subdivided into:			
Enstatite chondrites; dominant silicate enstatite, often almost iron-free; metal 20% to 35%, with Ni < 7% and often containing Si	11	6	
Olivine–bronzite chondrites♂, about equal amounts of bronzite and an olivine with 15 to 20 mole % Fe_2SiO_4; metal 16%–21%, with a nickel content of 7%–12%; total iron content ~26%	225	180	8
Olivine–hypersthene chondrites□, less pyroxene than olivine, which has 22% to 26% Fe_2SiO_4; metal < 10%, rich in Ni; total iron ~22%	273	170	12
Amphoterites◇, little pyroxene, olivine 28% to 32% Fe_2SiO_4; very little Ni-rich metal; total iron ~22%	39	10	
Olivine–pigeonite chondrites▲; iron-rich olivine (30–40 mole %) Fe_2SiO_4, little pyroxene, little or no metal; total iron 26%	15		
Carbonaceous chondrites, total iron 25%; characterized by the presence of 2% or more carbon, partly present as complex hydrocarbons, and of considerable hydrogen (hydroxyl groups and water);	20+		1
further divided into:			
Type I, in which sulfur is present as sulfates, largely water soluble; magnetic, iron being present mainly as nickelian magnetite, and silicates as amorphous hydrated magnesium silicate	5		
Type II, containing much elementary sulfur; nonmagnetic, iron being present mainly as a hydrated silicate related to the serpentines	13		
Achondrites	59§	9	1
divided into:			
Calcium-poor achondrites;	18	2	
subdivided into			
Enstatite achondrites (aubrites), composed of nearly pure enstatite with rare blebs of metal (kamacite) or oldhamite or both	7	1	
Hypersthene achondrites (diogenites), composed of hypersthene with 20–30 mole % $FeSiO_3$ with small amounts of accessory minerals	8		
Olivine achondrite (chassignite), over 95% olivine with 33 mole % Fe_2SiO_4	1		
Ureilites, consisting of grains of olivine and clinopyroxene, together with about 5% of metal, in a black matrix that appears to be severely shocked and contains about 1% of finely divided diamond	2	1	
Calcium-rich achondrites			
Nakhlites, containing 75% ferroan diopside and 15% olivine with about 66 mole % $FeSiO_4$, with a little feldspar and magnetite	1	1	
Augite achondrite (angrite), consisting of over 90% of a titanian augite with accessory olivine and troilite	1		
Eucrites consist of a plagioclase near An_{90} and pigeonite, with accessory chromite	26	3	1
Howardites differ from eucrites in containing hypersthene in place of augite, and rather less plagioclase	12	2	
Total, meteorites, all classes	785	1015	221

*Including 60 unclassified irons, 3 falls, 28 finds, 29 paired and doubtful. †Some authorities unite these two classes. ‡Including 39 octahedrites not further classified—6 falls, 31 finds, and 2 doubtful. §Including 3 unclassified achondrites. ‖Including one doubtful and unclassified stony iron. ¶Including 97 unclassified stones—39 falls, 15 finds, and 43 doubtful. ♀Including 59 unclassified chondrites—41 falls, 14 finds, and 4 doubtful. ♂Also referred to as H-group chondrites. □Also referred to as L-group chondrites; it should be noted that meteoriticists continue to place the boundary between bronzite and hypersthene at 20 mole % $FeSiO_3$. ◇Also referred to as LL-group chondrites. ▲Also referred to as carbonaceous chondrites, Type III, but not all members of the class are appreciably carbonaceous. +Including 2 not classed as Type I or Type II.
Source: B.H. Mason, *Meteorites* 1962.

blunt cone, and an oriented stone will usually escape fragmentation; the Middlesborough (Yorkshire) fall of 1881 is an excellent example. The fireball is often described as moving in a sinuous course; this motion is probably due to the breaking off of sizable fragments, sometimes described as "throwing off sparks," while a break into two or more pieces of comparable size is usually described as an explosion.

A meteorite may enter the upper atmosphere at any angle; should its fall be nearly vertical, the fireball will last only a short time, and it is perfectly possible for such a fall to pass unnoticed even in a fairly populous area; at the other extreme, a mass entering the atmosphere horizontally (that is, at grazing incidence) can give rise to a spectacular fireball visible over a long trajectory. In 1676 a fireball observed moving from the direction of the Adriatic passed over the whole width of northern Italy and fell into the sea off Leghorn; the fireball from which the Rochester (Indiana) meteorite of 1876 fell was visible over a course of 1,000 miles; and in 1913 a procession of between 30 and 60 fireballs was seen over a course from western Saskatchewan to southeast of Bermuda.

As the meteorite loses its initial kinetic energy to the fireball, it is slowed down and presently the fireball shrinks and is extinguished; for a mass of moderate size, the process usually occurs at about 17 to 25 kilometres (about 11 to 16 miles), and, from that point on, the meteorite falls as a dark body. If its atmospheric trajectory makes a small angle with the ground, the point of impact may be many kilometres from the point of extinction.

Over most of its path, the meteorite will be travelling at supersonic speed and will give rise to a typical shock wave, or sonic boom, which has been compared to the sounds of gunfire, of a heavy vehicle backfiring, or of a clap of thunder; there is commonly a series of echoes, as with a thunderclap. For distant observers, the boom and its echoes will naturally arrive some minutes after the fireball is seen; but in many cases whistling, whining, or hissing noises are reported to have been heard simultaneously with the sighting of the fireball or to have called the observer's attention to it, the typical boom being heard later. These sounds have not been fully explained, though it has been suggested that they are electrical in origin and are caused by the ionized mass of air; from their time of arrival, they cannot be normal sound waves.

As a rule, there is no further fragmentation once the fireball has extinguished, but the final breakup of the Barwell (Leicestershire, 1965) meteorite took place at a height of less than 10 kilometres (about 6 miles), and two fitting fragments without any trace of fusion crust were found about 600 metres (roughly 650 yards) apart.

Impact with the ground

The fragments from a multiple fall are usually distributed over a considerable area. The shape of the distribution pattern, an ellipse, depends principally on the entry angle and the fragmentation height; as a rule, larger, more massive pieces carry farther against air resistance, though there are rare exceptions (see Figure 2).

From B. Middlehurst and G. Kuiper (eds.), *The Moon, Meteorites and Comets;* University of Chicago Press (©1963)

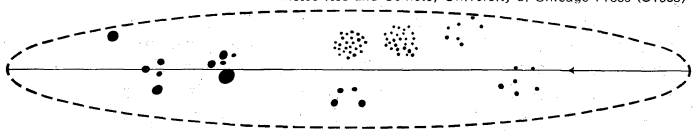


Figure 2: Distribution pattern of meteorite shower at Richardton, North Dakota. The arrow points from southwest to northeast.

The velocity of a meteorite at impact with the ground will, of course, depend on the angle of entry, the size and shape of the mass, and, in the case of a fragment, the height at which it became detached. Records indicate that, for masses of a few kilograms, very friable (easily crumbled) meteorites can land gently with little damage, while, at the other extreme, a 50-kilogram mass has been known to penetrate 1.25 metres into stiff clay, and one of 200 kilograms to nearly 3 metres, also in clay.

Since the kinetic energy increases as the cube of the equivalent diameter, but the air resistance increases only as its square, the buffering action of the air decreases rapidly as the mass increases, and a mass of 100 tons or more will have retained most of its entry velocity when it reaches the surface. Accordingly, while a moderate-sized mass will produce only a small impact pit, a mass arriving at the surface with a velocity of 0.5 to 2 kilometres per second will shatter, fragments of meteorite and soil or rock will fly in all directions, and an impact crater will form. The Sikhote-Alin shower produced some 200 craters of this type, and it appears that such craters can range from 9 to 100 metres in diameter.

When the mass is big enough to reach the ground at a speed above about 5 kilometres (3 miles) per second, the picture changes; with a meteorite having kinetic energy over 12,000 joules (a unit of energy, one watt being the rate of work equal to one joule per second) per gram, the shock waves on impact overcome the molecular cohesion of the meteorite and of the rock, so that an explosion crater results. Not only is the meteorite shattered but much is melted or vaporized and thrown out of the crater together with the solid fragments.

If the incoming mass breaks up in the atmosphere, the principal explosion crater may be accompanied by a number of impact craters in a typical scattering pattern, the principal crater being at the forward end of an ellipse; and complete masses may be found at the rear end of the ellipse. In the Henbury craters, central Australia, the scattering ellipse runs southwest to northeast.

Characteristics of a meteoritic crater

The primary characteristics of a meteoritic explosion crater include: a more or less circular form—though the marked jointing of the country rock (the bedrock) at the best known crater, Meteor Crater, Arizona, has resulted in a "bowed-square" outline; an upturned rim, the country rock all dipping outwards; and the presence within the crater of a breccia (sharp fragments embedded in a fine-grained matrix) of partly fused rock often containing high-pressure minerals formed by the shock wave, in particular, coesite and, in the regions of most intense shock, stishovite, both polymorphs of silica; below the partly fused breccia, there is a zone of shock-fractured but unmelted rock, passing downward and outward into unaltered country rock. In the case of ancient craters, this distribution of brecciation and fracturing, together with the absence of volcanic materials, is usually accepted as diagnostic of meteoritic origin. The peculiar structures known as shatter cones are widely, though not universally, regarded as shock-wave products and also diagnostic of meteoritic impact.

Chemical composition. Until 1969, when the first rock samples were brought back from the Moon, meteorites were the only extraterrestrial material man had to analyze in the laboratory, and their chemical and mineralogical composition was of great interest on that score alone. There are good reasons to believe that the relative abundance of the elements in the common chondritic meteorites approximates their relative abundance in the solar system as a whole, except for a number of the more volatile elements (hydrogen, nitrogen, and the noble gases, and, to a lesser degree oxygen, carbon, and sulfur). The relative abundances of the elements, excluding hydrogen, carbon, nitrogen, oxygen, and the noble gases, in the common chondrites, in the Sun (as deduced from spectroscopic data), and in the Earth's crust are to be found in Table 2.

The agreement between the meteoritic and solar abundances is for the most part within the probable errors of the determinations. The low solar lithium abundance is to be attributed to its consumption in nuclear reactions, but the relatively high iron figure for the common chondrites remains a subject of debate.

Comparison with the rocks of the Earth's crust

Compared with the common chondrites, the igneous rocks of the Earth's crust are markedly depleted in magnesium, chromium, manganese, iron, cobalt, nickel, germanium, and the platinum metals (all, except magnesium, siderophile elements; *i.e.*, those that would be concentrated in the metallic core usually assumed for the Earth) and enriched in the highly lithophile elements; *i.e.*, those that would tend to concentrate in Earth's crust. These include fluorine, aluminum, titanium, strontium, barium, zinc, thallium, and uranium. The meteorites appear to be depleted relatively to the crust and to the solar abundances in the more volatile elements rubidium, cesium, zinc, indium, tellurium, lead, and bismuth, and this depletion must be taken into account in any theory of the formation and cosmic history of the meteorites.

The above data are for the mean composition of the common chondrites, which are taken as a close approximation to the mean for all meteorites. Detailed studies,

Table 2: Abundances of the Elements
(Relative to 10^6 Atoms Si, Expressed on a Logarithmic Scale [Si = 6]).

	chondrites	igneous rocks	Sun		chondrites	igneous rocks	Sun
Li	1.6	2.67	1.46	Ag	1.0	2.97	2.6
Be	1.8	1.35	0.86	Cd	2.8	1.13	1.96
B	1.8	1.45	—	In	3.0	1.0	1.66
F	3.0	3.57	—	Sn	1.8	0.23	0.04
Na	4.63	5.10	4.80	Sb	1.0	1.25	0.4(?)
Mg	5.96	4.94	5.90	Tl	0.0	3.16	—
Al	4.98	5.49	4.70	I	~2.4	1.38	—
Si	6	6	6	Cs	2.8	0.58	—
P	3.7	3.59	3.84	Ba	0.60	1.87	0.60
S	4.99	3.21	5.8	La	1.6	1.11	—
Cl	2.5	2.63	—	Ce	1.8	1.52	—
K	3.52	4.83	3.20	Pr	1.2	0.60	—
Ca	4.65	4.96	4.65	Nd	1.9	1.23	—
Sc	1.5	1.6	1.3	Sm	1.4	0.64	—
Ti	3.30	4.97	3.18	Eu	2.9	1.86	—
V	2.3	2.4	2.20	Gd	1.6	0.62	—
Cr	3.8	2.3	3.5	Tb	2.7	1.76	—
Mn	3.86	3.25	3.4	Dy	1.6	0.45	—
Fe	5.78	4.96	5.07	Ho	2.9	1.87	—
Co	3.08	1.54	3.14	Er	1.3	0.18	—
Ni	4.39	1.78	4.41	Tm	2.6	1.08	—
Cu	2.3	1.94	3.54	Yb	1.3	0.20	0.0(?)
Zn	2.1	1.74	2.90	Lu	2.5	1.63	—
Ga	1.06	1.44	0.86	Hf	1.3	0.23	—
Ge	1.3	0.19	1.8	Ta	2.3	0.09	—
As	1.17	0.43	—	W	1.0	0.06	—
Se	1.2	1.0	—	Re	2.8	2.4	—
Br	~1.7	0.6	—	Os	1.9	4.7	—
Rb	0.4	2.14	0.98	Ir	1.5	4.7	—
Sr	1.60	2.72	1.10	Pt	0.1	3.4	—
Y	0.95	1.50	0.75	Au	1.1	3.0	—
Zr	1.20	2.24	0.7(?)	Hg	1.0	2.6	—
Nb	1.9	1.42	0.4(?)	Tl	4.5	1.8	—
Mo	0.38	0.0	0.4	Pb	2.8	0.87	1.83
Ru	0.1	3.0	0.0	Bi	~3.0	1.0	—
Rh	1.4	3.0	1.3	Th	2.4	0.7	—
Pd	0.3	2.6	1.7	U	~2.0	0.2	—

Source: A.G.W. Cameron, "Abundances of the Elements," in S.P. Clark, *Handbook of Physical Constants* (1966; updated to the end of 1969).

however, show a marked uniformity in chemistry within each class of meteorite, accompanied by great divergence in some respects between the several classes.

Considering the irons first, there has long been noted a general tendency among the octahedrites (see Table 1) for the thickness of the kamacite crystals (kamacite is a characteristic meteoritic nickel–iron alloy with less than 7 percent nickel) in the Widmanstätten structure (see above *Classification*) to diminish as the nickel content increases, but it is not a very close correlation—there is a good deal of overlap in nickel content among the several groups. A more significant division is by the amounts of the trace elements gallium and germanium; over 80 percent of the siderites examined fall into one or other of six classes, each with quite narrow limits of gallium and germanium content, the remainder being widely scattered. The necessary chemical analyses have, to date, been made on only a relatively small number (about 150) of the known siderites.

Group I, with 80 to 100 parts per million (ppm) of gallium and 300 to 400 ppm germanium, consists of the 12 coarsest and coarse octahedrites.

Group II, with 40 to 65 ppm gallium and 130 to 230 ppm germanium, includes all the 19 hexahedrites and nickel-poor ataxites studied so far, together with seven coarsest to medium octahedrites, mostly with distinctly low nickel contents.

Group IIIA, with 15 to 24 ppm gallium, 32 to 55 ppm germanium, and 7.4 to 8.7 percent nickel, includes 37 coarse and medium octahedrites.

Group IIIB, with 13 to 23 ppm gallium, 27 to 50 ppm germanium, and 9.2 to 10.7 percent nickel, includes 29 medium and fine octahedrites.

Group IVA, with 1.7 to 2.2 ppm gallium and 0.09 to 0.2 ppm germanium, are all fine octahedrites, 15 in number, nickel 7.4 to 9.0 percent.

Group IVB, with less than 1 ppm gallium and less than 0.12 percent germanium, are all nickel-rich ataxites, eight in number.

The 20 percent or so of irons that do not fall into any of the above groups vary widely in structure and nickel content, and their trace-element contents show little if any correlation among themselves or with their nickel content.

The chemistry of the stony irons is summarized in Table 1; while the pallasites vary little, the mesosiderites should probably be subdivided into a class of true mesosiderites, with granular nickel–iron (about 7 percent nickel) irregularly distributed in a groundmass of hypersthene and plagioclase, and a miscellany, including falls such as that at Mount Egerton, Western Australia, in which enclaves of varied chemical and mineralogical composition are present, and Estherville (Iowa), with its one- to three-centimetre nodules of metal.

Among the stones, the achondrites can be conveniently considered first, in view of their simpler mineralogy and chemistry, largely summarized in Table 1. In each class there is little variation in chemical composition, and, if plotted on a diagram, with percentage of calcium oxide and the molecular ratio of ferrous oxide to the sum ferrous oxide–magnesium oxide as co-ordinates, the several groups form distinct fields; in particular, the eucrites and howardites, sometimes classed together as pyroxene–plagioclase achondrites, are seen to be distinct—howardites with 5% to 8% calcium oxide and the ratio of ferrous oxide to the sum of ferrous and magnesium oxides as 30% to 40%; eucrites with 8% to 14% calcium oxide and the ratio of ferrous oxide to the sum of ferrous and magnesium oxides as 48% to 66%. The one angrite has more calcium oxide (24%) and titanium dioxide (2.4%) than any other meteorite and is also high in alumina, 8.9%.

As a group, the achondrites have little in common; their lack of chondrules, rather coarse crystallization, and the rarity or absence of nickel–iron make most achondrites difficult to distinguish from certain terrestrial rocks once the characteristic fusion crust is lost.

In contrast, the chondrites show a remarkable similarity in much of their gross chemistry. Despite the great range in mineralogical composition, if typical analyses of the various chondrite groups are recalculated to give atom percentages (excluding hydrogen, oxygen, carbon, and sulfur), the between-group difference for most elements is less than the within-group differences; the exceptions are that the carbonaceous chondrites are much richer in those relatively volatile trace elements that are depleted in the common chondrites relative to their cosmic abundance and that the olivine–hypersthene chondrites and amphoterites appear to be uniformly about 4 to 5 percent poorer in iron than the olivine–bronzite, olivine–pigeonite, and carbonaceous chondrites. The enstatite chondrites, which vary considerably in iron percentage, fall into two groups; one, richer in total iron (32 to 35 percent) than any other group of stones, contains clinoenstatite and cristobalite but no feldspar and is relatively oxidized (1 to 2 percent ferrous oxide; phosphorus pentoxide partly as apatite; part of the titanium still in oxidized phases); the other, with a wide range of total iron (22 to 34 percent), contains enstatite and plagioclase and includes the most highly reduced stones (that is, stones containing the least oxygen) known.

Differences between groups of chondrites

Apart from the division into iron-rich and iron-poor series and the relative enrichment of the carbonaceous chondrites in volatile trace elements and in carbon and hydrogen, the essential difference between the groups of chondrites is a discontinuous difference in degree of oxidation, from the fully oxidized Type I carbonaceous chondrites, in which sulfur appears as sulfate, to the most reduced enstatite chondrites, in which the metal contains silicon, titanium is present in a sulfide phase, and phosphorus is present as phosphide. Apart from the carbonaceous chondrites, the degree of reduction is conveniently measured by the ratios given in Table 1. The estimates of the degree of oxidation depend on analysis of the amount of oxygen present in metallic phosphorus or sulfur compounds (phosphates to phosphides, etc.). While the proportion of oxygen combined with phosphorus ranges from maximum, as phosphate in the olivine–hypersthene chondrites and more oxidized groups, down to mainly phosphides (phosphorus only; no oxygen) in the enstatite chondrites, both phosphide and phosphate may be present in an olivine–bronzite chondrite. In each of the classes, there is appreciable variation in the proportions of iron and magnesium oxides, but there are definite gaps between classes; within the enstatite–chondrite group, decrease in the proportions of iron and magnesium oxides is correlated with an increase in the phosphide–phosphate ratio and a transfer of titanium from oxidized phases to sulfides.

Since the total iron, nickel, magnesium, and silicon contents do not vary greatly, and since nickel does not enter into the silicates to any appreciable extent (less than 50 parts per million), the increase in the ratio of ferrous to ferrous-plus-magnesium oxide from the enstatite chondrites to the olivine–pigeonite chondrites necessarily entails a decrease in the amount of pyroxene, an increase in the amount of olivine, a decrease in the amount of metal, and an increase in the nickel percentage in the metal. In some olivine–pigeonite chondrites there is so little iron in the nonoxidized phases that their sulfur appears as pentlandite, a compound of iron, nickel, and sulfur, and iron disulfide, or pyrite, instead of the nearly nickel-free troilite characteristic of the common chondrites.

Mineralogy. The principal mineral constituents of meteorites, other than carbonaceous chondrites, all occur as terrestrial minerals—olivine, ortho- and clino-pyroxene, feldspar, troilite, and the two nickel–iron alloys kamacite and taenite (terrestrial taenite is known as awaruite). There is, however, a wide range of minor constituents, many of which are not known to occur naturally on Earth; some of these can be formed only under extreme reducing conditions and are known only from enstatite chondrites, enstatite achondrites, a few bronzite chondrites, and a few peculiar siderites. Table 3 summarizes the meteoritic minerals.

Following the discovery of the so-called organized elements in carbonaceous chondrites—small spherical and ellipsoidal structures, many with an envelope of insoluble, complex hydrocarbon nature—and the assumption

Table 3: Meteoritic Minerals*

Elements and alloys:
Kamacite, (Fe,Ni), the body-centred α structure; Ni < 7%
Taenite, (Fe,Ni), the face-centred γ structure; Ni > 20%
Graphite
Sulfur†
Diamond‡
Lonsdaleite, a hexagonal polymorph of diamond‡
Copper
Gold (two doubtful records)

Halides:
Lawrencite§, $FeCl_2$

Oxides:
Chromite, $FeCr_2O_4$; often aluminian, magnesian, or vanadian
Magnetite‖, Fe_3O_4
Tridymite¶, SiO_2
Quartz♀, SiO_2
Cristobaliteδ, SiO_2
Ilmenite, $FeTiO_3$
Spinel□, $MgAl_2O_4$
Perovskite□, $CaTiO_3$

Carbides, phosphides, nitrides, and silicides:
Schreibersite, $(Fe,Ni)_3P$, tetragonal
Barringerite◇, $(Fe,Ni)_2P$, hexagonal
Cohenite, Fe_3C (cementite)
Osbornite▲, TiN
Sinoite✦, Si_2N_2O
Perryite, $Ni_5(Si,P)_2$
Moissanite⊕, SiC (carborundum)

Sulfides:
Troilite**, FeS
Pentlandite, $(Fe,Ni)_9S_8$
Daubréelite††, $FeCr_2S_4$
Oldhamite‡‡, CaS
*Niningerite*δ, (Mg,Fe,Mn)S
Blende (=sphalerite)δ, (Zn,Fe)S
Alabanditeδ, (Mn,Fe)S
*Djerfisherite*δ, $K(Cu,Na)(Fe,Ni)_{12}S_{14}$
Brezinaite, Cr_3S_4
Chalcopyrite, $CuFeS_2$
Chalcopyrrhotine, (Cu,Fe)S, cubic
Pyrite, FeS_2
Valleriite, sulfide of Cu and Fe (doubtful identification)

Silicates:
Olivine, $(Mg,Fe,Mn)_2SiO_4$
Orthopyroxene,§§ $(Mg,Fe)SiO_3$
Clinopyroxene‖‖, $(Mg,Fe,Ca)SiO_3$
Plagioclase, usually with an anorthite content between 5% and 10%
Serpentine; the principal constituent of Type II carbonaceous chondrites
Merrihueite, $(K,Na)_2(Fe,Mg)_5Si_{12}O_{30}$
Roedderite, $(Na,K)_2(Mg,Fe)_5Si_{12}O_{30}$
Yagiite, $(Na,K)_2(Mg,Fe,Al)_5(Si,Al)_{12}O_{30}$
Cosmochlore, $NaCrSi_2O_6$
Krinovite, $NaMg_2CrSi_3O_{10}$
Ringwoodite, $(Mg,Fe)_2SiO_4$ with spinel structure
Cordierite, $(Mg,Fe)_2Al_4Si_5O_{18}$

Hydrocarbons:
Much of the carbon of Type I carbonaceous chondrites is present as an insoluble complex hydrocarbon of very high molecular weight

Phosphates:
Chlorapatite, $Ca_5(PO_4)_3Cl$
Whitlockite, $(Ca,Mg)_3(PO_4)_2$
Farringtonite, $Mg_3(PO_4)_2$
Panethite, $(Na,Ca,K)(Mg,Fe,Mn)PO_4$
Brianite, $Na_2CaMg(PO_4)_2$
Stanfieldite, $Ca_4(Mg,Fe,Mn)_5(PO_4)_6$
Hydroxyapatite, $Ca_5(PO_4)_3OH$
Sarcopside, $(Fe,Mn,Ca)_7(PO_4)_4F_2$
Graftonite, $(Fe,Mn,Ca)_3(PO_4)_2$

Carbonates:
Calcite¶¶, $CaCO_3$
Dolomite♀♀, $CaMg(CO_3)_2$
Ferroan magnesite♀♀, $(Mg,Fe)CO_3$

Sulfates:
Epsomite♀♀, $MgSO_4{\cdot}7H_2O$; probably formed terrestrially from meteoritic kieserite, $MgSO_4{\cdot}H_2O$
Gypsum♀♀, $CaSO_4{\cdot}2H_2O$
Blödite♀♀, $Na_2Mg(SO_4)_2{\cdot}4H_2O$

*Those peculiar to meteorites are printed in *italics*. Within each group of minerals those commonest in meteorites are listed first. In accordance with the usual mineralogical nomenclature, a pair or more of elements replacing one another isomorphously are written in parentheses, in order of abundance, and separated by commas—*e.g.*, $(Mg,Fe,Mn)_2SiO_4$; minor and trace substituents are omitted from the formulas. †Recorded from Type II carbonaceous chondrites only. ‡Recorded from ureilites and severely shocked irons only. §Occurs in irons and stony irons, which are then very liable to rapid and destructive rusting. ‖Nickeliferous magnetite is a principal constituent of Type I carbonaceous chondrites and is an accessory in olivine–pigeonite chondrites. ¶A principal constituent of the stony iron Steinbach, and an accessory in some other stony irons and some Ca-rich achondrites and some enstatite chondrites. ♀Recorded from some Ca-rich achondrites and enstatite chondrites and from a few irons. δRecorded from enstatite achondrites only. □Recorded from some olivine–pigeonite chondrites. ◇Recorded to date (1970) only from the Ollague iron. ▲Recorded from enstatite achondrites and enstatite chondrites only. ✦Recorded from enstatite chondrites only. ⊕Doubtful; recorded from the Cañon Diablo iron only and not confirmed. **The ferrous sulfide of meteorites is normally the nonmagnetic, stoichiometric FeS, rare terrestrially, but in some metal-deficient meteorites the magnetic, metal-deficient pyrrhotine may occur. The troilite of enstatite chondrites carries nearly all the titanium of the meteorite; that of the common chondrites contains traces of titanium. ††Recorded from irons and from enstatite chondrites and achondrites. ‡‡Recorded from enstatite chondrites and achondrites only. §§Subdivided into enstatite, with $FeSiO_3$ < 12 mole %; bronzite, with $FeSiO_3$ 10 to 20 mole %; and hypersthene, with $FeSiO_3$ > 20 mole %. ‖‖A variety of clinopyroxenes occur in meteorites, including diopside, augite, pigeonite, and calcian members of the clinoenstatite–clinohypersthene series. ¶¶Recorded from carbonaceous chondrites and some achondrites. ♀♀Recorded from carbonaceous chondrites only.

Hydrocarbons in carbonaceous chondrites

that these were of biological origin and possibly fossil algae, a great deal of work has been done on carbon compounds in meteorites. The greater part of the carbon in both Type I and Type II carbonaceous chondrites is present as a compound of very high molecular weight, insoluble in all solvents and containing carbon, hydrogen, oxygen, and probably nitrogen; pyrolysis (breaking down by heat) yields a wide range of hydrocarbon compounds, both aliphatic and aromatic. Solvent extraction of the meteorite yields traces of carbon compounds of many classes, including aliphatic and aromatic hydrocarbons. The presence of these organic compounds has been advanced as evidence of extraterrestrial life, but it would appear that purely inorganic processes could have given rise to them all. Some of the organized elements have been shown to be pollen grains and other terrestrial contaminants, and many are simply mineral grains coated with the insoluble hydrocarbon mentioned above.

Structure. The nickel-poor irons, hexahedrites and nickel-poor ataxites, consist wholly (apart from small inclusions of troilite and schreibersite) of kamacite, and the hexahedrites are single crystals of this alloy. Etching a polished surface usually reveals numerous fine, closely parallel lines called Neumann lines—also termed glide-twin lamellae; glide-twinning is a type of slippage of internal structural planes. The lamellae are always parallel to some but not all of a particular set of 12 crystal planes of the metallic crystal and clearly arise from directed deformation, but whether the causal impact was extraterrestrial or was the impact with the atmosphere or the ground is unknown (see Figure 3).

Iron meteorites presumably crystallized from molten material, and at temperatures above 800° C would all have the so-called face-centred (taenite) structure; at a temperature depending on the nickel percentage the body-centred phase will segregate on cooling in plates parallel to planes of the parent taenite crystal. With few exceptions, the Widmanstätten pattern (consisting of these segregation plates; see above *Classification*) retains the same orientation in meteorites over distances up to a metre; cooling from the melt to the segregation temperature must have been slow to allow the growth of such large single crystals. This large-scale segregation pattern forms only if the composition is such that transformation begins above 650°–700° C; in nickel-rich meteorites, segregation is microscopic (resulting in the mineral plessite), as it also is in residual patches of nickel-rich alloy in interstices of the Widmanstätten pattern (see Figure 3).

The Widmanstätten pattern

The segregation process involves diffusion of nickel from the growing kamacite plates into the residual taenite, and as the temperature falls this process slows and fails to keep pace with recrystallization, so that the residual taenite lamellae develop relatively nickel-poor cores and nickel-rich rims. From microprobe studies of this profile and from the composition of residual taenite, cooling rates of the order of 1° to 10° C per 1,000,000 years have been deduced.

The structure in the stony irons called pallasites consists of five-millimetre crystals or large fragments of olivine, a common terrestrial mineral, embedded in nickel–iron; because the melting point of the olivine is well above that of the metal, these meteorites must have originated as a close-packed mass of olivine crystals floating in molten metal under a solid ceiling.

The achondrites are fairly coarsely crystallized, often brecciated, and their mineralogy and texture suggests igneous differentiation products, though the details remain obscure.

The Type I carbonaceous chondrites are not, in fact, chondritic but consist of an intimate mixture of fine-grained nickeliferous magnetite (an iron oxide) and a hydrated amorphous silicate, veined with sulfates; Type II contain variable amounts of chondrules, which are practically iron-free olivine or pyroxene (an igneous-rock-forming mineral). Some researchers believe that Type II are derived from Type I by crystallization; others that the reverse is the case, alteration by liquid water being responsible for the degradation.

The common chondrites and olivine–pigeonite chondrites vary markedly in texture: at one extreme, the chondrules are sharply bounded spherules, often readily detachable from their matrix; the pyroxene is mainly of the crystalline structure called monoclinic; little or no feldspar is present, the composition of the olivine and pyroxene varies from crystal to crystal, and clear primary glass is present; this type is referred to by some meteoriticists as unequilibrated and, later, by others as Type 3 and primitive and corresponds approximately to the spherical

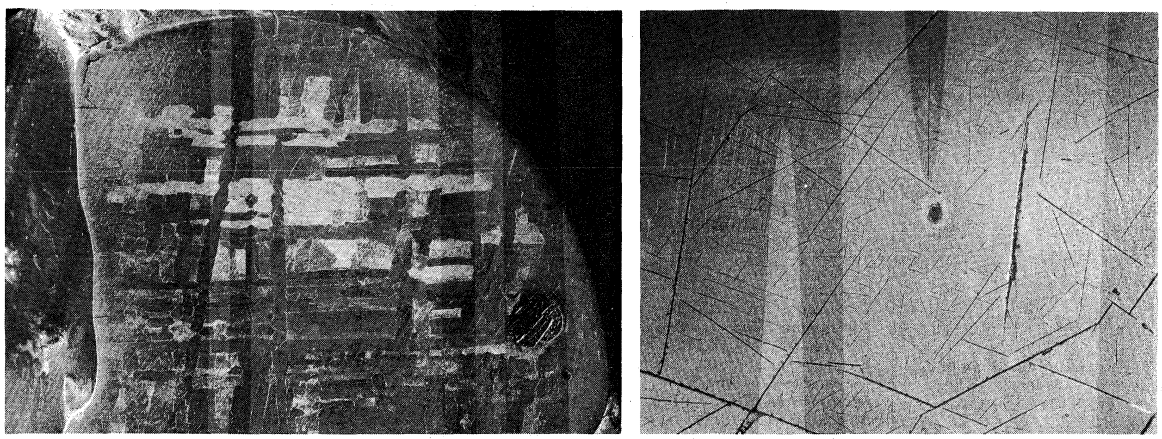


Figure 3: (Left) Widmanstätten pattern: The meteoritic iron Rowton, shown actual size, also has a heat-alteration zone near the surface. (Right) Neumann lines: Walker County; this hexahedrite has particularly well developed Neumann lines (approximately twice actual size).
By courtesy of the trustees of the British Museum (Natural History)

chondrites of the old Brezina classification (see Figure 4); at the other extreme, the chondrules are poorly defined, the pyroxene is orthorhombic, feldspar (a silicate of aluminum, calcium, and sodium) is present and glass absent, and the olivine and pyroxene are uniform in composition; this is the equilibrated, or Type 6 (two intermediate types being recognized), or recrystallized, one transitional type being recognized. The petrologic types 1 and 2 correspond to Type I and Type II carbonaceous chondrites respectively. It appears probable that the recrystallized type was derived from material approaching the primitive type by metamorphism at some stage in its history.

The five types of meteorite age

Meteorite ages. The age of a meteorite must be qualified. It can refer to the time since it arrived on Earth (terrestrial age), the time since it was an interior part of a larger body (cosmic-ray age), the time since it was last too hot to retain helium or argon (gas-retention ages), or the time since it was molten. A fifth age, derived from measurements of the products of decay of short-lived radionuclides assumed to be present in the primitive solar nebula (iodine→xenon, plutonium→xenon, lead→tellurium, palladium→silver ages), refers to the period between nucleogenesis and the formation of meteorite parent bodies. Times of the order of 10^8 years for this fifth age have been deduced, but measurements are few and their interpretation difficult.

From the time a meteorite is broken off a larger body or otherwise reduced to a fragment of less than about five tons, it is strongly irradiated throughout most of its bulk by cosmic rays, which induce various nuclear reactions; some of the nuclides produced are stable, some radioactive. When the meteorite reaches the Earth's surface, it is shielded from cosmic rays by the atmosphere; the radioactive nuclides decay, and their decay can in principle be used to determine the terrestrial age. There are complications—the mass must have been in orbit long enough for a steady state to have been reached, and in this state the concentration of each product will vary with distance from the preatmospheric surface—but these complications can be evaded, and some useful results have been obtained.

By courtesy of the trustees of the British Museum (Natural History

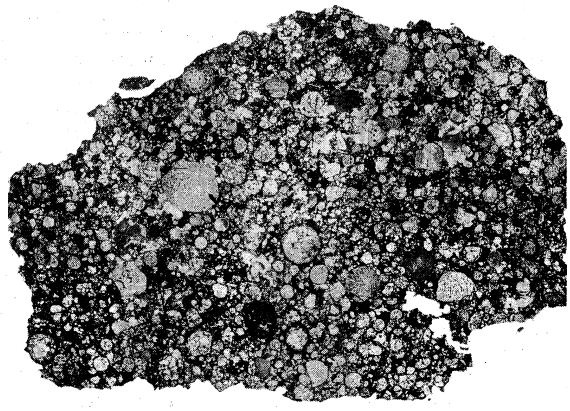


Figure 4: Chondrules in the stony meteorite Prairie Dog Creek (approximately 2¼ times actual size). The surface shown is a section through the meteorite.

Cosmic-ray ages. While a radioactive product will tend to attain a steady state under cosmic irradiation, a stable product will simply accumulate, and its amount can serve as a measure of the time during which irradiation has gone on; again there are complications, most of which can be overcome by making measurements both of a stable and of a suitable radioactive product nuclide. For chondrites, an age–frequency plot shows a very strong peak at 5×10^6 years, with a tail extending to about 40×10^6 years; for achondrites a generally similar curve, but with less pronounced peaks (at 5×10^6 and 11×10^6 years) has been obtained. For irons, much longer cosmic-ray ages have been found, with the hexahedrites, nickel-rich ataxites, and medium octahedrites forming distinct families at 10^8 to 2×10^8, 10^8 to 3×10^8, and 5×10^8 to 7×10^8 years, respectively.

Solidification ages. Solidification ages have been derived from the decay reactions of a rubidium isotope to strontium, of a rhenium isotope to osmium, or of two uranium isotopes to two isotopes of lead. Unfortunately, the chemical separation of parent and daughter nuclides, which is what the method really dates, tends to be very inefficient in meteorites, so that problems foreign to terrestrial geological dating arise. The subject cannot be briefly summarized, but its results are simple: all three methods agree in dating the solidification of both irons and stones at 4.4×10^9 to 4.6×10^9 years.

Gas-retention ages. At high temperatures, gases diffuse readily through silicate lattices; accordingly, the amount of radiogenic argon from the decay of potassium-40 or helium from uranium and thorium can furnish a measure of the time since the meteorite was last at a temperature at which gas leakage could occur. The results for irons appear to show ages much in excess of the solidification ages, but this is probably due to loss of potassium, which is readily leached out by water. For stones, both chondrites and achondrites, the ages from both methods tend to cluster around 4×10^9 to 4.5×10^9 years, indicating that cooling from solidification to gas retention was not very slow. There are, as were to be expected, many lower values, down to about 5×10^8 years, and, when the potassium–argon and uranium–helium ages do not agree, the latter is the lower—as is to be expected, since helium will be lost at a lower temperature than argon. The reheating is presumably due to collisions leading to the breakup of the parent body and to perihelion passages of the fragments.

Meteorite orbits. The only direct evidence available for the source or sources of the meteorites—Moon, minor planets, Mars, or outside the solar system—is the

orbit in which the object was travelling when it encountered the Earth. Unfortunately, observations of the atmospheric trajectories of meteorites are few and of low accuracy, especially the estimates of their velocities. For only one meteorite (Příbram, which fell in Czechoslovakia in 1959) have accurate data been obtained by photography from several observing stations, and this object had a highly elliptical orbit with perihelion (least distance from the Sun) at 0.8 astronomical unit (one astronomical unit is the mean distance between Earth and Sun, about 150,000,000 kilometres [93,000,000 miles]) and aphelion (greatest distance from the Sun) at about 4 astronomical units. Older calculations, based on very high velocities of entry, suggested hyperbolic (or open) orbits, but later studies indicate that orbits of all meteorites for which reasonable data are available are elliptical orbits of a similar general nature to that of Příbram. It also appears that most meteorites overtake the Earth and approach from directions not greatly inclined to the Earth's own motion.

Calculations of the probability of material broken off or ejected from the Moon, Mars, and the various families of comets and asteroids and of the average length of time such material may be expected to remain in orbit before being captured by the Earth have not been conclusive but tend to discount the Moon, Mars, or short-period comets as the parent body.

Origin of meteorites. For many years, meteorites and meteors (shooting stars) were taken to be essentially similar, differing only in size. Orbital evidence, however, suggests that the meteors are essentially cometary debris, while the meteorites almost certainly are not. (The fall of June 30, 1908, on the Tunguska River, Siberia, which was recorded on seismographs the world over but did not produce sizable craters, has been widely accepted as due to the head of a small comet, which exploded and disintegrated a few kilometres from the ground.) The age determinations show clearly that meteorites solidified at the same time or very shortly after the planets were formed and were broken off or ejected from a larger body relatively recently. The presence of diamond in the ureilites and some Canyon Diablo (Arizona) irons and certain details of the Widmanstätten structure have been adduced as evidence that the meteorites or some of them formed under high-pressure conditions in a body of lunar size or larger (the Moon and Mars have been favoured); but diamond can be formed by shock, and the effects of phosphorus on the iron–nickel system are rather similar to those of high pressure. On the other side, it is difficult to see how a body of lunar size could have cooled from a molten state in the time available.

Probable history of parent bodies

While there is no general consensus even on the size of the parent bodies and a wide divergence of opinion on details of these bodies' evolution, a coherent picture was beginning to emerge in the early 1970s, though it will doubtless undergo many modifications. The following scheme, one of many that cannot all be summarized here, appears satisfactory in many respects. The chondrules were formed in cool regions of the primitive solar nebula, probably by electrical discharges between dust aggregates, and cooled rapidly to a largely glassy state. The meteorite parent bodies, of medium asteroidal size, about 100–250 kilometres (about 60–150 miles) radius, aggregated cold from chondrules, metal and oxide grains, and carbonaceous material. Heat from the decay of short-lived radioactive nuclides melted a small core of the asteroid and heated a thicker zone short of complete melting; fractionation in the core produced achondrites, irons, and pallasites, while, farther out, thermal metamorphism produced the various petrological types of chondrite. A thin layer near the surface was altered by volatiles (carbon monoxide, water, and sulfur) distilled from the interior, to produce the carbonaceous chondrites, which also condensed such volatiles as zinc, cadmium, indium, mercury, tellurium, and lead distilled from the interior. These processes were virtually completed in about 10^8 years, about 4.5×10^9 years ago. Collisions between minor planets broke them up, and some of the fragments from each collision were attracted into Earth-crossing orbits. Because the life of a small body in such an orbit is fairly short, the sample of meteorites collected by the Earth shows a peak at material derived from one or more collisions about 5,000,000 years ago, material from older collisions being already largely lost. Heterogeneous meteorites such as Mount Egerton or Cumberland Falls (Kentucky) were probably

Table 4: A Selection of Historic and Other Notable Meteorites

meteorite	site	comments
Elbogen (Loket)	Bohemia	An iron of about 107 kg fell, probably about 1400; it was known for centuries as "the Bewitched Burgrave" and preserved at the castle; recognized as meteoritic 1811; medium octahedrite. 79 kg in Vienna.
Ensisheim	Alsace, France	A stone of about 127 kg fell November 16, 1492, 1130 hours; amphoterite. Specimens are in many museums and the main mass of 55¾ kg in the Rathaus, Ensisheim.
Hatford	Berkshire	At least 3 stones, totalling about 30 kg, fell April 9, 1628, 1700 hours. Though nothing is now preserved, the contemporary descriptions are so good its meteoritic nature is certain.
Krasnojarsk	Eniseisk, Siberia	A mass of about 700 kg was found about 1749; described by P.S. Pallas, 1772; pallasite. The etch pattern now known as the Widmanstätten pattern was first developed on a specimen by G. Thomson (W. Thomson, F.R.S.). Widely distributed; 515 kg in Moscow.
Lucé	Sarthe, France	A stone of 3½ kg fell September 13, 1768, 1630 hours; olivine–hypersthene chondrite. Discredited as "a stone struck by lightning," little has been preserved; 166 g in Vienna.
Siena	Tuscany, Italy	A shower of stones up to 3½ kg fell June 16, 1794, 1900 hours. Olivine–hypersthene chondrite. This fall stimulated interest in the problem of "stones from the sky." Widely distributed.
Wold Cottage	Yorkshire	A stone of about 25 kg fell October 13, 1795, 1530 hours. Olivine–hypersthene chondrite. This fall rallied English opinion to the reality of meteoritic falls. Main mass, 21 kg, in London (British Museum [Natural History]).
L'Aigle	Orne, France	A shower of stones, estimated to number 2,000–3,000 with an aggregate weight of about 37 kg, fell April 26, 1803, 1300 hours, and finally convinced the Paris Academy. Olivine–hypersthene chondrite. Widely distributed; 10½ kg in Paris.
Weston	Connecticut	A shower of several stones, together about 150 kg, fell December 14, 1807, 0630 hours; olivine–bronzite chondrite. The earliest American record. Little preserved; 15 kg in Yale University.
Cape York	West Greenland	Knives of meteoritic iron with bone handles were given to Capt. John Ross in 1818 by Eskimo. Three large masses were shown to Lieut. R.E. Peary in 1894 and transported to New York. Medium octahedrite. The mass of 31 tons in the American Museum of Natural History is the largest meteorite in any collection.
Oktibbeha County	Mississippi	A mass of 156 g found about 1854 in an Indian tumulus is the most nickel-rich meteorite known (60% Ni). Its meteoritic nature has been doubted but is established by a recent study by S.J.B. Reed.
Orgueil	Montauban, France	About 20 stones fell May 14, 1864, 2000 hours. Carbonaceous chondrite (Type I). The supposed microfossils and organic compounds in this meteorite have been the subject of extensive study and controversy. Main mass, 8.7 kg, in Paris.
Cañon Diablo	Arizona	Over 30 tons of pieces, from minute fragments to masses of 500 kg, have been collected since 1891 from around the crater now recognized as of meteoritic origin. Coarse octahedrite. Some pieces contain microscopic diamonds, probably produced by shock on impact.
Hoba	Grootfontein, South West Africa	A mass estimated at 60 tons was found in 1920; nickel-rich ataxite. The largest known meteorite. Preserved in situ.
Sikhote-Alin	Eastern Siberia	A shower of irons fell February 12, 1947, 1038 hours GMT, producing 106 small craters and impact holes; over 23 tons of material collected; coarsest octahedrite.
Norton County	Kansas	A shower of stones fell February 18, 1948, 1656 hours CST; over 100 have been recovered, including one of 1 ton—the largest known stony meteorite; enstatite achondrite. Main mass in Albuquerque, N.M.

formed by welding together of dissimilar fragments at the time of collision.

Objects formerly classified as meteorites. Tektites, small glassy fragments found in Czechoslovakia, Australia, Southeast Asia and the Malay Archipelago, West Africa, and the United States, were for a time regarded as glass meteorites, but this view is no longer tenable. There is a widely held opinion that they are formed from splashes of liquefied rock from major meteorite craters, but there are many difficulties with all present theories of origin (see further TEKTITE). The masses of nearly pure silica glass found on the surface in the eastern desert of Libya, near the Egyptian frontier, have also been supposed to be of meteoritic origin, but this is highly improbable, and their true origin also remains an enigma.

Meteorite collections. There are meteorite collections in many cities and universities, ranging from two or three specimens to large collections representing hundreds of falls. For many, catalogs have been published, and lists of such catalogs have been prepared. Some of the principal collections of the world are, with the number of falls represented at the last available count: British Museum (Natural History), London, 1,148 (1966); Field Museum of Natural History, Chicago, 835 (1946), but now probably more than 1,000; United States National Museum, Washington, D.C., 786 (1947), now probably more than 950; American Museum of Natural History, New York City, 850 (1964); Muséum d'Histoire Naturelle, Paris, 644 (1946); Naturhistorischen Hofmuseum, Vienna, 560 (1903); Arizona State University, Tempe (Nininger Collection), 536 (1950). A few of the most notable meteorites are listed in Table 4.

BIBLIOGRAPHY. Progress in the study of meteorites is so rapid that books are soon partly outdated, but the following are still useful: F. HEIDE, *Kleine Meteoritenkunde*, 2nd ed. (1957; Eng. trans., *Meteorites*, 1964); E.L. KRINOV, *Giant Meteorites* (Eng. trans. 1966); B.H. MASON, *Meteorites* (1962); B.M. MIDDLEHURST and G.P. KUIPER (eds.), *The Moon, Meteorites, and Comets* (1963); and J.A. WOOD, *Meteorites and the Origin of Planets* (1968). A good book for non-specialists, with excellent accounts of phenomena of fall, is by H.H. NININGER, *Out of the Sky* (1952). Scientific papers on meteorites are scattered through many journals, but two, *Meteoritika* (Russian) and *Meteoritics*, are devoted wholly to this subject. A record of known meteorites, with a summary of the principal literature on each fall, is provided by the *Catalogue of Meteorites*, published by the British Museum (Natural History), 3rd ed. rev. by M.H. HEY (1966).

(M.H.H.)

Meteorite Craters

Meteorite craters are depressions that result from the impact of a meteorite with any planetary body. Prior to the advent of space travel, meteorite craters were considered rare curiosities on Earth but were presumed to be commonplace on the Moon's surface. Some circular depressions associated with meteorites—such as Meteor Crater, Arizona; Odessa, Texas; four craters in Australia; Wabar in Arabia; Kaalijarv craters in Estonia; and the Campo del Cielo crater swarm in Argentina—were accepted as representative of occasional collisions between large iron meteorites and the Earth's surface. A few classic papers, known mainly to specialists, summarize the extent of scientific investigations up to the mid-1950s. Textbooks usually ignored the phenomenon of meteorite impact as a terrestrial process or treated it only briefly.

Since the launching of Sputnik in 1957, interest in and study of impacts between interplanetary materials (meteorites, micrometeorites, comets) and planetary surfaces have grown in proportion to the exploration of the planets made possible by the space probes. Ranger, Orbiter, Surveyor, and Apollo spacecraft have extended the estimate of the number of lunar craters from about 300,000 observable by telescope (at a 1.5-kilometre-diameter resolution) to roughly 3,000,000,000,000 with diameters greater than one metre. This includes those craters visible for the first time on the far side of the Moon. The space probe observations also allow the lunar craters to be classified by size and morphological characteristics and, at least inferentially, by modes of origin. Thus, most lunar scientists now recognize primary and secondary impact craters, volcanic calderas, maars, summit craters, and drainage or collapse craters as probable genetic types on the lunar surface.

The discovery of circular structures and actual depressions on the surface of Mars by Mariners 4, 6, and 7 suggests that cratering may be a universal natural process on terrestrial-like planetary bodies. Whether this conclusion can be extrapolated to impact cratering must await determination of the origin of Martian craters by later probes or direct landings; these may in fact be largely (or exclusively) volcanic or may be formed by still unspecified processes. It is reasonable, however, to assume that impacts occur frequently on all unprotected planetary surfaces in the solar system and, by extrapolation, on any similar bodies elsewhere in the universe. Therefore, Mercury is likely to display craters but Venus, with its gaseous envelope, would have few, if any, such structures. In general, the number of craters per unit of surface area correlates closely with the nature and density of the planetary atmosphere and its capability for erosion, as well as with the flux of bombarding particles over time in a planet's neighbourhood.

Most hypotheses for the origin of terrestrial (solid rock) planets assume a period of accretion of spherical and fragmented bodies of sizes varying from dust to planetesimals of multikilometre dimensions. It is probable, therefore, that collisions with energies sufficient to generate shock waves have occurred over much of the growth history of these planets. Many meteorites contain signs of shock damage, mostly inflicted during breakup of the parent body or during final impact with the Earth, but some aspects conceivably are related to the early accretionary process. It is impossible at present to state whether major, crater-producing impacts take place mainly after a planet approaches its final dimensions or as a continuing process while a protoplanet reaches a critical size capable of withstanding further large collisions. A partial answer may emerge when the relative ages of the cratered lunar surface are given more absolute values on the basis of radiometric dating of crustal rocks.

Numbers of craters on Earth and their recognition

Impact craters will necessarily be less frequent on the Earth than on the Moon. Friction from the Earth's atmosphere effectively burns up most of the smaller bodies entering it from outer space. Thus, any craters formed will tend to be larger than the average size of all entering meteorites. Over the last eon (1×10^9 years) meteorites large enough to survive the plunge through the air and then excavate big craters appear to have been too few to produce the expected population of impact structures on the continents. The combined effects of the hydrosphere and atmosphere in eroding the land surface since cooling of the crust and formation of oceans evidently have erased most meteorite craters between ten metres and about one to two kilometres. If such craters were more abundant in the first 2–3 $\times 10^9$ years of terrestrial history, owing to a greater meteorite flux in the early solar system, their traces have long since been removed by erosion, deposition, and metamorphism.

Since 1955, the list of definite and probable impact structures has tripled from the 20 to 27 proposed by investigators up until then. Another 50 structures that fall into the "suspicious" category have not been carefully checked in the field. The chief reasons for the sharp increase in numbers of identified craters have been formulation of the astrobleme concept, more intensive searching, and development of new criteria for recognizing impact structures.

The astrobleme was an important advance in the understanding of what to look for in relation to meteorite craters. An astrobleme (star wound) is, then, simply the scar of an impact crater—that is, any part of the structure that has survived erosion. Older structures may completely lack discernible crater rims but still retain some degree of circularity, have distinctive internal deformation features, or contain remnants of original rock fragments (breccia) shocked in the impact event. Thus, in the broader sense, a meteorite crater can be defined as any structure now at or near the surface containing evidence of a shock-producing impact in which the disturbance by deformation

and fragmentation of rock or soil is generally circular in outline (plan) and dies out rapidly in intensity upon reaching diameter-to-depth ratios of three to one or greater (Figure 1).

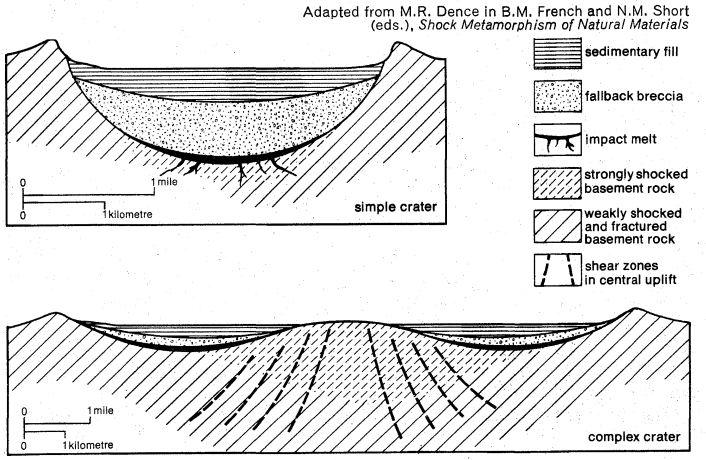


Figure 1: Major types of impact craters. Those larger than about six kilometres tend to be complex.

Once it was realized that astroblemes as well as actual crater depressions were clues to past meteorite impacts, the accelerated search given impetus by the space program could focus on any unusual structure meeting the general requirements of the above definition. A Canadian group at the Dominion Observatory in Ottawa led the way. These workers systematically examined aerial photos for circular features over all of Canada. The Precambrian shield of crystalline rocks, which outcrops over 50 percent of Canada, proved to be an excellent place to count impacts because of the improved probability of survival of larger structures in the shield terrane, where erosion rates have been low during the last 500,000,000 years (Figure 2). The group then applied geophysical methods, such as gravity and seismic surveys, to some of these structures. Several shallow-based disturbances were core drilled to provide samples of subsurface materials that exhibited evidence of shock pressure damage. Thus, from 1951, when the New Quebec (Chubb) Crater was first proposed as an impact crater, to the present, the number of probable meteorite craters in Canada has risen from 1 to 17. The Sudbury Basin can now be included among these; this structure, elongated by unrelated tectonic compression, is the source of over half the world's supply of nickel. Similar search techniques in other countries during the 1960s provided identification of many structural disturbances as rimmed impact craters (*e.g.*, Wolf Creek in Australia) and astroblemes (*e.g.*, Vredefort in South Africa).

Relation of numbers of craters and geological processes

The actual number of meteorite craters discovered on Earth is far below the numbers expected from calculations based on meteorite flux data and relations between meteorite mass and size of crater. Using data from several published estimates, there would be approximately 300 impact craters 2.5 kilometres and larger formed on the Canadian Shield in the last 10^9 years. The present value of 17 impact structures with diameters greater than 2.5 kilometres found to date affords a crude measure of the effectiveness of erosion in this ancient, stable area. Worldwide, the present land surface area would have been subjected to 130,000 impacts producing craters one kilometre or larger during the last 10^9 years. The gross discrepancy between that number and the 50–100 impact structures now known must be accounted for both by erosion and by the masking effect of younger sediments and metamorphism of older terranes.

Examination of meteorite craters in the field and studies of their rocks by laboratory techniques now provides new sets of criteria that serve as unique, unequivocal indicators of an impact event. Appropriate combinations of these criteria have never been found in craters or structures that are conclusively associated with volcanic or solution collapse origins. The criteria are direct and singular consequences of the impact process and thus are definitive in identifying meteorite craters as distinct from other crater types.

This article treats the processes associated with meteoritic impact, the characteristics of craters, and the effects of cratering on the Earth. For information on the nature and composition of meteorites, see METEORITE; for more detailed discussion of craters elsewhere in the solar system, see MOON, and MARS; for further information on shock processes and their effects on rocks, see WATER WAVES; ROCKS, PHYSICAL PROPERTIES OF; and GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES.

THE IMPACT CRATERING PROCESS

Another noteworthy advance in meteoritics since 1957 results from theoretical and experimental investigations of the cratering process. This includes field studies of nuclear and chemical explosions and of rocket missile impacts, laboratory-scale simulations of impacts using gun-impelled high-velocity projectiles, and the development of physical models of the sequence of crater formation using sophisticated computer codes. In addition, rocks and minerals can be subjected to shock effects under controlled conditions in crater explosions or laboratory experiments that reproduce most characteristics of shock metamorphism present in natural impact crater rocks.

A general model for impact cratering mechanics based

By courtesy of the Earth Physics Branch, Department of Energy, Mines and Resources, Ottawa, Ontario

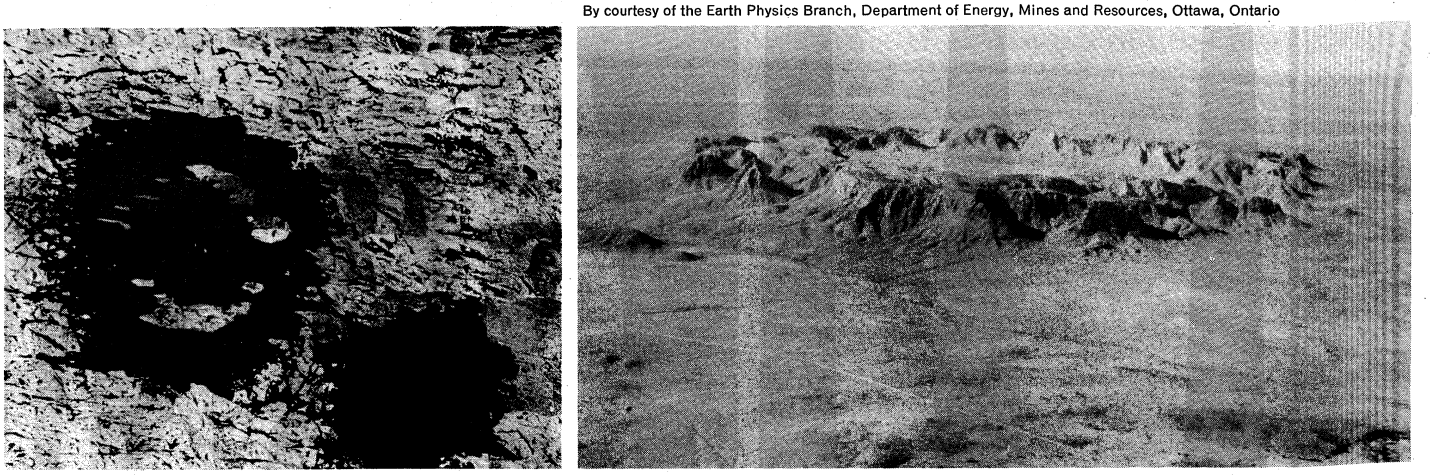


Figure 2: (Left) Clearwater Lakes impact structures in northern Quebec. West Clearwater Lake (left) is 30 kilometres wide; the ring of islands within it is part of a central uplift complex. The barren, jointed Precambrian granitic terrane is filled with small, linear lakes. (Right) Gosse's Bluff, a circular rim of hills composed of nearly vertical, intensely deformed sedimentary rocks. The inner rim shown here is only part of a 4.5-kilometre-wide central uplift containing many signs of shock damage, including shatter cones, which is part of a larger (19-kilometre-diameter) impact structure.

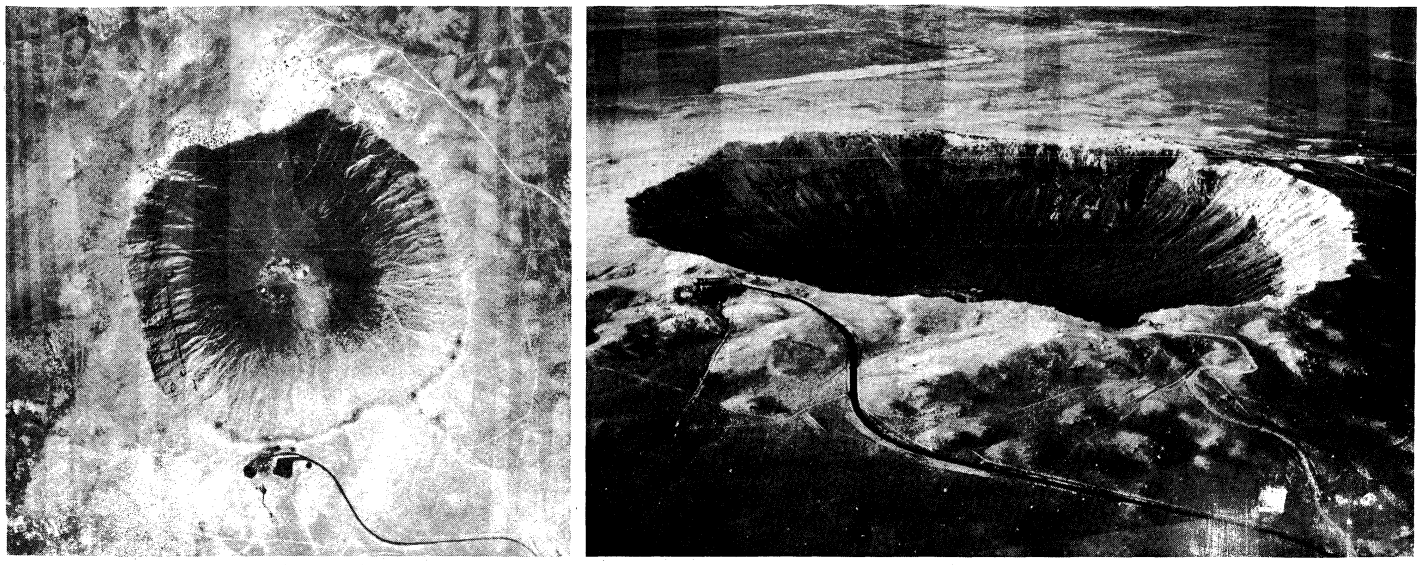


Figure 3: *Meteor Crater, Arizona.*
(Left) Vertical view of the 1.2-kilometre-wide crater showing its square outline (departure from circularity is controlled in part by joints in the sedimentary units). The hummocky, irregular surface built from ejecta deposits extends outward (note lighter patches) at least a crater diameter. (Right) Oblique view from the northwest into the crater. The limestone and sandstone beds below the rim have been severely faulted, overturned, and uplifted.
By courtesy of John S. Shelton

on studies of Meteor Crater, Arizona (Figure 3), requires a basic understanding of the nature of shock waves.

Nature of shock waves

A shock wave is a transient pulse, moving at velocities above that of elastic (seismic or sonic) waves in the same material or medium, which produces an almost instantaneous rise in pressure behind its advancing front. Ahead of this pressure rise, the medium is not yet affected in any way; behind it the medium is immediately compressed with a resultant decrease in volume (increase in density) until unloaded by decompression or rarefaction waves. Pressure magnitudes in shock waves exceed the dynamic elastic limits of the transmitting materials; for rocks, this limit falls between 20 and 100 kilobars (1,000 bars, or approximately 14,700 pounds per square inch, or 10^9 dynes per square centimetre). Nuclear explosions and impacts will generate pressure waves whose initial amplitudes can reach megabars in the materials undergoing shock. An important property of shock waves involves a possible change of phase upon meeting a free surface or another medium of differing density and wave propagating characteristics. Commonly, an initial compression wave (wave peaks are close together) will thus be converted to a rarefaction wave (farther apart), which, if reflected over the previous wave path, can unload the state of compression or place the medium into tension (stretching or tearing). As a shock wave diverges spherically from a point source, its expansion into an ever larger volume also reduces its magnitude so that the pressure at the moving front continually decreases with increasing radial distance from its source.

Hugoniot curves

Each material substance responds characteristically to shock waves of a given magnitude. This response is measured as a relative decrease in volume or increase in density. One specific change in volume or density results for each value of peak shock pressure. A plot of the different pressures attained and their corresponding changes in volume (density) produces a curve referred to as the Hugoniot. This represents an equation of state for the material under various shock loads, in which the total mass, momentum, and energy of the system are conserved during the shock event. These parameters, together with the velocities of the shock front and particles in motion behind it, are related by a series of conservation equations that allow calculation of pressures and volume changes if initial density and shock and particle velocities are experimentally measured. Once the Hugoniots are determined, one can calculate the partition of energy (appearing as heat, mechanical motion, etc.) in each material as shocks progress through interacting media.

Formation of Meteor Crater. According to E.M. Shoemaker's analysis of the formation of Meteor Crater, Arizona, an incoming iron meteorite (bolide) struck the layered rock units on the Colorado Plateau near Winslow, Arizona, some 15,000–40,000 years ago. The orbital velocity of this meteorite in space about the Sun could have been between 12–42 kilometres per second, to which the effect of Earth's velocity of 29.8 kilometres per second might be added if the meteorite were moving against it. Deceleration by the atmosphere reduced terminal velocity, but if it exceeded one to two kilometres per second an impact crater would result (a penetration crater is produced by meteorites reaching the surface in the range of one to two kilometres per second). In his calculations, Shoemaker used 15 kilometres per second as the terminal velocity.

Mass, energy, and impact of the meteorite

By assuming an equivalence between a shallow-buried nuclear explosion and the impact at Meteor Crater, the upward scaling of energy relations for nuclear cratering explosions to the dimensions of Meteor Crater (1.2 kilometres diameter) indicates that a total energy of 1.7 megatons (7.1×10^{22} ergs) was needed to excavate the volume of nearly 0.3 cubic kilometre at Meteor Crater (estimates as high as 17,500,000 megatons, or 7.2×10^{29} ergs, have been made for the 62-kilometre-wide Manicouagan structure in Quebec). From this, and from the assumed velocity of 15 kilometres per second, the kinetic energy of the meteorite can be calculated. This fixes its mass at 63,000 metric tons that, with a density of 7.85 grams per cubic centimetre for iron, would occupy a sphere of about 25 metres diameter. More recent estimates place the total energy as high as three to five megatons, which requires either a larger projectile or higher final velocity.

The sequence of events at Meteor Crater is depicted schematically in Figure 4. Immediately upon striking the ground, shock waves are imparted both to the rocks and the meteorite itself. The ground-traversing shock front outruns the penetrating meteorite and forms an expanding cavity. The rocks behind the front are strongly compressed and generally are set in radial motion outward from the region of penetration. Depending on the Hugoniots of the rocks present, varying fractions of this target material will be vaporized, melted, crushed, fragmented, or fractured; ultimately, different sections of the volume of rock actually excavated will be dispersed and mixed together. The shock generated in the meteorite at the interface with the ground moves back into it as a compression that, upon reaching boundary surfaces along the sides and rear, transforms to a rarefaction that reflects

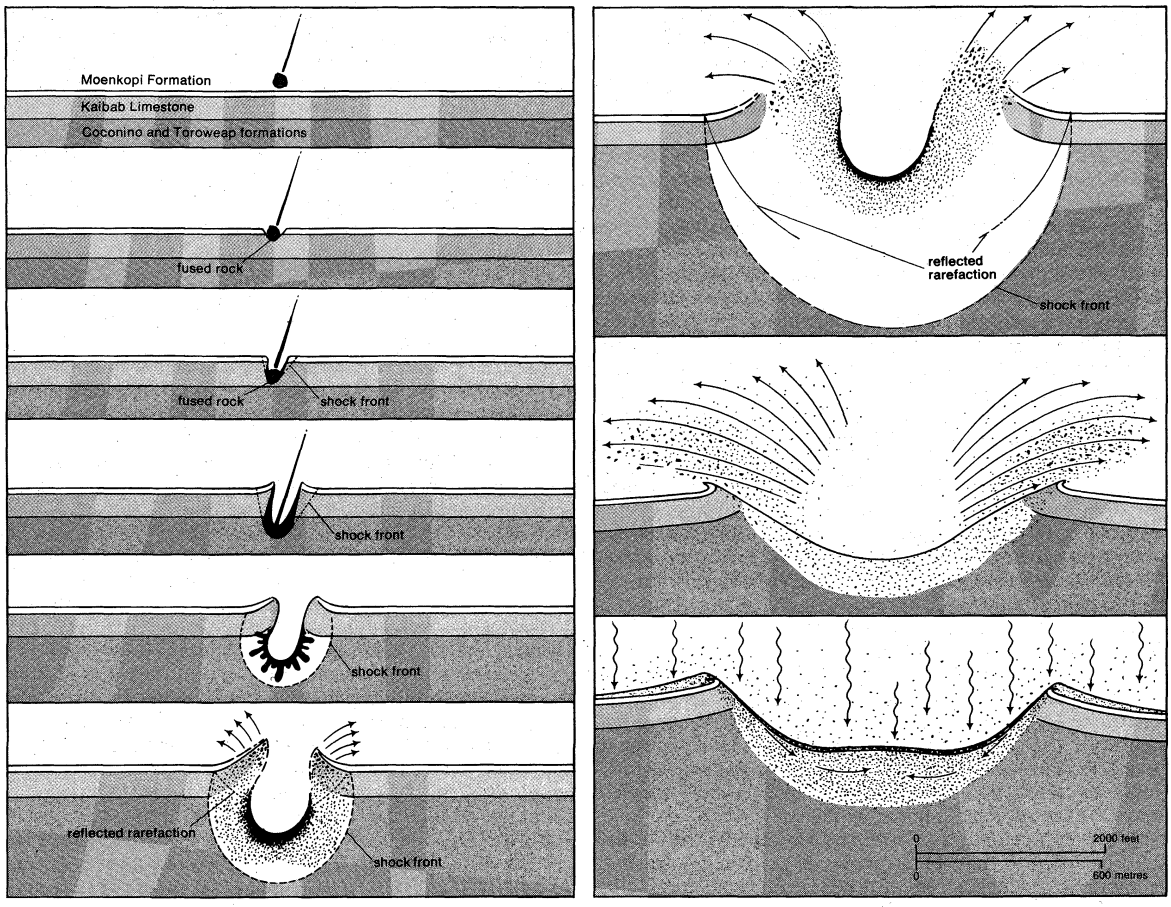


Figure 4: Stages in the formation of Meteor Crater, Arizona.
After E.M. Shoemaker in B.M. Middlehurst and G.P. Kuiper (eds.), *The Solar System*, vol. 4 (1963); University of Chicago Press

into the meteorite. Thus, some of the meteorite is vaporized, whereas other parts are melted or intensely heated and mechanically damaged as the entire body breaks up. A fraction of this meteorite is smeared along with the outward moving rock mass; much of the meteorite, however, ejects as fragments that escape the eventual crater.

The initially intense shock wave attenuates as it diverges outward. The net effect is that the degree of shock damage attributable to compression diminishes with increasing radial distance from the line of penetration. Only a small proportion of the total volume excavated undergoes strong to intense shock pressures; in contrast, the bulk of the rock experiences little or no permanent damage.

Excavation and brecciation of material

The actual excavation process requires rarefaction waves. These form as the advancing compression waves encounter free surfaces (ground–air interface) or discontinuities such as stratigraphic bedding surfaces, joints (rock fractures), or units with varying physical properties because of varying rock types. The rarefactions effectively place the rock medium under tension, resulting in a general disruption of the units. Because these now fragmented pieces were already in motion, they retain momentum sufficient to carry many of them forward and out of the developing crater as ejecta that follow trajectories established by their initial movements. Most such fragments move from the crater at low angles, and these are termed throwout deposits; at high angles the fragments fall back as crater-filling breccia deposits. Some, however, are carried downward without leaving to form fragmental debris that concentrates along the crater base. This base is defined by the boundary with the rupture zone, consisting of broken and displaced but not fragmented and ejected rock, that underlies the breccia fill. This boundary also represents the surface below which the compression and rarefaction waves become too weak to efficiently excavate material.

Within the rupture zone below the true crater floor and walls, the rocks can be folded, faulted, and intricately displaced. Especially in the rim, which marks the outer limit of surface excavation, layered units may be folded over completely to form structures that rival in complexity some of the great nappes (rock masses translated by thrust faulting) of the Alps. The overturning results when adjacent units below the uplifting free surface are peeled back as they go into motion in close succession. Massive rocks do not fold back but tend to uplift along joints, so that any pre-existing foliation is distorted. However, all crater-related deformation decreases abruptly away from the breccia-rupture zone contact, so that drilling beyond this contact would rapidly pass out of shocked or disturbed rock into ordinary, unmodified country rock.

For small craters, such as Meteor Crater, the entire formational process requires no more than a few seconds to completion. Craters that are tens of kilometres wide probably form in one to five minutes at most. The excavation caused by shock action occurs more rapidly than the later infilling process from fallback and slumping.

Conceptual crater model

Experimental studies. Experimental studies have added several details to the conceptual model. In the initial compression stage, materials behind the shock front experience pressures so high (many hundreds of kilobars) that their dynamic compressive strengths are greatly exceeded. Rocks behave under these conditions much like fluids (hydrodynamic state). The first rarefactions develop along the target-projectile interface and later the projectile itself. Decompression leads to the ejection of fluidized rock as a hot jet that, according to some workers, accounts for the glassy objects called tektites. As the spherical shock front advances outward, the compressed shell of material behind it mixes with the projectile (meteorite) and ejects steadily increasing amounts to begin the excavation stage. The complex geometric patterns of particle movements that follow progressive relaxation by rarefaction waves results in the lateral deflections of flowing materials from their earlier radial paths. Below the line of impact, particle movements still proceed mainly downward, but they deviate increasingly sideward until a tangential flow from the crater at low angles dominates the region of the ex-

panding rim. Ejection and crater growth is a continuous, orderly process involving a steady flow of materials along shear lines. Cratering ceases when stresses become too weak to break up materials or when momentum is insufficient to carry fragments beyond the excavation limits.

CRITERIA OF IMPACT ORIGIN

Crater geometry. The final shape of the crater depends on many factors. The depth to the visible crater floor (apparent crater boundary), relative to rim height, varies with the amount of material that backfills the crater as well as the degree of uplift or overturn of the rim. It has been suggested that the bulk of ejecta escape from the crater, but actual impact structures seemingly retain substantial fractions (up to 75 percent) of the initial excavated volume as breccia and melt deposits. Some of this is direct fallback or materials that fail to move out. A post-impact modification stage, however, begins even as cratering is being completed, and it continues for days to years afterward. Instability of steep crater walls leads to slumping and infill that, in effect, enlarges the rim-to-rim diameter. Erosion and inward drainage may bring some of the throw-out ejecta back into the crater. Isostatic adjustments, the regional response to reduction of crustal load at the crater site, can affect larger craters by pushing the entire structure upward.

Simple and complex craters

A simple crater is characterized by a well-defined lip, a bowl-shaped cross section, extensive fallback deposits that thicken toward the centre, and small to moderate amounts (1 to 2 percent) of impact melt concentrated along the central base. Its initial diameter to depth ratio is about five to one. When the crater diameter begins to exceed four to five kilometres, the rupture zone beneath the central base tends to uplift along shear lines. Large, complex craters have broad, domelike uplifts, which sometimes have topographic expression as central peaks. Rocks within this uplifted portion may be strongly shocked and intricately contorted. The mechanism of uplift is still conjectural: elastic rebound, isostatic adjustment, response to rarefaction waves reflected from deep crustal layers, and slump-induced rise of rubble are suggested mechanisms. Normally, in layered materials the uplift contains stratigraphic units from depths below the general level of the crater base, which indicates a real upward movement of rock. It has been proposed that large craters with diameter–depth ratios around ten to one and with central uplifts result from comet rather than meteorite impact. When a low-density cometary projectile strikes the ground, the dominant energy release remains near the surface; this causes lateral excavation to be more effective than is true of penetrating iron meteorites. Generally, complex craters develop poorly defined, relatively low rims and contain their breccia-melt deposits in an annular depression between rim and central peak.

Geophysical anomalies

Geophysical techniques are frequently applied to circular structures to determine their nature and origin. Commonly, impact craters and astroblemes show slight to strong negative gravity anomalies, because of brecciation and the creation of small cavities; but measurements of rock densities, compensation for possible infilling sediments, and topographic corrections are needed properly to assess the magnitudes and implications of such anomalies. Volcanic calderas, maars and diatremes, and even eroded volcanic necks and pipes may have comparable anomalies. Magnetic highs might suggest the presence of iron meteorite materials within the crater fill, but usually most of the projectile disperses outside the structure upon impact. Seismic reflection techniques can pick out the crater base contact and also may sense the local character of structural disturbance.

Shock metamorphism of rocks. Although shape, deformation, and geophysical anomalies are all useful criteria for locating meteorite craters, they do not, taken singly or together, comprise unambiguous proof of an impact origin. Some structures of nonimpact origin possess many features in common with meteorite craters. Astroblemes, in particular, may lack distinguishing surficial characteristics. The only truly definitive criteria so far advanced reside in the rocks themselves. Thus, rocks of any type are subject to remarkable changes during impact that, collectively, are described as shock metamorphism. The passage of transient pressure waves through rocks and their constituent minerals induces distinctive effects of very high pressures and temperatures and rapid strain rates that alter their composition and properties. These effects include the formation of high-pressure polymorphs (different forms of the same composition) that normally are not stable in the Earth's upper crust; fragmented, fractured and granulated rock masses; unusual intracrystalline markings; solid-state transformations from crystalline to glassy states; and selective melting and vesiculation of individual minerals. These changes are produced in microseconds to milliseconds by pressures ranging up to 500–750 kilobars (comparable to pressures at depths of 1,400–1,800 kilometres within the earth), temperatures that rise above 2,000° C (3,600° F), and strain rates millions of times faster than those operating during mountain building.

A clear picture of progressive shock metamorphism of rocks emerges from following the changes in granite, a common igneous rock containing quartz, feldspars and micas, as it undergoes increasing shock pressures. The effects of a nuclear explosion on granite surrounding the explosive device provide an example. In a contained (noncratering) explosion, pressures decrease systematically from the explosion centre as the waves diverge spherically. For a ten-kiloton explosion, maximum pressures 60 to 100 metres from the centre will be only five to two kilobars. No perceptible changes occur at these low pressures except where compressive waves encounter discontinuities that transform them into fracture-producing rarefactions. As peak pressures rise with ever-closer proximity to the explosion, microfractures generated in individual grains continue to increase in frequency and numbers per grain (Figure 5). At 20–30 kilobars, cleavage planes in mica start to bend and kink as though folded like layered strata; with higher pressures more individual mica crystals start to distort, and the intensity of deformation rises in all flakes thus affected. When the shock pressures in the passing waves exceed values of 50–75 kilobars (the dynamic elastic limit for many granitic rocks), changes not known even in equivalent rock types buried deep within the crust start to appear.

Shatter cones and mineral transformations

One such change is the development of shatter cones. These are peculiar external fracture surfaces in which close-spaced "grooves" seemingly radiate outward from the apex of a cone (horsetailing effect). Many such cones usually cluster together. Most have apices that point in a common direction that, when separate cone-bearing layers are rotated to their original position by later analysis, converge toward a single origin that appropriately coincides with the explosion centre or region of first impact.

In quartz grains, sets of lamellae-like planes form within crystalline zones over a pressure range from about 100 to 400 kilobars. These planar features are now known to be close-spaced bands (0.2–1.0 micrometre wide) of disordered crystalline material (the regular orientation of atomic units in a crystal lattice structure becomes distorted and bonds are bent and broken; the mineral substance is thus changed to the isotropic or glassy equivalent of the host grain). Planar feature sets tend to align along various atomic planes whose crystal indices are unusual or rare as crystal faces or mechanical slip zones in the corresponding unshocked mineral species. At higher pressures, as many as 5–15 distinct sets, following several different crystallographic orientations, appear to occupy all available space within a grain, although electron microscopy demonstrates some crystalline material to survive between planar discontinuities. Planar features presumably result from crushing or slipping of crystalline phases along preferred planes of weakness, as the grains are compressed rapidly into much smaller volumes.

In the higher pressure ranges over which quartz planar features are produced, coesite and stishovite, high-pressure forms of silica, will develop, presumably in association with the process of isotropization. Other shock-induced polymorphic transformations, such as diamond from other forms of carbon, have been discovered.

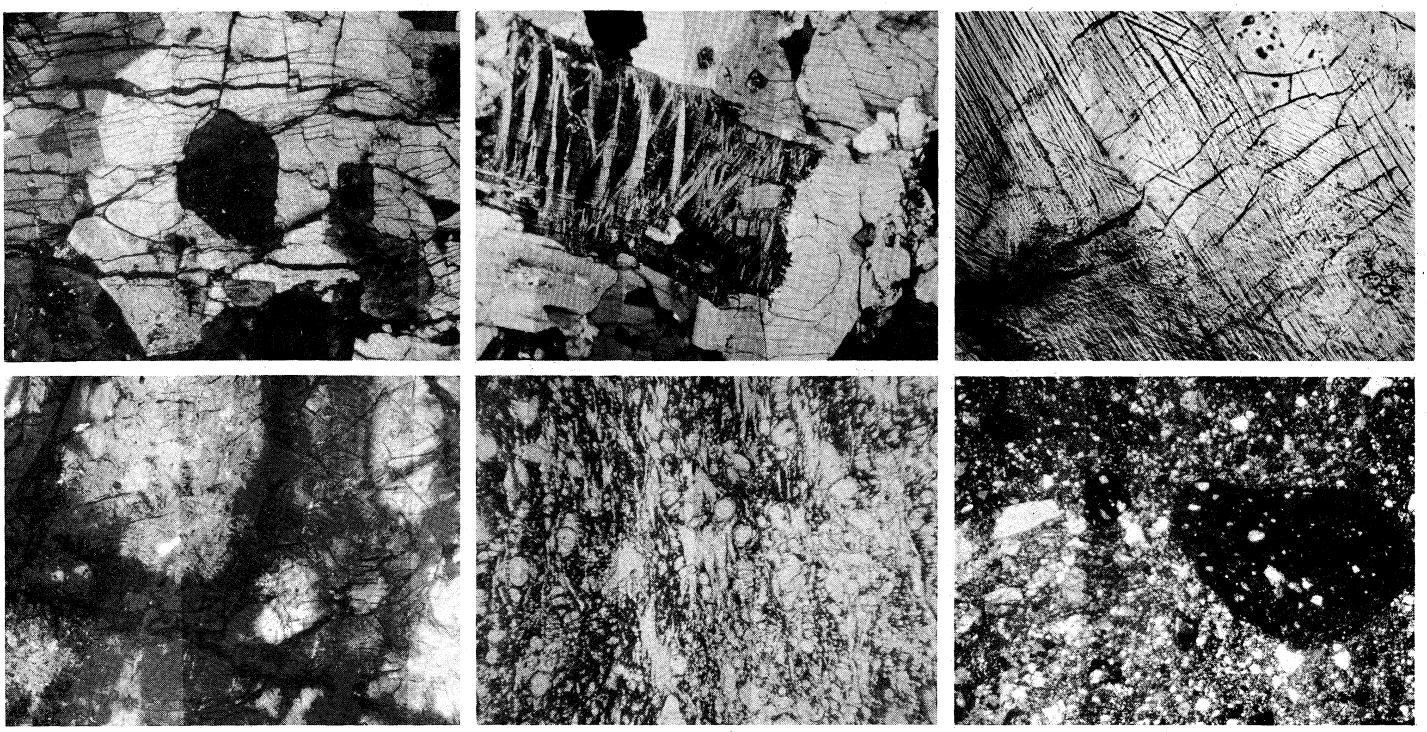


Figure 5: (Top left) Shock-induced fractures in a granite rock. (Top centre) Biotite mica flakes kinked by the action of shock waves. (Top right) Several sets of close-spaced planar features in a single grain of quartz in a shocked granite specimen. (Bottom left) Crystals of feldspar and perhaps quartz undergoing transformation from the crystalline state (white, light gray) to the disordered noncrystalline state (isotropic, dark gray) in a granite sample. (Bottom centre) Intensely shocked granite in which the rock has more or less melted (note some flow alignment) and is strongly vesiculated. (Bottom right) A "suevite" from the Brent impact structure, Canada; its texture is characterized by a mass of fragments broken by shock waves and containing various signs of shock metamorphism, mixed with glass (dark mass on right) and granulated material. Magnified about (top left, top centre, bottom left) 360 ×, (top right, bottom centre) 1,800 ×, (bottom right) 48 ×.
By courtesy of (all but bottom right) N.M. Short, (bottom right) Earth Physics Branch, Department of Energy, Mines and Resources, Ottawa, Ontario

Both feldspar and quartz within a shocked granite may contain planar features but some minerals have crystal structures that seemingly prevent formation of these discontinuities. At pressures around 350 kilobars (feldspar) and 425 kilobars (quartz), the density of planar features within grains reaches a maximum, and the grains become more or less completely isotropized. Thus, a grain may convert to a glassy state while retaining its original shape. These thetomorphs (Greek "adopted form") are unique to shocked rocks. Hand specimens containing thetomorphs may still display the original gross texture but will have noticeably reduced specific gravities.

Melting at high temperatures and pressures

At still higher pressures (450–600 kilobars), the effect of rapidly rising temperature from conversion of mechanical to thermal energy leads first to incipient melting of individual grains and then, eventually, to general melting and variable mixing of the whole rock material. The tectosilicates (quartz, feldspar) readily undergo such melting, accompanied by progressive vesiculation within individual grains. Mineral groups like olivines, pyroxenes, and micas (*qq.v.*) apparently degenerate structurally without actually melting until the remainder of the rock becomes superheated melt. Refractory minerals, like zircon and titanite, may partially decompose in the melt, trailing particles within the fluid.

When these shocked rocks are viewed in the microscope as thin sections, their distinctive character is immediately evident. The "suevite" texture, defined by rocks from the Ries Crater in Bavaria, consists of intensely shocked mineral and rock fragments mixed with moderately to imperceptibly shocked fragments held in a granulated matrix, often with blebs (small, irregular droplets) of melt. The identity of most constituents as nonvolcanic country rock (unless the impact happened in volcanics) distinguishes this impact material from most typical volcanic or tectonic breccias.

Other rock properties are drastically altered by shock. The indices of refraction of mineral components can be anomalously lowered, sometimes to values below those for any unshocked equivalents. X-ray analysis of individual grains may reveal that an apparently unstrained single crystal has been shattered to a fine aggregate, which produces a set of rings instead of spots in the diffraction pattern. Argon, derived from decay of radioactive potassium in feldspars or mica, can be driven off by the shock and attendant temperature rises, so that a rock may acquire a new age that reveals the time of impact.

If the deeply buried nuclear explosion in granite were replaced by a shallower explosion, the resulting crater should produce ejecta, showing a wide range of shock damage, which would mix on falling back to fill the excavated cavity. This crater would simulate, in many respects, the form, deformation style, and shock features that are the marks of an impact structure. Such an artificial crater, if recent, would differ mainly in the relative absence of postimpact alterations; natural impact craters generally contain tightly cemented breccia deposits characterized by notable geochemical changes.

Identification of circular structures. The various criteria for recognizing impact structures, as reviewed in this article, are as follows. On Earth, a circular ring, depression, or breccia zone is confirmed as a meteorite impact scar if it meets at least some of these definitive criteria. On the Moon and Mars, direct field investigation, including core drilling, is obviously still limited, so that proper identification of the literally billions of circular structures on these planets is presently impossible. Some of these structures are of one genetic type and differ mainly in their relative aging by erosive processes. Others appear to have nonimpact origins.

It is now clear from Orbiter, Surveyor, and Apollo investigations that volcanic rocks abound on the lunar surface. This implicitly suggests that volcanic cones, calderas, and maars should be interspersed with impact structures on the Moon. Two large circular structures in the inner ring of Mare Orientale, of apparently similar age, have been

proposed as the prototypes for impact and volcanic craters. The impact crater is defined by multiterraced walls, a rough floor, central peaks, a wide, hummocky surface beyond the rim, and ejecta rays extending out many crater radii. The volcanic caldera type shows a smoother floor, no central peaks (although this is possible), an unterraced inner wall, a smooth, relatively narrow ejecta (ash?) blanket, and no distinct large rays.

Ultimately, the question of origin of each lunar crater must rest on direct examination. If shocked rocks are collected *in situ* at a crater, its impact origin is substantiated. Shocked ejecta fragments should be prevalent over most of the lunar debris surface. Samples of breccias and regolith (soil-line rock debris covering the surfaces of both the maria and terrae or highlands) returned from all Apollo sites contain abundant signs of shock metamorphism. These materials evidently have participated in impact-crater-forming processes and tend to confirm the role of meteorites and comets in producing circular structures on the Moon. Repeated impacts over eons will sharply increase the proportion of more strongly shocked rock in the rubble.

A study of shock metamorphism of basalt, a prevailing lunar rock type, induced by a nuclear explosion in Nevada, indicates that the basalt is relatively resistant to shock damage below 300 kilobars. At higher pressures (~500 kilobars), feldspar constituents convert to thetomorphs and then melt with only localized fluid migration. Ferromagnesian minerals granulate, undergo variable decomposition, disperse within this molten feldspar, and then react with the melt to produce new mineral phases. Several crystal structure changes are recorded by optical and X-ray examination. In the extreme, dense basalt was converted by shock to a highly vesiculated, lightweight, partly glassy rock that resembles some volcanic cinders or dark pumice. Its true origin must be deciphered by confirming the shocked character of fragmental inclusions.

BIBLIOGRAPHY

Classic papers: G.K. GILBERT, "The Moon's Face: A Study of the Origin of Its Features," *Bull. Phil. Soc. Wash.*, 12:241–292 (1893); D.M. BARRINGER, "Coon Mountain and Its Crater," *Proc. Natn. Acad. Sci.* USA, 57:861–886 (1905); G.P. MERRILL, "The Meteor Crater of Canyon Diablo, Arizona: Its History, Origin and Associated Meteoritic Irons," *Smithson. Misc. Collns.*, 50:461–498 (1908); L.J. SPENSER, "Meteorite Craters as Topographic Features on the Earth's Surface," *Geogrl. J.*, vol. 81, no. 3, pp. 227–248 (1933); J.D. BOON and C.C. ALBRITTON, JR., "Meteorite Scars in Ancient Rocks," *Fld. Lab.*, vol. 5, no. 2, pp. 53–64 (1937); R.A. DALY, "Origin of the Moon and Its Topography," *Proc. Am. Phil. Soc.*, vol. 90, no. 2, pp. 104–119 (1946); R.S. DIETZ, "The Meteoritic Impact Origin of the Moon's Surface Features," *J. Geol.*, vol. 54, no. 6, pp. 359–375 (1946); J.J. GILVARRY and J.E. HILL, "The Impact of Large Meteorites," *Astrophys. J.*, vol. 124, no. 3, pp. 610–662 (1956).

Meteorite flux: E.M. SHOEMAKER, R.J. HACKMAN, and R.E. EGGLETON, "Interplanetary Correlation of Geologic Time," *Adv. Astronaut. Sci.*, 8:70–89 (1963); G. HAWKINS, "Impacts on the Earth and Moon," *Nature*, 197:781 (1963); W.K. HARTMANN, "Terrestrial and Lunar Flux of Large Meteorites in the Last Two Billion Years," *Icarus*, 4:157–165 (1965); D.W. PARKIN and D. TILLES, "Influx Measurements of Extraterrestrial Material," *Science*, vol. 159, no. 3818, pp. 936–946 (1968).

Mechanics of cratering: R. BJORK, "Analysis of the Formation of Meteor Crater, Ariz.," *J. Geophys. Res.*, 66:3379–3387 (1961); E.M. SHOEMAKER, "Impact Mechanics at Meteor Crater, Ariz.," in *The Solar System*, vol. 4, *The Moon, Meteorites, and Comets* ed. by B.M. MIDDLEHURST and G.P. KUIPER, pp. 301–336 (1963); D.E. GAULT, W.L. QUAIDE, and V.R. OBERBECK, "Impact Cratering Mechanics and Structures," in *Shock Metamorphism of Natural Materials*, pp. 87–99 (1968).

Shock metamorphism: B.M. FRENCH and N.M. SHORT (eds.), *Shock Metamorphism of Natural Materials* (1968); E.C.T. CHAO, "Shock Effects in Certain Rock-Forming Minerals," *Science*, 156:192–202 (1967); N.M. SHORT, "Shock Processes in Geology," *J. Geol. Educ.*, vol. 14, no. 4, pp. 149–166 (1966); and "Effects of Shock Pressures from a Nuclear Explosion on Mechanical and Optical Properties of Granodiorite," *J. Geophys. Res.*, 71:1195–1215 (1966).

Lists of terrestrial impact structures: J.H. FREEBERG, "Terrestrial Impact Structures: A Bibliography," *Bull. U.S. Geol. Surv. 1220* (1966); N.M. SHORT and T.E. BUNCH, "A World-wide Inventory of Features Characteristic of Rocks Associated with Presumed Meteorite Impact Structures," in *Shock Metamorphism of Natural Materials*, pp. 255–266 (1968).

Lunar impact structures: E.M. SHOEMAKER, "Interpretation of Lunar Craters," in *Physics and Astronomy of the Moon* ed. by Z. KOPAL, pp. 283–359 (1962); R.B. BALDWIN, *The Measure of the Moon* (1963); J. GEHRING, A.C. CHARTERS, and P.L. WARNICA, "Meteoroid Impact on the Lunar Surface," in *Lunar Surface Layer* ed. by J. SALISBURY and R. GLASER, pp. 214–265 (1964); J.F. MCCAULEY, "The Nature of the Lunar Surface as Determined by Systematic Geologic Mapping," in *Mantles of the Earth and Terrestrial Planets* ed. by S.K. RUNCORN, pp. 431–460 (1967); N.M. SHORT, "The Nature of the Moon's Surface: Evidence from Shock Metamorphism in Apollo 11 and 12 Samples," *Icarus*, 13:383–413 (1971).

(N.M.S.)

Meteorological Measurement

The behaviour of the atmospheric environment has long been a source of curiosity for mankind. Perception of warmth and cold, sunshine, winds, and precipitation must have been one of man's earliest sensations, yet it was many centuries before significant effort was expended on purposefully observing weather and on studying its causes.

The atmosphere as a fluid system. Only in the present century has the atmosphere been seen as a large fluid system, the behaviour of which is primarily determined by the need for a poleward transfer of heat. The weather (precipitation, wind, etc.) is primarily a result of a thermally driven, atmospheric, fluid motion on many time and space scales causing phase changes in the water substance—a small (4 percent) but vital component of the atmosphere. Thus clouds and precipitation (perhaps the two meteorological elements most frequently identified as "weather" by the layman) are a result of upward motion that favours cooling, and this in turn leads to condensation of water vapour into cloud droplets. Small-scale motion has important indirect effects on the growth of precipitation and even electrical activity within clouds. On a larger scale atmospheric motion is organized into planetary waves, cyclones, and anticyclones in the midlatitudes and into hurricanes and monsoons in lower latitudes. In all cases the distribution of resultant cloud cover is closely related to the configuration of motion in the system. The importance of cloud cover is not restricted to its ability to give precipitation: it also has a significant effect on the receipt of radiation by the Earth's surface. This effect is but one of the many mechanisms at work in the atmosphere and illustrates the complexity of a system with few, if any, simple, direct cause-and-effect processes.

Objectives of meteorological measurement. The basic object of meteorological measurement is to provide information on the behaviour of the complex atmospheric system in order to increase the understanding of the causes of weather and climate and to forecast their changes. But the scope of meteorological measurement has not always been so broad and yet, at the same time, so well defined. The fairly crude observations made prior to the 20th century were mostly the result of personal curiosity about the weather, and yet they were adequate to spawn the realization that the state of the atmosphere at a future time was, to some extent, predictable. This realization, together with the growth of aviation, has meant that weather forecasting holds prime place in the science of meteorology and, as such, has determined the pattern of meteorological measurement in this century. Most of the millions of meteorological measurements taken every day are for "operational" (routine) purposes—that is, to provide data for analysis that itself is the important first step in weather forecasting.

Since 1950 techniques of weather forecasting have undergone major changes. The forecasting problem in its widest sense has become the overall problem of meteorology—the elucidation of the workings of the general circulation of the atmosphere. A full understanding of the general circulation, including its component parts on many scales, will allow prediction of future states of its configuration—and this is the essence of forecasting.

This convergence of aims of routine and nonroutine observation means that meteorological measurement to-

Basic descriptive elements

day involves the description of the physical and dynamic state of the atmosphere on space scales ranging from millimetres to thousands of kilometres and time scales from seconds to years. The fluid state of the atmosphere suggests that five basic descriptive elements should be used: the gaseous constituents, density, pressure, temperature, and motion. Of these the first two (with the exception of water vapour) fluctuate very slowly, if at all, in time and, therefore, do not require frequent measurement. In contrast the last three are continuously variable in both space and time, and it is the variation of the elements themselves, their mutual interactions and their interactions with the water substance, that are observed. Continuous monitoring of the state of the atmosphere is of course impossible, so measurement largely involves the use of instruments (as opposed to visual estimation) in networks of varying density to take samples at different, yet internationally agreed times. The remainder of this article considers these methods in more detail.

HISTORY OF METEOROLOGICAL MEASUREMENT

True meteorological measurement, as opposed to simple perception of the vagaries of the atmosphere, began in the 17th century with the invention of the barometer and thermometer.

Barometers. Perhaps the most important discovery in meteorology was the principle of the barometer—the realization that changes in weather were related to fluctuations in the height of the mercury. The invention of the barometer in 1643 is attributed to Evangelista Torricelli, an associate of Galileo at Florence, who determined ". . . to make an instrument which might show the changes of the air, now heavier and coarser, now lighter and more subtle." By 1645 the idea had spread to France, and in 1663 the Royal Society of London had devised its own instruments using the same basic principle. A century later the principle of the aneroid barometer (see below *Atmospheric pressure*) was discovered, but not until 1843 was a satisfactory model produced.

Thermometers. Although the perception of heat and cold must have been more obvious to man than the weight of the air, the invention of the thermometer virtually coincided with that of the barometer. Early "thermoscopes" such as that devised by Galileo were defective in that they relied on the expansion of a gas, usually air, to indicate temperature change and had no scales. The barometer revealed the variability of the pressure of air and led to the adoption of the liquid-in-glass thermometer, which was being produced by Florentine glassblowers by mid-17th century. Of the many scales invented for the thermometer the most successful were those of D.G. Fahrenheit of Germany and Anders Celsius of Sweden.

Humidity measurements. One of the major stumbling blocks to the invention of the barometer was the widespread Aristotelian disbelief in the existence of vacuum. Water vapour also was slow in being recognized and accepted as fact. Once it was recognized that water vapour was a distinct substance, instruments to measure its amount in the atmosphere began to be invented.

The earliest humidity measuring instruments depended on the hygroscopic (moisture absorbing) properties of various substances. Changes in humidity affected either their weight or their volume. Sponge was frequently used in the 15th and 16th centuries—the principle being that changes in its weight in air gave a crude indication of the changes in humidity of the air. Although the first hygrometer was probably invented by Leonardo da Vinci, there is little doubt that he had access to the notes of Nicolas Cryfts, who in 1450 had suggested that a hygroscope could be made with the use of dried wool placed on scales. In the next three centuries many variations on this theme were conceived, culminating in the suggested use of seaweed in 1785. Hygrometers dependent on change of volume were similarly various. Wood was frequently used in the 17th century, and later it was found that gut, strings, and cords were good indicators of humidity changes. This eventually led to the discovery of the hair as a humidity sensor. Goldbeater's skin (ox gut) was probably first experimented with at the same time. It is still used today.

Humidity instruments dependent, on the one hand, on the formation of dew on artificially cooled surfaces (condensation hygrometers) and, on the other, on the amount of evaporation from a moist surface (psychrometers) were invented slightly later than the hygroscopic type. Ferdinand II of Tuscany probably discovered the condensation effect in 1655 and devised his own hygrometer working on this principle. A description of the cooling effect of evaporation was first recorded in the 17th century, but it was not until the late 18th century that the principle was used specifically to measure humidity. Theory and further development followed in the 19th century and continues today. Accurate measurement of humidity still challenges the makers of meteorological instruments.

Wind measurements. Wind direction was crudely recorded by the Greeks as early as 100 BC, but instruments to measure wind force as well as direction were only developed in the 17th century. The swinging-plate anemometer measured the deflection of a vertically hanging sheet of wood. This was an attempt to measure directly the dynamic pressure of moving air. The same basic principle is employed in the pressure-tube anemometer, conceived earlier but not constructed until the 1740s. The best known of this type is that invented in 1891 by the British meteorologist W.H. Dines, and still in use. The windmill effect, used in cup anemometers, was appreciated as early as AD 700 in Persia, but actual instruments were first constructed in France in the 17th century.

Swinging-plate anemometer

Rain gauges. Records of precipitation also have a long history. It is generally agreed that crude measurements were made with a container 18 inches wide in India about 400 BC, and that rain gauges were used in Korea in the 15th century. But the major developments again occurred in 17th-century western Europe, where self-recording gauges came into use.

Weather records and institutions. Whereas the 17th century saw the invention of the most important meteorological instruments for surface observation, the 18th century was a time of modification and usage to satisfy individual curiosity about the weather. The diaries of amateur observers provided invaluable information on changes in climate before the 19th-century establishment of national meteorological services. In Europe such institutions grew up in the period 1820–60 in Austria, Belgium, France, Greece, The Netherlands, and the United Kingdom. In the latter part of the 19th century, national organizations had been established in Japan, India, and Brazil.

At a very early date, the value of international cooperation was recognized. In 1853 a conference on maritime meteorology was held in Brussels to arrange the communication of meteorological data for the construction of the first weather maps. Nearly 20 years later a similar meeting laid down recommendations for the transmission of land-based data. By 1880 the directors of the national meteorological services had grouped themselves into the International Meteorological Organization (IMO). For over 70 years the IMO encouraged the improvement of observation and transmission of data for forecasting purposes. After World War II the IMO was reconstituted as the World Meteorological Organization (WMO), and in December 1951 the United Nations recognized the WMO as a specialized agency. Over 130 countries (virtually the entire world) are members of WMO and thus allow the free communication of standardized observations, mainly for forecasting purposes.

World Meteorological Organization

SURFACE MEASUREMENTS

Meteorological observations made at the Earth's surface are of two kinds: sensory observation—taken without the use of a measuring instrument and estimated by experience only; and instrumental observation—taken with the aid of measuring instruments and usually involving corrections and calculations to the actual readings. Three types of observing stations exist to record as fully as possible the state of the lower atmosphere. A synoptic station observes the following elements: present and past weather; wind direction and speed; amount, type, and height of base of cloud; visibility; air temperature; humidity; and atmospheric pressure. At certain synoptic stations

added observations are taken of pressure tendency; extreme temperatures; precipitation amount; state of ground; direction of cloud movement; and any special phenomena. Measurements of all these parameters are often taken at three-hour intervals, but at some stations certain measurements are taken every half hour. The other types of station are climatological and agricultural stations. At a principal climatological station, observations are usually made once per day of all or most of the following: weather, wind, clouds, visibility, temperature, humidity, pressure, precipitation, snow cover, sunshine, and soil temperature. Agrometeorological stations concentrate on air and soil temperatures, soil moisture, precipitation, sunshine, radiation, and evaporation amounts.

Cloud height: direction and speed of movement. Observations of clouds consist of visual estimations of their form and extent, instrumental measurement of their height above the ground, and visual estimations or instrumental measurements of their speed and direction of motion.

The identification of cloud forms is based on the specifications well illustrated in the WMO *International Cloud Atlas*. Ten basic forms constitute the basis of classification, but the infinite variety within these types means that recognition of an individual cloud-type requires some experience on the part of the observer. Cloud amount is also assessed visually, the amount of the visible dome over the observer that is cloud-covered being given in either eighths (*oktas*) or tenths.

The *International Cloud Atlas*

Although cloud-base height is often visually estimated by the experienced observer, it is possible, and preferable, to measure it (in metres) either with small pilot balloons or vertical light beams. The balloons, rising at a rate of about 130 metres per second, are tracked by binoculars or theodolite, and the duration of their visible ascent allows calculation of the cloud base. At night the measurement at one end of a base line of the angular elevation of the light spot produced on cloud base by a vertically pointed light at the other end of the base line provides a ready measure of cloud height. Instruments also exist to measure the speed and direction of cloud movement.

Sunlight duration. Although the WMO recognizes four main types of instrument for the measurement of sunshine duration, it recommends the use of the Campbell-Stokes recorder, which focusses the Sun's radiation to burn a trace on a chart. The length of the trace gives the duration of bright sunshine in hours. The instrument comprises a glass sphere about ten centimetres (four inches) in diameter mounted on a section of a spherical bowl such that the charts in the grooves of the bowl are scorched by the Sun's radiation. The sphere is mounted in differing positions, depending on whether it is in high, middle, or low latitudes, and three overlapping pairs of grooves are provided in the bowl to take cards suitable for different seasons of the year.

Visibility. The measurement of visibility is undertaken largely to meet aviation requirements, but the occurrence of ground fog has sufficient effect on man's other activities to justify its surface observation. As in the case of cloud observations, both visual and instrumental measurements of visibility can be made. The usual procedure is to record whether or not objects of a known distance (in kilometres and metres) from the observing station can be seen or not, but at night, visibility meters are useful. These instruments use photoelectric cells and are of two types, one measuring the attenuation coefficient of a long column of air, and the other measuring the scatter of light from a small volume of air.

Atmospheric pressure. The pressure of the atmosphere is the force it exerts per unit area on a given surface. This force is generally measured by either mercury or aneroid barometers, the units of measurement being millibars (1 millibar = 1,000 dynes/cm^2). A reading of pressure must be corrected for temperature and station altitude above mean sea level to make observations at different points comparable. The rise or fall of pressure in the previous three hours—the pressure tendency—is also recorded.

Barometer corrections

The mercury barometer works on the principle that the atmospheric pressure is balanced against the weight of a column of mercury. It is a sensitive, carefully calibrated instrument and once installed is not usually moved. The height of the mercury is read with the aid of a vernier scale (giving the pressure to a tenth of a millibar), and this reading is corrected with the aid of the attached thermometer reading and the knowledge of the height of the instrument above sea level. The number of readings per day ranges from one to 24 depending on the type of meteorological station.

Aneroid barometers are less accurate than mercury instruments but are particularly useful for their portability and their adaptability for use as barographs. An aneroid barometer has two basic parts: a closed, metal, evacuated chamber and a strong spring that prevents the chamber from collapsing. Variations in atmospheric pressure cause fluctuations in the size of the chamber. These fluctuations are recorded. If the recording is in the form of a pen trace on a chart, thus giving the variation of pressure with time, the instrument is known as a barograph.

Air temperature: maximum and minimum. A surface observation of air temperature refers to the shade temperature at 1.2 metres (four feet) above the ground. Only shade temperatures give the true air temperature: direct exposure of the thermometer to the sun's radiation would result in readings of the temperature of the glass of the thermometer rather than that of the surrounding air. To achieve this important objective, thermometers are housed in white, louvered screens (shelters) that allow adequate ventilation. The instruments themselves, used for actual, maximum, and minimum temperatures, are usually mercury-in-glass or alcohol-in-glass thermometers. The recommended scale is Celsius, but in some countries, notably the U.S., Fahrenheit is still used. Readings are estimated to a tenth of a degree. A maximum thermometer has a constriction in the bore below the lowest graduation. This constriction allows the mercury to expand freely up the bore but prevents its return when temperatures fall after the maximum is reached. The instrument is reset by gently shaking the mercury down through the constriction. A minimum thermometer usually has a dark index, about 2 centimetres (¾ inch) long, immersed in alcohol. As the temperature falls, the meniscus (curved surface) of the alcohol pulls down the index, which is left in the minimum reading position when the temperature rises.

Temporal variations in temperature are recorded on thermographs of which there are various types. They range from those dependent on the different expansion rates of two metals to those based on the variation with temperature of electrical resistance of fine wire elements. In all cases the changes in temperature are recorded by a pen arm on a chart wrapped around a rotating drum.

Earth temperature. Mercury-in-glass thermometers are used for both grass minimum readings and observations at depths of 5, 10, 20, 50, and 100 centimetres (2, 4, 8, 20, and 40 inches) in the soil. The grass minimum thermometer is simply a minimum thermometer exposed horizontally over short grass with its bulb just touching the tips of the blades of grass. The first three depths of soil are sampled by thermometers bent at right angles with their scales facing upward so that they can be read without being disturbed. At depths of 50 and 100 centimetres thermometers are usually suspended in tubes sunk into the earth.

Grass minimum thermometer

Precipitation. The estimation of precipitation amounts simply involves the exposure of a container for a given period, usually 24 hours, and the measurement of the depth of collected water (or liquid water equivalent of snow and hail). The most common rain gauge stands about 30 centimetres (12 inches) high and has an orifice of 13 centimetres (about 5 inches) diameter with a funnel leading to a glass jar, which itself stands in a metal can. Gauges designed for very wet areas usually have a greater capacity. The amount of rainfall in the 24-hour period is established with the use of a rain measure graduated in inches or millimetres, preferably the latter.

Autographic rain gauges provide two types of information on the variation of rainfall with time: first, the total amount of rain that has fallen since the time the record started and, second, the intensity of rainfall at any instant.

Very few gauges of the second type are in operational use. In the most common of the first type, the rain is led into a chamber containing a light float. The vertical movement of the float as the level of the water rises is recorded on the chart. When the chamber is full, a syphoning mechanism allows the whole operation to start again. A second type of rainfall recorder simply empties the known amount of water in the chamber by tipping and counts the number of tips.

Humidity. Water content of the air may be expressed in several ways: the vapour pressure (in millibars); the vapour density (in grams per cubic centimetre); the mixing ratio (in grams of water vapour per kilogram of dry air); and the more familiar relative humidity (in percent). Basically, the humidity of the air is related to the difference in the dry- and wet-bulb temperature readings. These readings are obtained by using a psychrometer, which has two thermometers mounted together, one with a wet cloth sleeve over its bulb. Reference to specially prepared tables gives relative humidity values. Changes in relative humidity are satisfactorily measured by the expansion or contraction of hairs and they are often used in the hygrograph—an instrument that provides a daily or weekly trace of temporal changes in relative humidity.

Wind speed and direction. Instruments for recording wind speed and direction are some of the most elaborate of the routine surface hardware. Wind direction (the direction from which the wind is blowing) is expressed in degrees, measured clockwise from geographical north, and assessed with the use of a freely turning wind vane. The wind speeds, corrected if necessary to the standard height of 10 metres (32.8 feet) are recorded in knots, and two types of instrument provide the relevant values: cup anemometers and pressure-tube anemometers. Cup anemometers work by using the wind to spin three hemispherical cups, driving a small electric generator; the resultant current is indicated on a meter, the scale of which is graduated in terms of wind speed. The pressure-tube anemometer uses a combination of pressure and suction forces to measure wind speed. Moving air enters an orifice, kept to the wind by a vane, and provides an upward push to a float in a chamber, while holes in the supporting vertical tube provide a suction force that pulls the float. A pen attached to the float gives a daily record of wind speed in knots on a chart.

Cup anemometers

Pollution and pollen count. The bases of measurement of smoke and pollen spores in the atmosphere are quite similar. In the case of smoke, a sample of air (usually about two cubic metres, or 70 cubic feet, per day) is passed through a filter paper, the smoke stain measured, and the reading converted into smoke concentrations of micrograms per cubic metre. The pollen count is measured by directing a continuous jet of air onto a stationary or moving slide that has an especially prepared surface. The pollen spores caught on the slide are counted for any desired time interval. Sulfur dioxide is perhaps the most dangerous of the routinely observed pollutants; it is usually measured simultaneously with smoke concentrations. The air that has just passed through the filter paper is passed through a chemical. The resultant acidity of the chemical gives an indication of sulfur dioxide concentration.

Radiation. Virtually all of the energy that drives the atmospheric circulation is received in the form of solar radiation. Three basic types of solar radiation are measured. Direct solar radiation is measured by a pyrheliometer; sky radiation and global (direct plus sky) radiation, by a pyranometer. But the Sun is not the only radiating body: both the Earth and the atmosphere emit thermal radiation (known as terrestrial radiation). Pyrradiometers measure both solar and terrestrial radiation, and a net pyrradiometer measures the net flux of downward and upward total (solar, terrestrial surface, and atmospheric) radiations through a horizontal surface.

As radiation is a process of continuous transfer of energy, it can be measured as a flux or as an amount received in a given time. Radiative flux per unit area is usually measured in milliwatts per square centimetre and radiation amount per unit area either in joules per square centimetre or milliwatt-hours per square centimetre. Many types of radiation instruments exist but most of them incorporate thermopiles (thermoelectric devices) as their sensitive elements and the outputs are measured as small electrical voltages. The exact nature of the equipment at any one station depends on observational requirements and these vary widely among national meteorological centres.

Evaporation. The measurement of water loss from the Earth's surface is of little value to routine weather forecasting but is vital to the elucidation of the water balance of the atmosphere. The term evaporation means water loss from a water surface; and the term evapotranspiration is used to cover the combined loss of water by evaporation and transpiration (water loss from leaf surfaces of plants that have taken up water from the ground). Both types of water loss are measured in millimetres of liquid water lost per unit time (an hour or a day) from a unit area.

Evapotranspiration

Evaporation is measured by observing the change of level of water in a pan or tank. The pan may be sunken, so that the water surface is nearly level with the ground; or above the ground to allow free air circulation; or floating in a lake to provide estimates of lake evaporation.

An evapotranspirometer is a container of soil and vegetation from which the water loss is measured by weighing or accounting for all incoming water at the surface and all outflow from the bottom of the container. Several designs exist, and research is still active in an attempt to devise instruments that will accurately measure water loss by the evapotranspiration process.

UPPER-AIR MEASUREMENT

For obvious reasons, information on conditions in the free atmosphere is far less complete (in both numbers of elements observed and density of observing networks) than on conditions at the surface. Routine upper-air observation by radiosonde (balloon-borne radio transmitter) is restricted to temperature, humidity, pressure, and wind speed and direction at selected levels between the surface and an altitude of about 18 kilometres (11 miles). Above that level rocket sondes are used. Data on other meteorological phenomena—particularly clouds—are available with the use of sferic equipment, radar, satellites, and aircraft.

Sferic measurements. When lightning flashes in a thunderstorm, electromagnetic waves are produced, which are known in meteorology as atmospherics or sferics. As these waves are fairly easily detectable for a distance of thousands of kilometres, sferics give a convenient way of locating thunderstorms.

The most common type of sferic equipment is the cathode-ray direction finder. This unit has two aerials at right angles. On receipt of the sferic, the direction from which the waves came is displayed on a cathode-ray tube. The information is either photographed or noted by the observer. A single direction is, of course, of little value, so a sferic network usually comprises from two to four stations 500–1,000 kilometres (about 300–600 miles) apart. The simultaneous recording of a lightning flash by these stations allows a "fix" on the storm to be made by simple intersection methods. In thundery weather the sferics stations usually operate for ten minutes before each hour and plot as many fixes as is possible.

Cathode-ray direction finder

Radar measurements. Radar can be used to "see" precipitation particles in the same way it can be used to "see" aircraft. The development of radar meteorology has provided information on the dynamics of clouds and upper air winds and has been most valuable in short-term forecasting.

Basically, a radar set comprises a transmitter of wave energy, a receiver of reflected energy, and a display system so arranged that the object under surveillance can be conveniently located. In meteorology the object under surveillance is most often a cloud or collection of clouds, and the energy is reflected by precipitation-sized raindrops or ice crystals in the cloud. For a given wavelength the amount of backscattered (reflected) energy is proportional to the sixth power of the drop or crystal diameter. For a given drop-diameter the reflected energy increases with a decreasing wavelength of the transmitted energy, but short wave-

length radars suffer from high attenuation—a weakening of outgoing and reflected wave by intervening clouds and raindrops—and are not used as widely as their properties would initially appear to warrant.

Radar observations of clouds (or strictly, the precipitation-sized particles in the clouds) are usually displayed in two basic ways on a cathode-ray tube. The horizontal location of "echoes" is shown by the plan position indicator (PPI), usually a full circle display with the radar station at the centre, with azimuth markings every 20°, and concentric range markers a constant distance apart. A variation is the constant-altitude plan position indicator (CAPPI): this provides PPI pictures of echoes at various altitudes other than the surface and has proved of great value in cloud physics research. The second basic type of display is the range–height indicator (RHI), which gives a section of precipitating clouds along a chosen direction. Both types of display may have a horizontal range over 150 kilometres (about 93 miles); the vertical range of the RHI is about 15 kilometres. Although most radar information is transmitted and stored in the form of either verbal descriptions or photographs, research is currently being directed toward the automatic numerical coding of echoes.

Plan position indicator (PPI)

Radar is also used to track targets attached to balloons. The slant range and direction of the target at different times provide a measure of the upper-air wind speed and direction.

Radiosonde measurements. Radiosondes are used to measure pressure (in millibars), temperature (in degrees Celsius), and humidity (in percent relative humidity) up to heights of 16 to 20 kilometres (10 to 12.5 miles)—occasionally to 30 kilometres (19 miles). The instrument comprises meteorological sensors, a radio transmitter, and a device to allow control of the transmitter from a distance. The whole is suspended beneath a large balloon that rises until it bursts; the sonde descends by parachute.

Sensors of radiosondes vary from country to country but certain types are most frequently used. For pressure, aneroid capsules are most satisfactory; for temperature, thermometers are usually of the bimetallic or resistance type; and for humidity, hair or goldbeater's skin is recommended. Of the four types of sonde listed by the WMO, that using the variable audio-frequency principle is most used. As the balloon rises with this instrument, the sensors sample the pressure, temperature, and humidity at height intervals of about 120 metres (390 feet) and activate inductors that vary the modulation frequency. The ground station receives the signal transmitted from the battery-powered sonde and compares its frequency (with the aid of an audio-frequency oscillator) with calibration curves to establish the conventional units. The calibration curves themselves are a result of subjecting the sensors to varying, but known, values of pressure, temperature, and humidity in the laboratory. Despite these efforts, it is recognized that quite large errors do occur in radiosonde measurements (particularly of humidity).

Often the radiosonde also provides useful information on upper winds. If the transmitted signal is sufficiently continuous with time to allow regular observations of direction and elevation of the instrument, upper wind speed and direction can be calculated.

Rocket soundings. Above the 30-kilometre (19-mile) level, rockets must be used to take instruments to the middle atmosphere (30–100 kilometres; 19–63 miles). Once there, the actual procedure of measurement of temperature and winds is very similar to that employed with balloon-borne sondes. The temperature element is basically similar to the radiosonde type and is ejected from the rocket at altitudes of about 70 kilometres (43 miles). When used by day, corrections of up to 20° C (36° F) must be made for solar radiation at a height of 65 kilometres (40 miles). To obviate this, many rocket soundings are made at night. Because the air density is so low at these altitudes (65 kilometres), the parachute and sensor fall at 200 metres (660 feet) per second, a speed that may cause dynamic heating of as much as 15° C (27° F). Such errors must, of course, be eliminated by correction procedures. The wind speed and direction are determined from radar observations of the partially metallized parachute. There is no pressure sensor on the sonde: pressures are calculated on the basis of the pressure at the 20-kilometre level. Another method of measuring the wind characteristics is to disperse fine wires from a rocket and track the resultant "cloud" by radar. This can give good wind measurements in the layer at a height of 65–90 kilometres.

Dynamic heating

The very small values of air density at high levels would, at first sight, appear to be of little meteorological importance. In fact, the measurement of density allows the calculation of temperature and pressure profiles and as such allows the useful upward extension of knowledge of the thermal and dynamic structure of the atmosphere. Radar tracking of small spheres falling at supersonic speeds in the 45–100-kilometre layer allows the calculation of drag on their surfaces and from this the density can be determined. A more explicit way of measuring temperature is to eject grenades into the high atmosphere, to record their explosion at vertical intervals of about five kilometres (about 16,400 feet) and to measure the travel time of the shock wave from each explosion to the ground. This travel time depends on mean air temperature and mean wind in the column between the ground and the grenade. The effects of temperature and wind can be separated and mean winds and mean temperatures separately determined.

Satellite observations. The advent of meteorological satellites in the 1950s and 1960s was perhaps the most spectacular development in the history of meteorological measurement. These satellites can provide efficient and economic global observation of different types of radiation emitted by the Earth and atmosphere, and recently strides have been made in the deduction from these radiation measurements of the vertical distribution of temperature and humidity in the atmosphere.

Perhaps the most striking results of satellite observations are the cloud photographs. The first pictures came from TIROS I in 1960, and since 1966 automatic picture transmission has made such pictures available to forecasters soon enough to be of value to a forecast. The quality of pictures has continually increased through to NIMBUS IV, launched April 8, 1970, and these show in great clarity the organization and structure of cloud systems ranging from large cyclonic vortices thousands of kilometres across to individual shower clouds only one kilometre across. Both NIMBUS satellites and Applied Technology Satellites (ATS) provide continuous surveillance of the globe: the former by having a polar orbit and by being Earth oriented and the latter by being geosynchronous (*i.e.*, the orbit is synchronized with the rotation of the Earth). A satisfactory global coverage of cloud pictures could probably be achieved by using a combination of three or four geosynchronous and one or two polar orbiting satellites, with the former relaying pictures received from the orbiting satellites to those points of the Earth not in direct view of the latter.

TIROS I

Both solar and terrestrial radiation amounts are measured by radiometers on satellites. Maps of Earth surface and cloud-top temperatures have been constructed from measurements of long-wave radiation. These measurements have provided useful estimates of the net amount of energy available to the Earth and atmosphere and some indication of the day-to-day changes in the heat balance of the system.

In 1969 a further important step was made in meteorological satellite observation. The Satellite Infra-Red Spectrometer (SIRS) on the NIMBUS III satellite permitted the calculation of the vertical temperature distribution through the atmosphere directly below the spacecraft. The results showed remarkable agreement with radiosonde soundings and already they are being used in the U.S. Weather Bureau's numerical weather analysis and forecasting programs. A later model on NIMBUS IV provided similarly encouraging results.

Aircraft observations. Aircraft observations can provide a valuable supplement to information gathered by conventional methods. An aircraft provides a moving observation platform to obtain information on temperature, humidity, pressure, altitude, wind direction, and speed

and, occasionally, ice accretion, turbulence, visibility, and cloud types, amounts, and heights. Usually aircraft observations are taken for a specific research purpose; *e.g.*, frontal structure and cloud dynamics, but in parts of the world where surface stations are few and far between, they fill a significant gap in operational networks.

MEASUREMENT TECHNIQUES IN MICROMETEOROLOGY AND MESOMETEOROLOGY

All measurements covered earlier are primarily relevant to large-scale atmospheric physical and dynamic processes. Micrometeorology and mesometeorology are concerned with atmospheric behaviour on space scales ranging from millimetres to tens of kilometres.

In micrometeorology interest is focussed on the distribution of temperature, humidity, water vapour pressure, wind speed, and evaporation with height above the ground surface up to a few decametres (under 100 feet). In this way the vertical transfers of heat, water, and momentum from ground to atmosphere and vice versa can be evaluated. The vertical profiles are established by attaching instruments to various levels of a tower and taking frequent, or occasionally continuous, observations. In choosing equipment for any project, it is important to assess what degree of instrument sensitivity is required. Most micrometeorological investigations employ fine temperature-sensing elements connected to suitable sensitive recorders, and with slight modifications they also allow the measurement of humidity. Wind speed may be measured by sensitive, small-cup anemometers or hot-wire anemometers. The latter works on the principle that moving air causes heat loss from a constantly heated wire, and changes in this heat loss are monitored on a gauge suitably calibrated to give the wind speed.

The instruments employed in mesometeorology are basically similar to those used in routine synoptic observations. But because mesoscale atmospheric systems have horizontal dimensions of about 150 kilometres and temporal dimensions of a few hours, most of the instruments are of the recording variety (occasionally of increased sensitivity) and they are deployed in networks of greater density than the routine synoptic grids. The organization of one major mesoscale research project exemplifies these points. In an area of 50 by 65 kilometres (31 by 40 miles) the following instruments were installed: 33 microbarographs, 10 hygrothermographs, 10 wind recorders, 165 rain gauges, and 165 hail gauges. Also, clouds were carefully photographed so they could be analyzed stereoscopically, and as a result of this observational onslaught, valuable new information and insight was provided on convective mesosystems.

High-density networks

WEATHER-STATION NETWORKS AND COMMUNICATIONS SYSTEMS

It is pertinent to consider how the stations employing the instruments described earlier are distributed in space and how frequently observations are made so that a coherent picture of atmospheric behaviour emerges.

Surface observations at synoptic stations are most widely taken every three hours from midnight Greenwich Mean Time (GMT). Nearly 8,000 stations supposedly conforming to this regulation are to be found on land areas and they are supplemented by over 4,000 weather, merchant, and passenger ships. Impressive as these numbers appear, there are huge gaps in the network, and their elimination is one of the major future tasks of meteorology.

Upper air observations by radiosonde are taken every 12 hours, again starting at midnight GMT. The network in the Northern Hemisphere, although not as complete as is desired, is quite impressive: it is in the Southern Hemisphere that knowledge of the free atmosphere suffers badly from a lack of data. The network of rocket-firing stations is the most sparse of all. This is, of course, due to the relatively recent innovation of this observation technique and the expense involved in each ascent. It is possible that further meteorological rocket development may be rendered unnecessary by the increasing capability of satellites.

The networks of radar sets are most dense in the U.S., western Europe, and Japan, partly because such stations are quite expensive to run. Although widely used on a routine basis, particularly in the United States, there are no international regulations for radar observations as there are for all surface and radiosonde measurements. Consequently the data tend to be treated as of an ephemeral nature and are not stored as carefully as more conventional recordings.

Once the conventional surface and upper-air measurements are made they are coded and transmitted to the local (often national) meteorological centre by telephone, radio, or teleprinter. Here they are decoded and analyzed and the resultant forecast in cartographic form is often transmitted back to outstations by facsimile machine. Much data and many satellite photographs are also transmitted internationally.

TRENDS IN METEOROLOGICAL MEASUREMENT

Perhaps the most heartening result of the rapid meteorological technological advance in the 1960s was the realization that observation of the whole atmosphere was possible. The continued development of remote sensing devices (a major trend in itself) will allow many of the large gaps in the observation network to be filled in the early 1970s. Such is the aim of the Global Atmospheric Research Program (GARP)—a massive international effort to increase understanding of the atmosphere.

Global Atmospheric Research Program

Major additions will be made to the radiosonde network —particularly in the tropics and Southern Hemisphere. Substantial increases will also be made in the numbers of commercial ships and aircraft capable of making meteorological observations. At the surface, it is hoped that automatic weather stations (on both land and sea) will consistently provide good quality synoptic data in those areas that at present have too sparse a network.

But it is the development of satellite observation capability that will provide the greatest amount of information. Much effort has been expended on developing the Global Horizontal Sounding Technique—a system whereby satellites track constant-altitude balloons as they are carried around the Southern Hemisphere. The technique is at present under review because of the short life of the balloons. More encouraging are the derivations of wind speeds from cloud movements recorded by geosynchronous satellites and of temperature profiles in the atmosphere below the vehicle. These developments are of paramount importance as they will allow the automatic and frequent observation of the parameters vital to the working of the most recent mathematical models of the atmosphere. Perhaps the time will soon come when a global model of the atmosphere will exist that is constantly being updated by a stream of data from an adequate observation system. When that time arrives a measure of true understanding of the behaviour of the atmosphere will exist.

BIBLIOGRAPHY. Authoritative works on the history of meteorological instruments include: W.E.K. MIDDLETON, *History of the Barometer* (1964), *A History of the Thermometer and its Uses in Meteorology* (1966), and *Invention of the Meteorological Instruments* (1969). Other works relating to history are R. T. GLAZEBROOK, *Dictionary of Applied Physics*, vol. 3, *Meteorology, Metrology, and Measuring Apparatus* (1923); and L. HOWARD, *The Climate of London*, 3 vol., 2nd ed. (1803). A useful summary of the evolution of meteorology is contained in W.N. SHAW, *Manual of Meteorology*, vol. 1, *Meteorology in History* (1926).

The following books, published by the WORLD METEOROLOGICAL ORGANIZATION, Geneva, contain the international requirements for surface and upper-air observation practice: *International Cloud Atlas* (1956), *Guide to Climatological Practices* (1960), and *Guide to Meteorological Instrument and Observing Practices*, 3rd ed. (1969). More details on instruments are available in the following publications of the METEOROLOGICAL OFFICE, London: *Handbook of Meteorological Instruments* pt. 1, *Instruments for Surface Observations* (1956), and pt. 2, *Instruments for Upper-air Observations* (1961).

Radar usage is well covered in L.J. BATTAN, *Radar Meteorology* (1959); and "Use of Ground-Based Radar in Meteorology (Excluding Upper-Wind Measurements)," *Tech. Note Wld. Met. Org.* no. 78 (1966). High-altitude observations and satellite capabilities are treated in "Meteorological

soundings in the upper atmosphere," *Tech. Note Wld. Met. Org.* no. 60 (1964); E.C. BARRETT, *Viewing Weather from Space* (1967); L.F. HUBERT and P.E. LEHR, *Weather Satellites* (1967); and W.K. WIDGER, *Meteorological Satellites* (1966).

(B.W.A.)

Methodists

Methodist Christianity is a highly institutionalized example of a form of religious life and thought that has appeared from time to time during the history of the Christian Church.

Nature and significance. Methodism is marked by a warm acceptance of the doctrines of historic Christianity, without a stringent insistence on the need for doctrinal conformity or a great interest in theological speculation; by an emphasis on those doctrines that indicate the power of the Holy Spirit to confirm the faith of the believer and transform his personal life; by insistence that the heart of religion lies in personal relationship with God; by simplicity of worship; by the partnership of ordained ministers and laity in the worship and administration of the church; by a concern for the underprivileged and the betterment of social conditions; and (at least in its British form) by the formation of small groups for mutual encouragement and edification. Methodist churches of the two traditions—British and American—present considerable differences. Those in the British Isles and the British Commonwealth stand third in size to the churches of the Anglican Communion and the Roman Catholic Church. They are composed mostly of members of the middle class, though not of its most affluent section. They have affinities both to Anglicanism and to English Nonconformity (of which, nevertheless, they have rarely been an integral part). All are nonepiscopal. Those of the American tradition, spread over all those parts of the world where American influence is strong, are much larger in size than their British counterparts and receive considerable support from the wealthiest elements in society while remaining predominantly middle class. They are episcopally governed but in close relationship with the nonepiscopal representatives of American Protestantism.

For the last three-quarters of a century there have been determined efforts to bring the two traditions together, since they share the same doctrinal emphases, the same sort of connectional system of government (by which all ministers are "in connexion" with the central authority of the church, and the agencies of the church are centrally administered), and the same allegiance to John Wesley.

History. *Origins.* John Wesley, the founder of Methodism, was born in 1703. After ordination in the Church of England, he was elected a Fellow of Lincoln College at Oxford in 1726. In the following year he left Oxford temporarily to act as curate to his father, the rector of Epworth. Back in Oxford, to which his younger brother Charles had now come, he found himself a member and soon the leader of a group of earnest students pledged to frequent attendance at the Sacrament (Holy Communion), serious study of the Bible, and regular visitation of the filthy Oxford prisons. The members of this group received the sobriquet of Methodists and the group itself that of the Holy Club.

In 1735 both John and Charles Wesley set out for Georgia to be pastors to the colonists and missionaries (it was hoped) to the Indians, at the insistence of the founder of the colony, James Edward Oglethorpe. They were unsuccessful in their pastoral work and did no missionary work. The brothers returned to England conscious of their lack of genuine Christian faith. They looked for help from members of the Church of the Brethren, who were staying in England for a while before joining Moravian settlements in the American colonies; among these Peter Böhler was especially important. On May 24, 1738, John Wesley's *Journal* narrates that he "felt" his "heart strangely warmed" and continues, "I felt I did trust in Christ, Christ alone, for salvation; and an assurance was given me that He had taken away *my* sins, even *mine*, and saved *me* from the law of sin and death." Charles Wesley had reported a similar experience a few days previously.

Moravian influences

Some months later, John Wesley was invited by his friend George Whitefield, also an Anglican clergyman who had undergone a "conversion experience," to come to the city of Bristol and help to preach to the colliers of Kingswood Chase, just outside the city, where human conditions were at their lowest. Wesley came and found himself, much against his will, preaching in the open air. This enterprise among the outcasts of society was the beginning of the Methodist Revival. Whitefield and Wesley at first worked together, but later separated on doctrinal grounds: Whitefield believed in double predestination; Wesley regarded this as an erroneous doctrine and insisted that the love of God was universal.

Under the leadership at first of Whitefield and afterwards of Wesley, the movement rapidly gained ground among those who felt themselves neglected by the Church of England. Wesley differed from contemporary Anglicans not in doctrines but in emphases: he claimed to have reinstated the Biblical doctrines that a man may be assured of his salvation and that, by the power of the Holy Spirit, he is capable of attaining "perfect love" for God and his fellows in this life. Wesley's helpers included only a few ordained clergymen and notably his brother Charles, who wrote more than 6,000 hymns to express the message of the Revival and to assist both the public and private meetings of the Methodists. Many lay helpers were recruited.

The Methodists formed a "Society" within the Church of England, and Wesley never wished them to leave that church. But relations with Anglicans were often strained, and in 1784, when there was a shortage of ordained ministers in America after the Revolution, the Bishop of London refused to ordain a Methodist for America. Wesley, acting in an emergency and on biblical principles that allow (as he thought) a presbyter to ordain, consecrated Thomas Coke as superintendent and ordained two others as presbyters. In the same year, by a Deed of Declaration, he appointed a Conference of 100 men to govern the Society of Methodists after his death.

British Methodism from the death of Wesley to the present. The definite break with the Church of England came in 1795, four years after Wesley's death. After this, English Methodism, with vigorous outposts in Ireland, Scotland, and Wales, rapidly developed from a Society into a Church, with its system centred in the annual Conference, which controlled all its affairs. The country was divided into districts and the districts into circuits, or groups of congregations. The ministers were appointed to the circuits, and each circuit was led by a superintendent.

Organization of the Methodist Church

Much power remained in the hands of the local trustees, who sometimes opposed the wishes of the Conference. The sacraments were administered by the ordained ministers, but "local preachers"—laymen—conducted the majority of other services.

This tightly knit system enabled the Wesleyan Methodist Church to grow rapidly throughout the 19th century, at the end of which it counted 450,000 members. The growth was largest in the expanding industrial areas, where Methodism enabled men and women to overcome economic depression by spiritual means: and often, by its encouragement of thrift and simple living, raised their economic status as well.

These developments, combined with the autocratic habits of some leading Methodist ministers, alienated many of the more ardent and democratic spirits, and there was a series of schisms. The Methodist New Connexion broke off in 1797, the Primitive Methodists in 1811, the Bible Christians in 1815, and the United Methodist Free Churches in 1857.

The smaller Methodist groups were in closer contact with the working classes than the Wesleyans and provided the leadership in early trade unionism to an extent disproportionate to their size. The Wesleyans were at first conservative in politics but in the second half of the century identified themselves more and more with the liberalism of the foremost statesman of the century, William Gladstone.

A movement to reunite the Methodist groups began about the turn of the century and reached success in two stages. In 1907 the Methodist New Connexion, the Bible

Christians, and the United Methodist Free Churches joined to form the United Methodist Church; and in 1932 the Wesleyan Methodist Church, the Primitive Methodist Church, and the United Methodist Church came together to form the Methodist Church.

The Methodist Church has shared with the other English churches in the numerical decline that seems to have begun about 1910 and has affected both the lay membership and the ordained ministry. In 1969, 651,139 members and 3,149 ministers were reported. During the same period Methodists have not ceased, however, to exercise influence on national and local life and to plan fresh attempts at evangelism. But the chief development has been in the steady improvement of relations with the Church of England, culminating in a scheme for reunion in two stages. This was accepted in principle by both churches in 1965 and, in a slightly modified form, approved by the Methodist Conference in 1969 and 1971. Its implementation has been prevented by the failure of the governing body of the Church of England to register a sufficiently large vote in its favour.

Methodism in America. Methodism was brought to America by immigrants from Ireland converted by John Wesley. Wesley also sent preachers to strengthen the societies, and by far the most successful of these was Francis Asbury, a blacksmith, who arrived in 1771. He covered vast distances in his preaching journeys and organized the movement on Wesley's concepts, adapting them to the needs both of the settled communities and of the frontier. Wesley damaged the effect of his enterprise by taking the side of the English government at the time of the Revolution, but Asbury restored the situation by aligning himself with the new American republic. Wesley sent the men whom he had ordained as presbyters, with Thomas Coke as superintendent, to help Asbury. The Methodist Episcopal Church was constituted in 1784 and regarded itself as autonomous. Asbury and Coke, to Wesley's great displeasure, allowed themselves to be called bishops.

The next 50 years saw a remarkable advance led by the circuit riders who preached to the frontiersmen in simple terms and went with them to the furthest reaches of the American expansion, while colleges and schools were built in their wake. The advance was halted by the slavery issue, which bitterly divided the Methodists of the North and the South. In 1844 it split the Methodist Church into two churches: the Methodist Episcopal Church and the Methodist Episcopal Church, South (organized in 1845). After the Civil War both churches increased rapidly in numbers and in material resources. They became gradually assimilated to the general pattern of American Protestantism, and, when it was clear that the old issues no longer divided them, they began to move together. Certain antipathies had, however, outlived the issues that caused them, and it was not until 1939 that they came together to form The Methodist Church. The Methodist Protestant Church, a smaller group, joined in the same union. Their educational work, principally in the form of colleges and universities, has been widespread and influential; and their contribution to social service and to the enunciation of the "social gospel" has been considerable.

The church in the South lost its Negro members before and during the Civil War. At the time of the union, the Central Jurisdiction was formed embracing all of the Negro members wherever they lived, alongside the other jurisdictions that were determined by geography. The Central Jurisdiction was abolished in 1968; and black Methodists are now fully integrated in the higher departments of the church's life, though by no means in all the local congregations.

Merger with United Brethren

The Evangelical United Brethren Church, itself a union of the Church of the United Brethren in Christ and the Evangelical Church, was united with the Methodist Church in 1968 to form the United Methodist Church. Its origins lay in the work of German-speaking Protestant ministers who had been attracted to Methodist doctrines and church order in the early 19th century but had not been able to bring their people into the Methodist Church because of a Methodist rule that only the English language could be used in church services. With the disappearance of the language difficulty, the way to union was open, unimpeded by any doctrinal or liturgical difficulties. In 1970 the United Methodist Church had a membership of about 11,000,000.

Methodism in Canada. Most of the early Methodist preachers in Canada came from America, but the War of 1812 led to the coming of preachers from the British Wesleyan Church and the subsequent arrangement by which Upper Canada was subject to the General Conference of the U.S. and Lower Canada to the British Conference. In 1828 the church in Upper Canada became independent and in 1833 was united with the British Conference. The union was dissolved in 1840 and restored in 1847. A further union with the smaller Methodist bodies working in Canada was consummated in 1874, producing the autonomous Methodist Church of Canada.

Discussions with the other nonepiscopal churches with a view to union began in 1908, and in 1925 the United Church of Canada came into existence. Some Presbyterians remained outside, but in recent years the Episcopal Church has begun union talks with the United Church.

Missions and ecumenism. *Methodist world missions.* Thomas Coke initiated the missionary activities of British Methodism by his eloquence and ceaseless travels. The first area in which they took root was the West Indies; then followed Sierra Leone and southern Africa. The Gold Coast, French West Africa, and Nigeria received missionaries not much later, though the climate in many parts of Africa took a heavy toll of missionary lives.

Methodism in India

In India converts were very few until about 1880, when a mass movement swept many thousand low-caste Indians in the south into the Methodist and other churches. In China missionary work had a checkered career, though there were mass movements there also. The last English missionary left China in 1949. In Australia the Methodist Church began in 1815 and, like the Methodist Church in South Africa, became independent before the end of the 19th century. The movement toward autonomy became a flood after World War II; only a few small churches remain under the control of the Methodist Missionary Society. Most of the autonomous churches are negotiating for united churches in their countries; and the Church of South India, including Anglicans, Methodists, Congregationalists, and Presbyterians, has been in existence since 1947.

American Methodists have been equally enthusiastic for missionary activity, and their greater resources have carried them over still larger areas of the earth's surface. North India, Mexico and most of the other countries of Latin America, Cuba, Korea, Japan, Taiwan, and many parts of Africa not reached by the British possess Methodist churches of the American tradition. The movement toward autonomy was somewhat slower in starting in these areas than in the British sphere of influence, but is now well under way; and the General Conference of the United Methodist Church, which is responsible for them all, is formulating elaborate plans for combining fraternal relations among them with their newly found and highly valued independence.

There are minority Methodist churches in most European countries. The one in Italy is of English origin, that in Germany is of mixed English and American origin, though ordered in the American manner; the rest are all derived from American Methodism, though they exhibit many similarities in spirituality to the English type.

Methodism in the ecumenical movement. Methodists from all countries have played a full part in worldwide ecumenical activities, and the Methodist churches have joined in all union negotiations. In America the church responded to the call to discuss reunion contained in the Blake–Pike proposals of 1960, in which the Presbyterian leader Eugene Carson Blake called on the Methodists and other denominations to join the Episcopalians and United Presbyterians in planning a union of churches that was to be "truly Catholic and truly Reformed." The Methodist Church subsequently became involved in the Consultation on Church Union formed to accomplish this goal.

Teaching, practice, and organization. All Methodist churches profess allegiance to the Scriptures and to the

doctrines of the historic creeds and hold themselves to be in the tradition of the Protestant Reformation. They emphasize, though less now than in the past, the teaching about assurance and Christian perfection that is associated with John Wesley. Especially in America, there is a strong tendency to interpret this teaching as an insistence on the personal aspects of religion and as leading to its social application. The worship of both Methodist traditions is partly liturgical, partly spontaneous. In Britain many of the services in the Anglican *Book of Common Prayer*, more or less modified, have remained in use, especially the Order of Holy Communion. The *Book of Offices* (1936) marked a revision of liturgies in accordance with the liberalism then prevalent. Since then the Liturgical Movement has greatly affected all the churches, and all Methodist liturgies are again in processes of drastic revision. The majority of services, however, still follow the old pattern of interlaced hymn, prayer, lessons, and sermon, though experimental worship is on the increase. Wesley adapted *The Book of Common Prayer* for use in America as *The Sunday Service of the Methodists in North America*. This has been frequently revised, and the latest revision, *The Book of Worship for Church and Home* (1966), shows the influence of the Liturgical Movement. There is great freedom in nonliturgical services. Hymns are important in all branches of Methodism, and in Britain those of Charles Wesley are still dominant.

In all the churches of the British tradition the annual Conference is the supreme authority for doctrine, order, and practice. All ministers have parity of status, but special functions are exercised by the president and secretary of the Conference, the chairmen of districts, the secretaries of departments, theological tutors, and superintendents. District affairs are regulated by Synods, Circuits by Quarterly Meetings, local Societies by Leaders' Meetings. The American tradition is episcopal; the bishops are elected by the Jurisdictional Conferences, which, like the General Conference, meet every four years. Each diocese has an annual Conference and is divided into District Conferences, each with its superintendent. The dioceses are combined into five Jurisdictions that cover the nation. The circuit system is not developed as it is in England. Each local church has its Church Local Conference and various Boards. A minister is ordained first deacon, then elder.

In America the African Methodist Episcopal Zion Church and the African Methodist Episcopal Church (each about a million strong) antedate the explosion of the slavery question; the Colored (now "Christian") Methodist Episcopal Church was founded as a result of it. All three are exclusively Negro but follow the doctrine and organization of the United Methodist Church.

BIBLIOGRAPHY. R.E. DAVIES and E.G. RUPP (eds.), *History of the Methodist Church in Great Britain*, vol. 1 (1966), authoritative for the period till the death of Wesley; E.S. BUCKE *et al.* (eds.), *The History of American Methodism*, 3 vol. (1964), comprehensive and authoritative; V.H.H. GREEN, *The Young Mr. Wesley* (1961), a critical study of Wesley's Oxford days; MARTIN SCHMIDT, *John Wesley: A Theological Biography*, 2 vol. (1962–71); A.C. OUTLER (ed.), *John Wesley*, 2nd ed. (1970), theological introduction and select writings; C.W. WILLIAMS, *John Wesley's Theology Today* (1960), a systematic exposition of theological ideas; W.R. CANNON, *The Theology of John Wesley* (1946), a comprehensive account with emphasis on the doctrine of justification; U. LEE and W.W. SWEET, *Short History of Methodism* (1956), concerned mainly with the American form of Methodism; R.E. DAVIES, *Methodism* (1963), a survey, chiefly from the British point of view, and *Methodists and Unity* (1962), relation of British Methodism to the ecumenical movement; G.F. MOEDE, *The Office of Bishop in Methodism* (1964), a history of constitutional developments in American Methodism; G.G. FINDLAY and W.W. HOLDSWORTH, *History of the Wesleyan Methodist Missionary Society*, 5 vol. (1921–24), comprehensive up to 1914; W.C. BARCLAY, *History of Methodist Missions*, 3 vol. (1949–57), the American counterpart to the last-mentioned book; J.H.S. KENT, *The Age of Disunity* (1966), on division and reunion in British Methodism; J.M. MOORE, *The Long Road to Methodist Union* (1943), reunion of American Methodism; CHURCH OF ENGLAND, PROVINCE OF CANTERBURY, *Conversations Between the Church of England and the Methodist Church: A Report* (1963); ANGLICAN-METHODIST UNITY COMMISSION, *Towards Reconciliation* (1967); M. DEANESLY and G.G. WILLIS, *Anglican-Methodist Unity: Some Considerations Historical and Liturgical* (1968).

(R.E.D.)

Metternich

The Austrian statesman Klemens Fürst von Metternich, the most resolute champion of conservative principles of his time, played a decisive role in bringing about the victorious alliance against Napoleon, took a leading part in the Congress of Vienna, and brought Austria to the front rank among the European powers. The descendant of an old Rhenish noble family, Metternich was born at Coblenz, in the electoral territory of Trier, on May 15, 1773, the son of Franz Georg Karl, Graf von Metternich-Winneburg and Countess Beatrix Kagenegg. His father was then the Austrian envoy to the Rhenish principalities of the empire, and Metternich spent his youth in the Rhine–Moselle region, for which he retained a lifelong affection.

By courtesy of the Staatliche Kunstsammlungen, Dresden

Metternich, black and white chalk drawing by Anton Graff, *c.* 1803-05. In the Kupferstichkabinett, Dresden, East Germany.

Early years and first diplomatic position

In 1788 he entered the University of Strasbourg, where he studied diplomacy; but the spread of the French Revolution prompted him to leave Strasbourg in 1790 and enter the University of Mainz. Before the French revolutionary troops entered Mainz, he went to Brussels in the Austrian Netherlands, where his father was then chief minister. In 1794 he undertook a diplomatic mission to England, where he published a pamphlet calling for a general arming of the German people; but in October he rejoined his father, who had in the meantime fled to Vienna as the French invaded the Netherlands. In Vienna he occupied himself with natural, scientific, and medical studies, in which he always kept a lively interest and which he later did much to encourage.

In September 1795 Metternich married the countess Eleonore Kaunitz, heiress and granddaughter of the former Austrian state chancellor Wenzel Anton, Graf von Kaunitz. This marriage gave him the link with the high nobility of Austria and the access to high office he had long desired. After having represented the Catholic Westphalian counts of the empire at the end of the Congress of Rastatt (1797–99), which ratified compensation for the German princes ousted by the French from their possessions on the left bank of the Rhine, he was in 1801 appointed Austrian minister to the Saxon court at Dresden; and there he formed his friendship with Friedrich von Gentz, the German publicist and diplomat. Serving as Austrian minister in Berlin after 1803, Metternich failed to persuade Frederick William III of Prussia to join Austria in the war of 1805 against France but gained a profound insight into the internal brittleness of the

Prussian state, whose speedy ruin he predicted. While he was in Berlin, he attended lectures by Johann Gottlieb Fichte, the German philosopher, and the poet and critic August Wilhelm von Schlegel.

Ministry during the Napoleonic Wars. In 1806 Metternich served as Austrian minister to France. In contact with Napoleon's sister Caroline Murat and other ladies of Parisian society, he won a reputation for licentiousness; but nevertheless from those ladies and from his relations with the foreign minister Talleyrand and with the Russian envoy, he obtained excellent reports on the state of affairs in France. Although Metternich's successes in the negotiations leading up to the Franco-Austrian Treaty of Fontainebleau were insignificant, he used his time to acquire a deep insight into the emperor Napoleon I's character. Yet he overestimated the impact of the Spanish rising of 1808 on the Napoleonic system, and his optimistic reports did much to induce Austria to undertake the disastrous War of 1809 against France. After the Battle of Wagram, he tried to obtain favourable terms in the peace negotiations but was rebuffed by Napoleon.

Minister for foreign affairs

On October 8, 1809, the emperor Francis (at that time Francis I of Austria but no longer Holy Roman emperor) appointed Metternich minister of foreign affairs. Six days later the oppressive Treaty of Schönbrunn was signed with France. Austria was now in urgent need of a respite, which Metternich obtained by forming the project of a marriage between the archduchess Marie-Louise, a daughter of Francis I, and Napoleon, whose vanity Metternich cleverly exploited. It is not clear how far he expected that this marriage would restrain Napoleon from further campaigns of conquest, but at least he achieved a relationship between France and Austria loose enough to preserve Austria's freedom of action: Austria neither joined the Confederation of the Rhine, a league of German princes under Napoleon's protection, nor became one of the client states of the Napoleonic system. Utterly exhausted and debt ridden, Austria could hardly have resisted any further demands of Napoleon, but it was then no longer the main object of Napoleon's hostility.

As early as 1811, in order to promote Austria's internal development, Metternich wanted the state to be reorganized on federal lines instead of continuing under the centralized system that the emperor Joseph II had imposed. Yet Metternich could never overcome the objections of his strictly absolutist emperor. At the same time the enthusiasm for arming the nation and for a German national rising against Napoleon that he had felt as late as 1809 began to be superseded by a firm dislike for all popular movements. Agreeing with the Emperor on this, he now came to regard these manifestations as a menace to the multinational Habsburg state. He became the strictest exponent of the doctrine of the balance of power in Europe—a doctrine instilled into him originally by Koch, latterly by his diplomat friend Gentz.

When Napoleon launched his invasion of Russia in 1812, Metternich obtained the status of an independent contingent for the Austrian forces under Prince Karl Schwarzenberg that accompanied the French Army. The disaster that befell Napoleon's army came as a surprise to Metternich. On January 30, 1813, Schwarzenberg concluded an indefinite armistice with the Russians. But in view of the inadequacy of Austria's armaments, Metternich could not make up his mind to change over to war on Russia's side against Napoleon. Resisting all ill-considered projects, in particular those of the archduke John (who was put under house arrest for planning a premature anti-French rising in the Alps), Metternich firmly adhered to neutrality while Austria secretly rearmed. He even drew Saxony into the neutral camp for a time. When, later in 1813, Saxony's return to the French side and Napoleon's victory over the Russians and Prussians at Bautzen shook Metternich's will to make war and stiffened Napoleon's attitude, Metternich mediated an armistice between France, Russia, and Prussia. Even so, in the subsequent Treaty of Reichenbach, June 24, 1813, between Austria, Prussia, and Russia, Metternich undertook to bring Austria into the war against France if Napoleon rejected the peace terms that he was offering.

Metternich's middle course

By dominating the negotiations with the French during the summer of 1813, Metternich gained more time for rearming. At this point he was not interested in the annihilation of Napoleon's power, which the emperor Francis likewise, out of consideration for his daughter Marie-Louise, was unwilling to destroy altogether. Metternich also distrusted the Russian emperor Alexander I and feared that after the collapse of France, Europe would be at Russia's mercy. Napoleon's obstinacy frustrated the attempt at a settlement; but when in August Austria finally declared war on France, Metternich, by his superior conduct of negotiations, had won for his country the leadership both in the political and in the military field. In October 1813 the hereditary title of prince was bestowed on him by the Austrian emperor.

In opposition to the plans of the Prussian minister Karl, Freiherr vom Stein, and of the Russian emperor, Metternich promised the South German states of the Confederation of the Rhine that if they went over to the allies they would not forfeit the position they had achieved on Napoleon's side. This promise alone showed that while he was striving for a solution compatible with the interests of all parties, he also wanted to gain the South German states as allies against the Prussian–Russian designs of aggrandizement. After Napoleon's defeat at Waterloo and abdication, Metternich rejected as unrealistic the proposals of Baron Stein and others for the resuscitation of the Holy Roman Empire. The first Treaty of Paris (May 30, 1814) stipulated nothing more for Germany than a loose confederation of states.

Leadership of the Congress of Vienna. The Congress of Vienna (September 1814–June 1815) was the climax of Metternich's work of reconstruction. The very fact that it was held in Vienna was in itself a great success for him. He had precise ideas about the basis for a new order in Europe but knew from the start that he would have to modify them substantially if he was to salvage even a small part of his plans against the opposition of self-interested princes. He wanted to secure Austria's predominance by forming two confederations, one German and the other Italian, with Austria as the leading power in both. Within Germany, he proposed the creation of a hereditary German imperial title, and he thought that Austria and Prussia should share the task of protecting Germany's western frontier. Friendship with Prussia on the one hand and with Bavaria on the other thus seemed to him to be the prerequisite of success. Supported by the British foreign secretary Lord Castlereagh, Metternich sought to prevent the elimination of France, which he saw as a necessary counterweight against Russia. Likewise, he resisted the territorial aggrandizement of Russia and Prussia and objected in particular to Prussia's designs for annexing all of Saxony.

The congress became a splendid social event. By an unbroken chain of festivities Metternich kept the visiting monarchs in a mood that made them disinclined to interfere very persistently in the real work of the statesmen. Facile and not averse to amorous adventures, Metternich brilliantly mastered his dual role of social representation and political leadership.

Partial success of Metternich's policy

Yet Metternich only partly succeeded in his plans: the German imperial project came to nothing because Francis steadfastly refused to support it; the Italian confederation did not materialize; and the German confederation, when it at last did come into being in June 1815, was based only on a brief and noncommittal federal act derived from a Bavarian compromise proposal. In European affairs, however, Metternich was more successful: he achieved equality of status for France; he obtained a reduction of the Prussian demands on Saxony; and, in particular, he blocked the farther reaching demands of Russia. Both Russia and Prussia, in fact, were held in check by the common front of Austria, England, and France that Metternich had created.

Metternich's moderation produced a long-lasting European order. This, however, must be ascribed to his diplomatic capability rather than to his political foresight. Austria's status in the German confederation had been strengthened; but the Emperor's refusal of the German

crown meant that Prussia, with equal status in the confederation, would be able to counterbalance Austria.

Role in the German confederation. Not even within the Austrian Empire was Metternich able to prevail with the idea (already at the root of his plans of 1811) of overcoming the spirit of national revolution by revitalizing the old historical regions and the privileges they enjoyed in preabsolutist times. His attempt to organize the newly acquired Italian provinces according to historical principles was frustrated by the emperor Francis, who, though accepting Metternich's ideas, united two incompatible regions in a completely unhistorical "Lombardo-Venetian kingdom" and so destroyed Metternich's hopes of counteracting pan-Italian nationalism. Moreover, as Metternich had feared, the initially strong pro-Austrian mood turned into its opposite.

The reconstruction of Austria took shape entirely in the spirit of the emperor Joseph II, on centralist and absolutist lines, without regard to national differences and without the establishment of departmental ministries, which Metternich had demanded. After many futile remonstrances Metternich eventually yielded to the obstructionism of the Emperor, who detested innovation and stood jealously on his dignity. The re-establishment of the ancient diets of the estates of Tirol and Galicia complied to some extent with Metternich's idea of resurrecting the provincial diets in order to create a counterweight against the growing forces of liberal and nationalist opinion that was demanding a central parliament. He condemned, however, the repressive measures by which the police minister tried to achieve these aims.

The domestic affairs of Austria created difficulties for Metternich at the Frankfurt am Main Bundestag (federal assembly), which opened in 1816. He had originally intended to use this assembly to oppose revolutionary thought all over Germany. Pointing to the examples of Tirol and Galicia, he attempted in 1817–18 to encourage the German states to introduce constitutions resurrecting the historical provinces and to set up their own diets instead of a central parliament. In summer 1818, however, Bavaria and Baden promulgated constitutions that reflected not Metternich's ideas but those of limited monarchy similar to that outlined by the French charter of June 1814; and in 1819, when revolutionary activity culminated in the murder of the dramatist August von Kotzebue and when the opening sessions of the Bavarian and Baden assemblies proved stormy, Metternich decided to stifle these unmanageable liberal currents.

Failure of Metternich's constitutional ideas

He managed to convince the Prussian chancellor Prince Karl August von Hardenberg that his prescription for provincial diets was right. Then, assured that Prussia would not follow the South German example, he could quietly watch how parliaments created against his advice fulfilled his predictions and discouraged the liberal inclinations of German princes. Consequently, at the ministerial conferences of Carlsbad and Vienna in 1819–20, Metternich, to the surprise of the South German states, did not attempt to undo the new constitutions but simply curtailed the activity of the federal assembly, which had become an inconvenience to him. Reorienting his German policy, he began to rely not on the assembly but on the common interest of the princes whom he led to share his point of view by personal contact. Henceforth it was no longer the privileges of Austria as granted by the federal act, but Metternich's personality that guaranteed Austria's predominance in the German confederation.

Years of decline. Metternich had hoped that a system of congresses, at which the great powers would concert their actions, would maintain order and peace in Europe. At the congresses of Aix-la-Chapelle (1818), of Troppau (1820), of Laibach (1821), and of Verona (1822) his international reputation was at its zenith, but the disruption of the forum of great powers became evident when Great Britain abandoned the policy of intervention against revolutions in other countries: Viscount Castlereagh prepared the way for this change at Troppau, and George Canning, his successor as British foreign secretary, brought Metternich's influence on western Europe to an end by insisting on the right of national self-determination for the South American colonists in revolt against Spain and for the Greek insurgents against Turkey. With Alexander I's death (1825) it seemed likely that Metternich's influence on Russia would likewise come to an end; and Prussia's jealousy of Austria's dominance was causing further difficulties, when in 1830 the July Revolution in France, followed by insurrections in Belgium, in Poland, and in Germany, appeared to justify again Metternich's dismal prognoses and served to convince the three eastern powers, Austria, Prussia, and Russia, that they should stand together by his principles.

Effect of Britain's new policy of non-intervention

Metternich had been appointed Austrian state chancellor on May 25, 1821, but his influence in Austria was decisively restricted by the appointment of Franz Anton, Graf von Kolowrat, as minister of state and head of the cabinet conferences (1826). Kolowrat opposed an orderly governmental organization and won great ascendancy over the emperor Francis. Metternich, as his influence dwindled, yielded to an increasingly irritating vanity and to a passion for theorizing that caused his utterances to become increasingly prolix and, at times, to verge on the ridiculous. For the sake of legitimacy and despite general apprehensions, he gave Francis the disastrous advice to recognize his feebleminded eldest son, the archduke Ferdinand, as heir to the throne. In 1835, therefore, when Ferdinand succeeded his father, Metternich, at first together with the archduke Louis (Ludwig), took the chair at the "conference" or council of state that assumed the functions of a regency. In 1836 it seemed as if he would at last be able to carry out his idea of well-ordered government, but at the decisive moment Kolowrat managed to convert the archdukes John and Louis to his own theories. Thenceforth Metternich's authority was confined to external affairs. His vanity tempted him to disguise the waning of his influence by accepting responsibility for decrees that neither came from him nor accorded with his views. He thus became a hated symbol of repression and reaction and, eventually, on March 13, 1848, had to resign, as the first victim of the revolution. He made his way with difficulty into exile in England but returned to Vienna in 1851. Having taken no further direct part in politics, Metternich died in Vienna on June 11, 1859.

The marriages of Metternich

After his first wife's death in 1825, Metternich had married Baroness Antoinette Leykam in 1827. After her death in 1829, he had married Countess Melanie Zichy-Ferraris in 1831, who died in 1854. His son by his second marriage, Richard, Fürst von Metternich, was Austrian ambassador in Paris from 1859 to 1870 and one of the foremost diplomats of his time.

Assessment. Metternich's personality has long been the subject of dispute. His great merits cannot be denied. On the other hand, his failure to overhaul the government and administration of the Habsburg monarchy remained an omission that could never be compensated. He always went the way of least resistance and so allowed others to turn Austria into a prison in which free speech was impossible and any denunciation was enough to cut short a promising career. Although on several occasions he put before the Emperor projects that show his insight into the growing difficulties, he never drew the least consequence from the Emperor's rejection of his plans. Even in the regency council he was content to play the part of the philosopher. The complaisance associated with his predominantly diplomatic talents prevented him from playing the part of the great Austrian statesman.

The rapid decline of Austria's institutions from 1826 sapped Metternich's international standing, but his vanity would not let him see how hollow the foundations of his prestige and of his system had become. He considered himself the "father confessor of cabinets" and the "physician of sick governments in the entire world." The "Metternich system," which he himself so often cited and about which so much has been written, postulated the existence of eternally valid maxims of supreme wisdom in statesmanship, the most important of them being the European balance of power, sole guarantee of international peace. For Metternich, the old cultural and political order of Europe was a living organism that he wanted to preserve

against the destructive forces of revolution. His convictions were strengthened by the ease with which his statecraft could circumvent the liberals of the time, whose exaggerated manner of expressing themselves often had an air of dilettantism and whose political immaturity seemed to him to threaten the delicate mechanism of the balance of power. He therefore founded his system on restraint of all those imponderable forces that could not be reconciled with the old-established art of statesmanship: imponderables such as the coming-of-age of peoples, the awakening of nations, and the emergence of social problems.

The struggle against the spirit of the time and against the political aspirations of the bourgeoisie was for Metternich also a struggle against a rising tide of political dilettantism. He hoped that economic affluence would keep the bourgeoisie away from politics, and the return to preabsolutist conditions seemed to him the necessary basis for antirevolutionary action. Liberalism and the revolutionary spirit he merely regarded as a not entirely unjustified reaction against the absolutist anarchy for which the emperor Francis stood and which he, Metternich, condemned. In this absolutist anarchy he saw the root of all evil. It was his tragedy and also his fault that he never built a foundation on which his system could really develop and prove its worth. After the Revolution of 1848 Metternich asserted again and again that his influence on Austria's domestic policy had been insignificant. Although his statement is probably true, it cannot be conclusively confirmed; the records of the Austrian Council of State were lost in a fire in 1927.

Assessment

From the time of Joseph II there had been a discrepancy between the specific concerns of Austria and the old universalist tradition of the Holy Roman Empire; and the same discrepancy was evident between Austria's internal policy and Metternich's foreign policy, which was devised to assume responsibility for all of Europe. Whereas Francis and his entourage thought along narrow Austrian lines, Metternich stood entirely in the tradition of the universal monarchy—not so much from devotion to the past glory of the emperors as from the conviction that the revolutionary spirit was an international disease that could be fought successfully only by disregarding national frontiers. For ten years after Napoleon's downfall Metternich, by sheer force of personality, made the world think of Austria as a real power: his failure came when the feet of clay could no longer be hidden. For Germany, the fatal effect of his influence was loss of contact with the liberal and progressive West. In recent years, research on Metternich has mainly been pursued by United States historians, some of whom still consider Metternich an unmitigated reactionary. Whether he was or not can only be established with certainty by examination of the entire body of relevant documents in Vienna, in particular of Metternich's instructions to the Austrian ambassadors, in which he often freely expressed his thoughts.

BIBLIOGRAPHY. The standard biography is H. VON SRBIK, *Metternich der Staatsmann und der Mensch*, 3 vol. (1925–54); an opposing and unfavourable view of Metternich is given in V. BIBL, *Metternich, der Daemon Oesterreichs* (1936). The works by E.L. WOODWARD, *Three Studies in European Conservatism: Metternich, Guizot, the Catholic Church in the Nineteenth Century* (1929); A. MAY, *The Age of Metternich, 1814–1848* (1933); A. CECIL, *Metternich, 1773–1859* (1933); and E. LAUBER, *Metternichs Kampf um die europaeische Mitte: Struktur seiner Politik von 1809 bis 1815* (1939), have pursued lines of inquiry opened up by Srbik. More recent studies have offered various other views of Metternich. Among them are E.E. KRAEHE, *Metternich's German Policy*, vol. 1, *The Contest with Napoleon, 1799–1814* (1963); and G. DE BERTIER DE SAUVIGNY, *Metternich et la France après le Congrès de Vienne*, vol. 1, *De Napoléon à Decazes 1815–1820* (1968), vol. 2, *Les Grands Congrès—1820–1824* (1970). The dominant position of Metternich in Europe is treated in R.A. KANN, "Metternich: A Reappraisal of His Impact on International Relations," *Journal of Modern History*, 32:333–339 (1960); P.W. SCHROEDER, *Metternich's Diplomacy at Its Zenith, 1820–1823* (1962); H.F. SCHWARZ (ed.), *Metternich, the Coachman of Europe* (1962); D.E. EMERSON, *Metternich and the Political Police: Security and Subversion in the Habsburg Monarchy (1815–1830)* (1968); and A.G. HAAS, *Metternich, Reorganization and Nationality, 1813–1818* (1963), which describes a fundamental aspect of Metternich's plans for internal reform in Austria.

(K.O.v.A.)

Mexico

Mexico is a federal republic in North America. Officially known as the United Mexican States (in Spanish, Estados Unidos Mexicanos), it covers a total area of 761,600 square miles (1,972,544 square kilometres). It is bordered by the United States to the north, the Gulf of Mexico and the Caribbean Sea to the east, British Honduras and Guatemala to the southeast, and the Pacific Ocean to the south and west. Mexico in 1971 had a population of more than 50,000,000; its rate of population growth—about 3.5 percent a year—is one of the highest in the world.

This "land of great solitudes and the baroque," as it has been called, is geographically positioned in the Earth's great desert region. No major river crosses it, and water is scarce throughout most of its territory. Areas of benign, temperate climate are determined by altitude rather than latitude. One such area is the Mexican Plateau, the country's most densely populated region. Mexico City, the capital, with a population of about 8,600,000 (1970) is located here, in clear view of the peaks of Popocatépetl, rising 17,883 feet (5,452 metres), and Ixtaccihuatl, 17,338 feet (5,286 metres); it is the only megalopolis in the world without either direct access to the sea or an important river passing through it.

Cultural unification

Historians and sociologists are presently agreed that transculturation (the transmission of cultural characteristics from one cultural group to another) between the mestizos (of mixed Indian and European blood) and the criollos (Europeans born in America) has reached noteworthy proportions. The statement that "even though there may not be Indian blood in one's veins it is always present in one's spirit" is also true in the opposite direction, in that the Indian cannot escape absorbing European culture.

Cultural unification of the population has been aided by the existence of excellent air and land transport networks, connecting the 29 states, two territories, and the Federal District (Mexico City), which together make up the Mexican federal republic. Throughout the past four decades uninterrupted peace has permitted marked economic growth—calculated in 1970 to be occurring at a rate of 7.7 percent. Even though foreign investment is continuously growing, it has not dampened domestic saving, which has also greatly increased; a growing degree of industrialization is consequently taking place. This last factor, together with continuing governmental concern since the 1920s for the improvement of education, as well as a constant decrease in the degree of interference in national affairs by the church and the military since that time, have permitted consolidation of social and political stability (see MEXICO CITY; for associated physical features, see CARIBBEAN SEA; MEXICO, GULF OF; and RIO GRANDE; for historical aspects, see MEXICO, HISTORY OF).

THE LANDSCAPE

Mexico is a country of extraordinary physical variety. In many ways it is a bridge between the natural phenomena of the broad North American continent, characterized by vast ranges and plains and by climatic extremes, and the tropical features of Central America and the Caribbean region.

Mexico is widest in the north, where the frontier—running roughly northwest to southeast, from the Pacific to the Gulf of Mexico—covers a distance of about 1,300 miles. To the south, the width of the country diminishes until the Isthmus of Tehuantepec is reached; the isthmus is a neck of land about 140 miles wide, running from west to east and separating the Pacific to the south from the Gulf of Mexico to the north. To the east, the isthmus broadens toward the Guatemalan and British Honduran borders: Mexico's eastern extremity consists of the Yucatán Peninsula, which projects northeastward; it is about 200 miles wide and has a coastline 700 miles in length. One of the most striking features of Mexico is the Baja

California Peninsula. From Mexico's northwestern extremity, this long fringe of land, from 25 to 150 miles wide, runs parallel to the Pacific coast in a southeasterly direction for about 760 miles. It is separated from the mainland proper by the Gulf of California, a long arm of the sea.

About 300 miles to the south of the Baja California Peninsula and about 450 miles west of the coastal port of Manzanillo lie the Islas de Revillagigedo—a group of volcanic islands.

Relief features. With the exception of the Yucatán Peninsula and the eastern and northwestern coastal plains, which represent no more than 14 percent of its total area, Mexico is a country of broken relief composed of mountain ranges, tablelands or plateaus with bolsons (depressions or valleys having no outlet), and short depressions, canyons, and innumerable valleys. The average altitude exceeds about 3,300 feet, but almost 71 percent of the total area lies at about 1,600 feet above sea level.

Mountain ranges and plateaus. The Sierra Madre Occidental (Western Sierra Madre) and the Sierra Madre Oriental (Eastern Sierra Madre) extend along Mexico's coasts from north to south. In the isolated area between them lie two vast tablelands—the Mesa del Norte (Northern Plateau) and the Altiplanicie Mexicana (Mexican Plateau). These have enclosed basins in Chihuahua and Coahuila, north of the inland towns of Zacatecas and San Luis Potosí, followed to the south by wide valleys such as the Bajío de Guanajuato (Guanahuato Shoal) and the Michoacán.

The Sierra Volcánica Transversal

The Sierra Volcánica Transversal (Transversal Volcanic Mountain Range) is a range of volcanic mountains that extends from central Nayarit state in the west to central Veracruz state in the east. The nation's highest peaks belong to this range—Orizaba, rising to 18,700 feet (5,700 metres), and Popocatépetl, which reaches 17,883 feet (5,452 metres). Its numerous volcanoes enclose a belt of valleys, bolsons, plateaus, and lakes. On its southern slope the Neo-Volcánica Cordillera (New Volcanic Range) plunges abruptly in the direction of the valley of the great Río Balsas, traversing the states of Michoacán, Guerrero, and Morelos. Between the Río Balsas Valley and the Pacific Ocean, the Sierra Madre del Sur Range extends along the southern coast to the Isthmus of Tehuantepec, where it joins the complex mountain range of the Sierra Soconusco and the Chiapas Highlands.

The characteristics of north central Mexico are repeated in Chiapas: a central depression and plateau bounded by mountain ranges on either coast, which descend to the alluvial coastal plains.

The coastal plains. The coastal plain along the Gulf of Mexico is bounded inland by the Sierra Madre Oriental, the Sierra Madre del Sur, and the Chiapas Highlands. It is almost featureless along the Gulf but rises westward to foothills that are usually below 650 feet in elevation. Wide beaches, sandbars, swamps, and lagoons are common. There are no good natural harbours, although the Laguna Tamiahua shelters coastal craft.

The western coastal plain lies between the Sierra Madre Occidental and the Gulf of California and the Pacific Ocean. It is widest in the north, where its broad basins are separated by low hills and mountains. Here the coast is broken by numerous bays and lagoons. The southern portion of the Pacific coast is bordered inland by the Sierra Madre del Sur. It forms a narrow alluvial apron and contains good natural harbours such as those of Acapulco and Manzanillo.

The Yucatán Peninsula. The coastal lowlands attain their maximum width of about 275 miles in the limestone platform of the Yucatán Peninsula. The terrain is generally flat to rolling, and attains the greatest elevation in its centre. There is little surface drainage because precipitation penetrates the porous and soluble limestone. Subterranean erosion has produced caverns and sinkholes; the latter are formed when cavern roofs collapse.

The western and northern coasts of the peninsula are marked by lagoons and sandbars, but coral reefs dot the eastern coast. The islands of Cozumel and Mujeres lie off the peninsula's northeastern tip.

Drainage. Mexico has few important rivers. The humid northeast trade winds, as well as tropical hurricanes, blow onto the exterior slopes of the Sierra Madre ranges, causing great increases in the volume of the mountain rivers, which almost invariably flow through deep canyons into either the Pacific Ocean or the Gulf of Mexico.

The river systems

The central plateau region and the desert-like Baja California area suffer from a lack of large streams. The northern bolsons impede the development of a naturally flowing river network, and the scanty lakes of the plateau's interior watershed are subject to intense evaporation. The Río Conchos, the most important of the plateau's rivers, flows northward into the Rio Grande (*q.v.*; known to Mexicans as the Río Bravo del Norte), which forms part of the frontier between Mexico and the United States and empties into the Gulf of Mexico.

In the Salado area of northwest Mexico, the Chiapas Highlands, and the Sierra Madre Oriental, much moisture drains into the subsoil as a result of the terrain's limestone quality. This also occurs in the northern and central regions of the Yucatán Peninsula, a tropical land lacking fixed rivers.

On the Isthmus of Tehuantepec, however, the Grijalva-Usamacinta and the Papaloapan river systems account for almost 40 percent of the total volume of Mexico's rivers, flowing across the alluvial plain of the Gulf, where flood and swamp conditions prevail because of inefficient drainage. Another 15 percent crosses the valleys of the southern region, and 8 percent flows in the northwestern streams that irrigate the lower valleys of the Río Colorado. The major southern river, the Río Balsas, rises in Puebla state and flows in a generally westward direction across Guerrero state and between Guerrero and Michoacán states to its mouth on the Pacific.

Soils. Lateritic soils (soils with a high content of iron oxides and aluminum hydroxide) are found on the Yucatán Peninsula and on the Pacific coastal strip south of Cabo Corrientes. Yellow and red subtropical forest soils occur on the Gulf coast and on the west coast to the north of Cabo Corrientes. In the central plateau region, the soils gradually change in a northward direction from humid prairie soils to brown soils characteristic of the steppe (extensive and usually treeless plains), and to the gray and brown desert soils of the northwest.

The climate. *General climatic factors.* The Tropic of Cancer cuts Mexico almost in half. This factor permits polar air fronts to invade the north, east, and central regions, a phenomenon responsible for the great drops in temperature and increased rainfall associated with the *nortes,* or northern gales, that occur during the winter season. The northern regions belong to the desert belt of the hemisphere, and their climate is characterized by western anticyclones (systems of winds that rotate about a centre of high atmospheric pressure). Since, however, the country is practically surrounded by the waters of the Pacific, the Gulf of Mexico, and the Caribbean, most of the surface is subject to humid trade winds from the east between May and August and to tropical hurricanes originating in both oceans which may occur from August to October. These winds from the east, southeast, and south follow varied patterns but often cross the high Sierra Madre and deposit most of their moisture on the seaward mountain slopes. The extensive interior plateaus and lowlands thus receive minimal rainfall throughout the year.

Regional variations. Half of Mexico experiences a predominantly arid climate and receives less than 24 inches of precipitation annually. Continental climatic characteristics, such as hot springs and summers and relatively severe winters, are also in evidence. These characteristics are evident in such regions as the Mesa del Norte and Baja California. The degree of drought diminishes in the Mexican Plateau, but aridity is also apparent in the tropical coastal areas of the Balsas Lowlands, the Grijalva Uplands, and North Yucatán.

Except in Baja California, up to 75 percent of the rainfall occurs during the summer and early autumn, usually in the form of storms. The heaviest precipitation is registered on the southern seaward slope of the Sierra Madre

Oriental in north Oaxaca and Chiapas, in the Tuxtlas Massif, and in the Sierra de Soconusco near the Honduran border in the southwest, where the annual rainfall exceeds 100 inches and occasionally reaches 180 inches. The lowest precipitation—two to four inches—falls in the Baja California Peninsula.

Temperature patterns

Average high and low annual temperatures vary considerably between the coastal regions, the valleys at medium altitudes, and the plateaus. In Manzanillo, a port located on the hot, tropical Pacific Coast, the mean temperature is approximately 79° F (26° C), with annual variation fluctuating between 73° and 84° F (23° and 29° C). Tepic, at an altitude of 3,000 feet, has an average of 68° F (20° C) with a greater variation—from 57° to 82° F (14° to 28° C). In the northern continental climate of Chihuahua at 4,700 feet there is a strong contrast between mean temperatures of the months of June (79° F [26° C]) and of January (50° F [10° C]). The difference is less in the high Mexican Basin, at 7,500 feet, where the low reaches 52° F (11° C), the high 73° F (23° C), and the yearly average is 61° F (16° C). The low extremes reach several degrees below freezing in regions of continental climate, with frosts occurring in the plateau area, the arid northwest, and the high Sierra Madres.

Plant and animal life. In the arid northern, northwestern, and central regions, plants structurally adapted to a limited water supply predominate. These include cactus, agave, cassava, mesquite or scrub, and brush plants. In the south and east, in Yucatán, and on the western coasts, tropical vegetation such as dense deciduous or rain forests, savanna (grassy parkland) vegetation, grazing-land vegetation, and spiny plants are evident. The central mountains and the Sierra Madres, at altitudes of between 5,000 and 13,000 feet, support predominantly coniferous and mixed forests in which the holm, or German oak, is widespread. The mountain slopes contain regions of complex transition between different vegetation types; above the pine zones, cold pasture lands appear on the slopes of the large volcanoes. On the coasts, large mangrove swamps are common.

Cactus species are especially varied, particularly on the southern plateau and in Baja California. Great variety also characterizes the tropical forests, which contain valuable hardwoods, epiphytes (plants that derive their moisture and nutrients from the air and rain), and palms. From the southeast of Veracruz to the south of Yucatán, the plant life is unusually rich and varied.

Animal life includes examples from both the North and South American regions; the approximate dividing line between the two regions passes, at medium altitudes, along the seaward slopes of both the Sierra Volcánica Transversal, and the Sierra Madres. Species, however, are mixed in the Balsa Depression, the southern part of the Central Plateau, and in mountain areas.

Birds, reptiles, insects, and a variety of mammals—including rodents, monkeys, armadillos, tapirs, anteaters, pumas, and wolves—are plentiful. The nation's waters, especially the Gulf of California and the Pacific northwest, contain a wealth of fish and crustaceans.

Traditional regions. Various geographic, ethnic, and historic influences, together with a limited economic development, have resulted in the existence of a vast number of small natural, historic, and economic regions. Similarity of conditions permits the grouping of many of these small regions into larger ones. The "traditional" regions, which are enthroned in popular folklore, do not, however, possess clearly defined borders.

Baja California. A desert climate prevails, except in the high, temperate areas of the major mountain ranges of Sierra de Juárez and Sierra San Pedro Mártir and in the subtropical area of the far south. While the area, therefore, is not forested, those parts of it bordering the Pacific Ocean and the Gulf of California, as well as those parts watered by the Lower Colorado River or subterranean rivers, are rich in vegetation. Because the region borders the United States, it has a completely mixed population. The people are concentrated in oases; in agriculturally developed and irrigated areas, such as the Mexicali Valley and Santo Domingo; in the border belt with its tourism and trade centres, such as Tijuana and Ensenada; or in the numerous fishing ports along the coasts.

The northern regions. The northwest is an extensive area, stretching between the Sierra Madre Occidental to the east and the Gulf of California to the west and between the Desierto de Altar in the north and the tropical plain of Nayarit state in the south. High population densities occur in the highly productive irrigated valleys and in prosperous commercial and industrial border areas, such as Nogales–Cananea. Small Indian farming communities are located in the Yaquí, Mayo, and Fuerte valleys.

The nation's largest region, the north, is cut off from both coasts by the Sierra Madre ranges; it extends from the United States south through Zacatecas and northern San Luis Potosí states. It is an arid area, broken by bolsons and mountains. The plateau regions are virtually uninhabited, and population is concentrated in such irrigated areas as La Laguna, Delicias, and Juárez Valley; in the numerous mining areas from Parral and Durango to Nueva Rosita; and especially in such industrial cities as Monterrey.

The northeast is located between the Sierra Madre Oriental and the Gulf of Mexico and stretches from the United States border to Tampico. Population concentrations occur in valley regions; the agricultural, commercial, and oil-producing area of Bajo Bravo; the plantation areas of Montemorelos-Victoria; and the southern subtropical regions containing the industrial cities of Mante and Tampico. As in the north, the region is primarily inhabited by mestizos and criollos.

The western and central regions. In the western and central regions, where less arid conditions prevail and where black soil is often found at higher altitudes, there is a high rural population density in areas such as the Bajío, the Tarasca Plateau, central Jalisco state, and the southern part of the San Luis-Aguascalientes area. The rural landscape is being transformed, however, by the accelerated growth of such cities as Guadalajara, Léon, Salamanca, and Querétaro. Many rural migrants have been attracted to the volcanic regions of the high central valleys, where industry is located in or near Mexico City, Puebla, Toluca, and Sahagún. Although racial mixing has been intense, rural pockets of Indian groups may still be found in Michoacán, Mezquital, Puebla, and Toluca. As a result, there are great contrasts between the life-style of the rural inhabitants of the valleys and mountains and that of the urban populations.

The southern regions. The south is an extremely mountainous region that stretches along the Pacific Ocean from Jalisco state to the Isthmus of Tehuantepec and also extends inland to Puebla state. Arid tropical conditions predominate in the narrow valleys of Guerrero and southern Puebla states, while more humid conditions are found in the Sierra Madre del Sur (southern Sierra Madre), and in the state of Oaxaca. A land of both *minifundios* (small—often very small—farms) and commercial plantations, the south is primarily inhabited by Indians. There are only a few important cities, among which are the communications and tourist centre of Acapulco and the commercial centre of Oaxaca.

The Isthmus of Tehuantepec

The narrow Isthmus of Tehuantepec presents two different physical regions. Veracruz state, to the north, has a damp tropical climate and is populated mainly by mestizos. The state of Oaxaca has a semi-arid climate and an undeveloped agricultural economy and is marked by the cultural influence of the Zapotec Indians. It is sparsely populated; some communities, such as those in the Coatzacoalcos Mountains, are virtually isolated.

The Huastecas region covers territory in the states of San Luis Potosí, Veracruz, Puebla, Hildalgo, and Querétaro. It is composed both of the complicated mountain systems of the Sierra Madre Oriental and the flat coastal lands of Veracruz state. Its population is composed of Indians in remote areas and mestizos in developed areas, such as Valles, Tuxpan, Pánuco, and the oil fields of Naranjos-Pozo Rica and Ebano.

The east, or Gulf region, stretches along the Gulf of

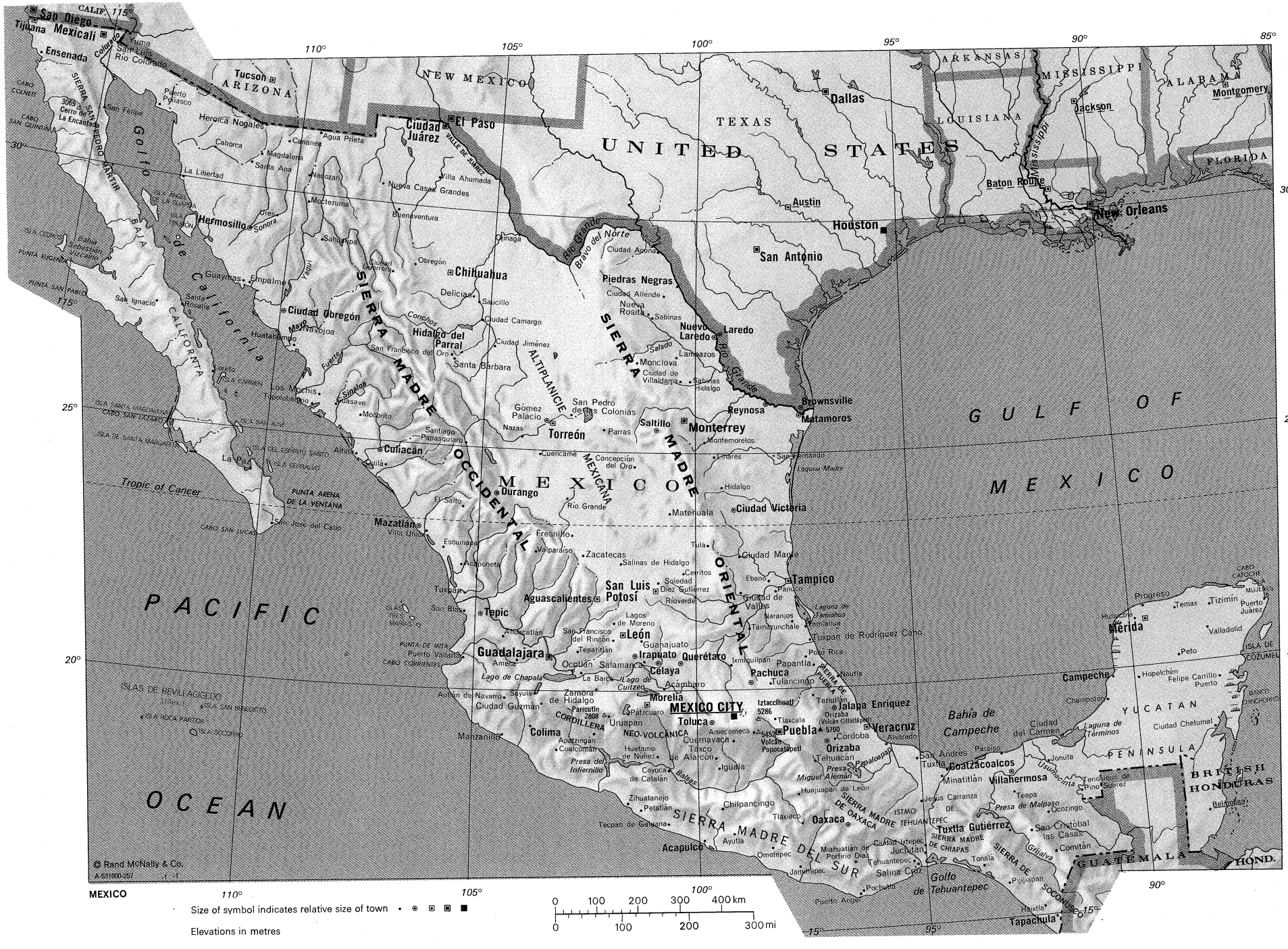
MAP INDEX

Cities and towns

Acámbaro.....20·02n 100·44w
Acaponeta.....22·30n 105·22w
Acapulco [de Juárez].....16·51n 99·55w
Agua Prieta....31·18n 109·34w
Aguascalientes.21·53n 102·18w
Ahuacatlán.....21·03n 104·29w
Altata.....24·38n 107·55w
Alvarado.....18·46n 95·46w
Ameca.....20·33n 104·02w
Amecameca [de Juárez].....19·07n 98·46w
Apatzingán.....19·05n 102·21w
Autlán de Navarro.....19·46n 104·22w
Ayutla.....20·07n 104·22w
Buenaventura..29·51n 107·29w
Caborca.....30·37n 112·06w
Campeche.....19·51n 90·32w
Cananea.....30·57n 110·18w
Celaya.....20·31n 100·37w
Cerritos.....22·26n 100·17w
Champotón.....19·21n 90·43w
Chihuahua.....28·38n 106·05w
Chilpancingo [de los Bravos]....17·33n 99·30w
Ciudad Acuña..29·18n 100·55w
Ciudad Allende.28·20n 100·51w
Ciudad Camargo.....27·40n 105·10w
Ciudad Chetumal.....18·30n 88·18w
Ciudad del Carmen.....18·38n 91·50w
Ciudad de Valles.....21·59n 99·01w
Ciudad de Villaldama.....26·30n 100·26w
Ciudad Guerrero.....28·33n 107·30w
Ciudad Guzmán.....19·41n 103·29w
Ciudad Ixtepec..16·34n 95·06w
Ciudad Jiménez.....27·08n 104·55w
Ciudad Juárez..31·44n 106·29w
Ciudad Mante (Mante).....22·44n 98·57w
Ciudad Obregón.....27·29n 109·56w
Ciudad Victoria.23·44n 99·08w
Coalcomán.....18·47n 103·09w
Coatzacoalcos..18·09n 94·25w
Colima.....19·14n 103·43w
Comitán.....16·15n 92·08w
Concepción del Oro.....24·38n 101·25w
Córdoba.....18·53n 96·56w
Coyuca de Catalán.....18·20n 100·39w
Cuencamé [de Ceniceros].....24·53n 103·42w
Cuernavaca....18·55n 99·15w
Culiacán.....24·48n 107·24w
Delicias.....28·13n 105·28w
Durango.....24·02n 104·40w
Ebano.....22·13n 98·22w
El Salto.....23·47n 105·22w
Empalme.....27·58n 110·51w
Ensenada.....31·52n 116·37w
Escuinapa [de Hidalgo].....22·51n 105·48w
Felipe Carrillo Puerto.....19·35n 88·03w
Fresnillo.....23·10n 102·53w
Gómez Palacio..25·34n 103·30w
Guadalajara....20·40n 103·20w
Guanajuato.....21·01n 101·15w
Guasave.....25·34n 108·27w
Guaymas.....27·56n 110·54w
Hermosillo.....29·04n 110·58w
Heroica Nogales.....31·20n 110·56w
Hidalgo.....24·15n 99·26w
Hidalgo del Parral.....26·56n 105·40w
Hopelchén.....19·46n 89·51w
Huajuapan de León.....17·48n 97·46w
Huatabampo....26·50n 109·38w
Huetamo de Núñez.....18·35n 100·53w
Huixtla.....15·09n 92·28w
Hunucmá.....21·01n 89·52w
Iguala.....18·21n 99·32w
Irapuato.....20·41n 101·28w
Ixmiquilpan.....20·29n 99·14w
Jalapa Enríquez.19·32n 96·55w
Jamiltepec.....16·17n 97·49w
Jesús Carranza.17·26n 95·02w
Jonuta.....18·05n 92·08w
Juchitán [de Zaragoza].....16·26n 95·01w
La Barca.....20·17n 102·34w
Lagos de Moreno.....21·21n 101·55w
La Libertad.....29·55n 112·43w
Lampazos.....27·01n 100·31w
La Paz.....24·10n 110·18w
León [de Los Aldamas].....21·07n 101·40w
Linares.....24·52n 99·34w
Loreto.....26·01n 111·21w
Los Mochis.....25·45n 108·57w
Magdalena.....30·38n 110·57w
Mante, see Ciudad Mante
Manzanillo.....19·03n 104·20w
Matamoros.....25·53n 97·30w
Matehuala.....23·39n 100·39w
Mazatlán.....23·13n 106·25w
Mérida.....20·58n 89·37w
Mexicali.....32·40n 115·29w
Mexico City.....19·24n 99·09w
Miahuatlán de Porfirio Díaz...16·20n 96·36w
Michoacán.....17·55n 93·53w
Minatitlán.....17·59n 94·31w
Mocorito.....25·29n 107·55w
Moctezuma.....29·48n 109·42w
Monclova.....26·54n 101·25w
Montemorelos...25·12n 99·49w
Monterrey.....25·40n 100·19w
Morelia.....19·42n 101·07w
Nacozari [de García].....30·24n 109·39w
Naranjos.....21·21n 97·41w
Nautla.....20·13n 96·47w
Navojoa.....27·06n 109·26w
Nazas.....25·14n 104·08w
Nogales, see Heroica Nogales
Nueva Casas Grandes.....30·25n 107·55w
Nueva Rosita...27·57n 101·13w
Nuevo Laredo...27·30n 99·31w
Oaxaca [de Juárez].....17·03n 96·43w
Obregón.....28·47n 106·52w
Ocotlán.....20·21n 102·46w
Ocozingo.....16·54n 92·07w
Ojinaga.....29·34n 104·25w
Ometepec.....16·41n 98·25w
Orizaba.....18·51n 97·06w
Pachuca [de Soto].....20·07n 98·44w
Pánuco.....22·03n 98·10w
Papantla [de Olarte].....20·27n 97·19w
Paraíso.....18·24n 93·14w
Parras.....25·25n 102·11w
Pátzcuaro.....19·31n 101·36w
Petatlán.....17·31n 101·16w
Peto.....20·08n 88·55w
Piedras Negras.28·42n 100·31w
Pijijiapan.....15·42n 93·14w
Pochutla.....15·44n 96·28w
Pozo Rica.....20·30n 97·27w
Progreso.....21·17n 89·40w
Puebla [de Zaragoza].....19·03n 98·12w
Puerto Ángel...15·40n 96·29w
Puerto Juárez...21·11n 86·49w
Puerto Peñasco.31·20n 113·33w
Puerto Vallarta..20·37n 105·15w
Querétaro.....20·36n 100·23w
Quilá.....24·23n 107·13w
Reynosa.....26·07n 98·18w
Río Grande.....23·50n 103·02w
Ríoverde.....21·56n 99·59w
Sabinas.....27·51n 101·07w
Sabinas Hidalgo.....26·30n 100·10w
Sahuaripa.....29·03n 109·14w
Salamanca.....20·34n 101·12w
Salina Cruz.....16·10n 95·12w
Salinas de Hidalgo.....22·38n 101·43w
Saltillo.....25·25n 101·00w
San Andrés Tuxtla.....18·27n 95·13w
San Blas.....21·31n 105·16w
San Cristóbal de las Casas.....16·45n 92·38w
San Felipe.....31·00n 114·52w
San Fernando...24·50n 98·10w
San Francisco del Oro.....26·52n 105·51w
San Francisco del Rincón.....21·01n 101·51w
San Ignacio....27·27n 112·51w
San José del Cabo.....23·03n 109·41w
San Luis Potosí.22·09n 100·59w
San Luis Río Colorado.....32·29n 114·48w
San Pedro de las Colonias...25·45n 102·59w
Santa Ana.....30·33n 111·07w
Santa Bárbara..26·48n 105·49w
Santa Rosalía...27·19n 112·17w
Santiago Papasquiaro...25·03n 105·25w
Saucillo.....28·01n 105·17w
Sayula.....19·52n 103·37w
Soledad Díez Gutiérrez.....22·12n 100·57w
Tamazunchale..21·16n 98·47w
Tamiahua.....21·16n 97·27w
Tampico.....22·13n 97·51w
Tapachula.....14·54n 92·17w
Taxco de Alarcón. .....18·33n 99·36w
Teapa.....17·33n 92·57w
Tecpan de Galeana.....17·15n 100·41w
Tehuacán.....18·27n 97·23w
Tehuantepec...16·20n 95·14w
Temax.....21·09n 88·56w
Tenosique de Pino Suárez...17·29n 91·26w
Tepatitlán [de Morelos].....20·49n 102·44w
Tepic.....21·30n 104·54w
Teziutlán.....19·49n 97·21w
Tijuana.....32·32n 117·01w
Tizimín.....21·10n 88·10w
Tlaxcala [de Xicoténcatl]....19·19n 98·14w
Tlaxiaco.....17·16n 97·41w
Toluca [de Lerdo].....19·17n 99·40w
Tonalá.....16·04n 93·45w
Topolobampo...25·36n 109·03w
Torreón.....25·33n 103·26w
Tula.....23·00n 99·43w
Tulancingo.....20·05n 98·22w
Tuxpan.....21·57n 105·18w
Tuxpan de Rodríguez Cano.....20·57n 97·24w
Tuxtla Gutiérrez.....16·45n 93·07w
Ures.....29·26n 110·24w
Uruapan [del Progreso].....19·25n 102·04w
Valladolid.....20·41n 88·12w
Valparaíso.....22·46n 103·34w
Veracruz [Llave].....19·12n 96·08w
Villa Ahumada..30·37n 106·31w
Villahermosa....17·59n 92·55w
Villa Unión.....23·12n 106·14w
Zacatecas.....22·47n 102·35w
Zamora de Hidalgo.....19·59n 102·16w
Zihuatanejo.....17·38n 101·33w

Physical features and points of interest

Ángel de la Guarda, Isla, *island*.....29·20n 113·25w
Arena de la Ventana, Punta, *point*.....24·04n 109·52w
Baja California, *peninsula*.....27·30n 113·00w
Balsas, *river*....17·55n 102·10w
Bravo del Norte (Rio Grande), *river*.....25·57n 97·09w
California, Golfo de, *gulf*.....28·00n 112·00w
Campeche, Bahía de, *bay*..20·00n 94·00w
Carmen, Isla, *island*.....25·55n 111·10w
Catoche, Cabo, *cape*.....21·36n 87·07w
Cedros, Isla, *island*.....28·10n 115·15w
Central Plateau, see Mexicana, Altiplanicie
Cerralvo, Isla, *island*.....24·17n 109·52w
Chapala, Lago de, *lake*.....20·15n 103·00w
Chinchorro, Banco, *bank*...18·35n 87·20w
Citlaltépetl, Volcán, *volcano*.....19·01n 97·16w
Colnett, Cabo, *cape*.....31·00n 116·20w
Colorado, *river*..31·45n 114·40w
Conchos, *river*..29·32n 104·25w
Corrientes, Cabo, *cape*....20·25n 105·42w
Cozumel, Isla de, *island*.....20·25n 86·55w
Cuitzeo, Lago de, *lake*.....19·55n 101·05w
Espíritu Santo, Isla del, *island*.24·30n 110·20w
Eugenia, Punta, *point*.....27·50n 115·03w
Fuerte, *river*....25·50n 109·25w
Grande, Rio, see Bravo del Norte
Grijalva, *river*...18·36n 92·39w
Infiernillo, Presa del, *reservoir*...18·35n 101·45w
Iztaccíhuatl, *mountain*.....19·11n 98·39w
Juárez, Valle de, *valley*.....31·40n 106·15w
La Encantada, Cerro de, *mountain*.....31·00n 115·24w
Madre, Laguna, *lagoon*.....25·00n 97·40w
Madre de Chiapas, Sierra, *mountains*.....16·20n 93·50w
Madre de Oaxaca, Sierra, *mountains*.....17·45n 96·40w
Madre Occidental, Sierra, *mountains*.....25·00n 105·00w
Madre Oriental, Sierra, *mountains*.....22·00n 99·30w
Madre del Sur, Sierra, *mountains*.....17·00n 100·00w
Malpaso, Presa de, *reservoir*...17·10n 93·40w
Mayo, *river*.....26·45n 109·47w
Mexicana, Altiplanicie, *plateau*.....24·00n 102·00w
Mexican Plateau see Mexicana, Altiplanicie
Mexico, Gulf of..25·00n 90·00w
Miguel Alemán, Presa, *reservoir*.....18·13n 96·32w
Mita, Punta de, *point*.....20·47n 105·33w
Mujeres, Isla, *island*.....21·13n 86·43w
Neo-Volcánica, Cordillera, *mountain range*.....19·30n 100·00w
Orizaba, Pico de, see Citlaltepétl, Volcán
Pacific Ocean...20·00n 110·00w
Papaloapan, *river*.....18·42n 95·38w
Parícutin, *volcano*.....19·28n 102·15w
Popocatépetl, Volcán, *volcano*.....19·02n 98·38w
Puebla, Sierra de, *mountains*..19·56n 97·45w
Revillagigedo, Islas de, *islands*.....19·00n 111·30w
Roca Partida, Isla, *island*.....19·01n 112·02w
Salado, *river*....18·44n 103·36w
San Benedicto, Isla, *island*.....19·18n 110·49w
San José, Isla, *island*.....25·00n 110·38w
San Lázaro, Cabo, *cape*.....24·50n 112·18w
San Lucas, Cabo, *cape*.....22·50n 109·55w
San Pablo, Punta, *point*...27·12n 114·30w
San Pedro Mártir, Sierra, *mountains*.....30·45n 115·13w
San Quintín, Cabo, *cape*.....30·21n 116·00w
Santa Magdalena, Isla, *island*.....24·50n 112·15w
Santa Margarita, Isla de, *island*..24·25n 111·50w
Sebastián Vizcaíno, Bahía, *bay*.....28·00n 114·30w
Sinaloa, *river*....25·18n 108·30w
Soconusco, Sierra de, *mountains*.....15·30n 92·35w
Socorro, Isla, *island*.....18·45n 110·58w
Sonora, *river*....28·50n 111·33w
Tamiahua, Laguna de, *lagoon*.....21·35n 97·35w
Tehuantepec, Golfo de, *gulf*..16·00n 94·50w


MEXICO
Size of symbol indicates relative size of town
Elevations in metres
0 100 200 300 400 km
0 100 200 300 mi
© Rand McNally & Co.
A-531600-257
UNITED STATES
MEXICO
GULF OF MEXICO
PACIFIC OCEAN
TEXAS
NEW MEXICO
ARIZONA
LOUISIANA
ARKANSAS
MISSISSIPPI
ALABAMA
FLORIDA
Golfo de California
Tropic of Cancer
SIERRA MADRE OCCIDENTAL
SIERRA MADRE ORIENTAL
SIERRA MADRE DEL SUR
ALTIPLANICIE MEXICANA
CORDILLERA NEO-VOLCANICA
PENINSULA DE YUCATAN
Bahía de Campeche
Golfo de Tehuantepec
Río Grande
Río Bravo del Norte
Río Colorado
Río Mississippi
MEXICO CITY
Guadalajara
Monterrey
Puebla
Tijuana
Mexicali
Ensenada
San Diego
Tucson
El Paso
Ciudad Juárez
Dallas
Houston
San Antonio
Austin
New Orleans
Baton Rouge
Jackson
Montgomery
Hermosillo
Chihuahua
Ciudad Obregón
Mazatlán
Culiacán
Durango
Torreón
Saltillo
Nuevo Laredo
Laredo
Reynosa
Matamoros
Brownsville
Piedras Negras
Ciudad Acuña
Tampico
Ciudad Victoria
Veracruz
Jalapa Enriquez
Aguascalientes
San Luis Potosí
León
Irapuato
Querétaro
Pachuca
Morelia
Toluca
Acapulco
Oaxaca
Tuxtla Gutiérrez
Villahermosa
Coatzacoalcos
Orizaba
Campeche
Mérida
Tapachula
Colima
Tepic
Zacatecas
Durango
Culiacán
Mazatlán
Hermosillo
Obregón
Delicias
Ciudad Camargo
Hidalgo del Parral
Ciudad Jiménez
Monclova
Ciudad de Villaldama
Nuevo Laredo
Sabinas Hidalgo
Ciudad Frontera
Ciudad Acuña
Nueva Rosita
Ciudad Allende
Ciudad Guerrero
Nuevo Casas Grandes
Villa Ahumada
Ciudad Juárez
Agua Prieta
Cananea
Nogales
Puerto Peñasco
Caborca
Santa Ana
Magdalena
Nadozari
Moctezuma
Buenaventura
Cinaga
Guaymas
Empalme
Navojoa
Huatabampo
Los Mochis
Topolobampo
Sinaloa
Mocorito
Culiacán
Santiago Papasquiaro
Nazas
Gómez Palacio
San Pedro de las Colonias
Parras
Torreón
Cuencamé
Concepción del Oro
Saltillo
Monterrey
Montemorelos
Linares
San Fernando
Laguna Madre
Matamoros
Reynosa
Brownsville
Hidalgo
Río Grande
Matehuala
Ciudad Victoria
Tula
Ciudad Mante
Ebano
Tampico
Panuco
Ciudad de Valles
Naranjos
Tamazunchale
Tuxpan de Rodriquez Cano
Poza Rica
Papantla
Nautla
Tulancingo
Pachuca
Iztaccihuatl 5286
Teziutlán
Orizaba
Tlaxcala
Puebla 5700
Córdoba
Veracruz
Alvarado
Jalapa Enriquez
Volcán Popocatepetl 5452
Amecameca
Cuernavaca
Toluca
Uruapan
Morelia
Patzcuaro
Pariacutin 2808
Zamora de Hidalgo
Sayula
Autlán de Navarro
Ciudad Guzmán
Manzanillo
Colima
Tepic
San Blas
Zacatecas
Aguascalientes
San Luis Potosí
Soledad Diez Gutierrez
Río Verde
Cerritos
Salinas de Hidalgo
Fresnillo
Valparaíso
Tuxpan
Acaponeta
Escuinapa
Mazatlán
Villa Unión
El Salto
Durango
Cuencamé
Concepción del Oro
Mapimí
Laguna de Tamiahua
Tamiahua
Tuxpan de Rodriquez Cano
Poza Rica
Papantla
Nautla
Tulancingo
Pachuca
Ixmiquilpan
Querétaro
Irapuato
Celaya
León
Guanajuato
San Francisco del Rincón
Lagos de Moreno
Tepatitlán
Ameca
Ocotlán
Salamanca
La Barca
Lago de Chapala
Lago de Cuitzeo
Acámbaro
Morelia
Zamora de Hidalgo
Sayula
Autlán de Navarro
Ciudad Guzmán
Manzanillo
Colima
Tepic
San Blas
Guadalajara
Puerto Vallarta
Cabo Corrientes
Punta de Mita
Islas Tres Marías
Islas de Revillagigedo
Isla San Benedicto
Isla Roca Partida
Isla Socorro
Cabo San Lucas
San José del Cabo
Punta Arena de la Ventana
La Paz
Isla del Espíritu Santo
Isla Cerralvo
Isla de Santa Margarita
Isla Santa Magdalena
Cabo San Lázaro
Isla San José
Isla Carmen
Loreto
Isla Angel de la Guarda
Isla Tiburón
Bahía Sebastián Vizcaíno
Punta Eugenia
Punta San Pablo
Isla Cedros
Cabo Colnett
Cabo San Quintín
Ensenada
Tijuana
Mexicali
San Diego
Yuma
San Luis Río Colorado
Puerto Peñasco
San Felipe
Cerro de la Encantada 3069
Sierra San Pedro Mártir
Golfo de California
Sierra Madre Occidental
Sierra Madre Oriental
Sierra Madre del Sur
Altiplanicie Mexicana
Cordillera Neo-Volcánica
Península de Yucatán
Bahía de Campeche
Golfo de Tehuantepec
Río Grande
Río Bravo del Norte
Río Colorado
Río Mississippi
Valle de Juárez
Río Conchos
Río Fuerte
Río Yaqui
Río Mayo
Río Sinaloa
Río Lerma
Río Balsas
Río Papaloapan
Río Usumacinta
Río Grijalva
Río Suchiate
Presa del Infiernillo
Presa Miguel Alemán
Presa de Malpaso
Lago de Chapala
Lago de Cuitzeo
Laguna de Tamiahua
Laguna Madre
Golfo de California
Golfo de Tehuantepec
Bahía de Campeche
Mexico City
Guadalajara
Monterrey
Puebla
Tijuana
Mexicali
Ensenada
San Diego
Tucson
El Paso
Ciudad Juárez
Dallas
Houston
San Antonio
Austin
New Orleans
Baton Rouge
Jackson
Montgomery
Hermosillo
Chihuahua
Ciudad Obregón
Mazatlán
Culiacán
Durango
Torreón
Saltillo
Nuevo Laredo
Laredo
Reynosa
Matamoros
Brownsville
Piedras Negras
Tampico
Ciudad Victoria
Veracruz
Jalapa Enriquez
Aguascalientes
San Luis Potosí
León
Irapuato
Querétaro
Pachuca
Morelia
Toluca
Acapulco
Oaxaca
Tuxtla Gutiérrez
Villahermosa
Coatzacoalcos
Orizaba
Campeche
Mérida
Tapachula
Colima
Tepic
Zacatecas
110°
105°
100°
95°
90°
85°
30°
25°
20°
15°

MAP INDEX (continued)

Tehuantepec, Istmo de, *isthmus*.....17·00n 94·30w
Términos, Laguna de, *lagoon*.....18·35n 91·30w
Texcoco, Lago de, *lake*.....19·30n 90·00w
Tiburón, Isla, *island*.....29·00n 112·20w
Tres Marías, Islas, *islands*...21·25n 106·28w
Usumacinta, *river*.....18·22n 92·40w
Yaqui, *river*.....27·37n 110·39w
Yucatan Channel.....21·45n 85·45w
Yucatan Peninsula.....19·30n 89·00w
Zumpango, Lago de, *lake*.....19·46n 99·09w

Mexico coastal plain and extends inland to the foothills of the Sierra Madre Oriental. It is mainly composed of plains, hills, and the exterior slope of the Veracruz, Tabasco, and northern Oaxacan and Pueblan ranges, which have rainy tropical climates and dense forests or savannas of great potential wealth. Mestizos and small groups of Indians live in the rich alluvial valleys of the Papaloapan and the Grijalva-Usamacinta river systems, on the plantation lands of medium altitudes in Veracruz and Puebla states, and in the oil-field and manufacturing districts of Minatitlán, Coatzacoalcos, Ciudad Pemex, Orizaba-Córdoba, and Veracruz.

Except for its northern section, Chiapas state is relatively isolated from the rest of the country. Its physical regions include the tropical coastal area on the Pacific Ocean, the central Grijalva Depression, the almost uninhabited Lacandonia jungles, and the temperate and cold regions of the Sierra Madre ranges and the Mexican Plateau. In the Chiapas Highlands, around San Cristobál de Las Casas, there is a large concentration of Indians—the Tzeltales and Tzotziles—who are engaged in animal husbandry.

The southeast, which consists entirely of the Yucatán Peninsula, is a region containing the tropical forests of the territory of Quintana Roo and Campeche state in the south and communal farms (*ejidos*) and agave plantations in the north. Mérida, where agave is processed to yield the henequin fibre, is located in the northwest. The region is populated by Mayan peasants, the descendants of a people who, centuries ago, created a great civilization.

The landscape under human settlement. Mexico continues to be a predominantly agricultural nation, which at the same time is experiencing a rapid, and almost totally disproportionate, growth in its cities. The urbanization process, which began in about 1940, has changed the complexion of the country, although in 1970 about 50 percent of the population continued to live in localities of fewer than 10,000 inhabitants.

The origins of rural settlement

Rural life. Because the Spanish colonial economy was based primarily on mining and only secondarily on agriculture, western and central Mexico became the most densely populated regions. In the more arid areas, settlement took the form of missions, mining posts, and small towns. The Indians were divested of their lands in the valleys and were relegated to the mountains or the forests, where they settled in small poverty-ridden villages. During the presidency of Porfirio Díaz from 1877 to 1911, large haciendas, or ranches, based on the production of grain, were established in central Mexico, while cattle ranches were established in the north. The peons, or peasants, settled in the vicinity of the haciendas. After 1910, important changes were effected in regional population patterns. The haciendas were divided; large areas were irrigated in La Laguna, Bajo Bravo, and in the northwest, creating high rural population densities on what had previously been almost uninhabited plains.

The rural population is also dense in the oil-producing regions of the east, on the tropical plantations of Yucatán and other coastal areas, and in the traditionally Indian areas of the southern valleys. In general, rural communities are widely scattered. There are more than 90,000 settlements, the majority of which are ranches or hamlets with less than 500 inhabitants.

In irrigated regions where agriculture is mechanized and where productivity is high, rural life offers opportunities for the improvement of the peasants' standard of living. In contrast to this, in the areas of traditional peasant agriculture—the rainy areas of central Mexico and the western mountains, the tropical southern regions, and Yucatán—the dietary and housing conditions of the peasants, especially of the Indians, reflect their poverty, isolation, and abandonment.

Urban regions. The urban population is concentrated in the central regions, which extend across the country between the 19th and 22nd parallels north latitude. Large industrial and commercial centres have also developed in the north (Monterrey, Chihuahua, and Torreón), in the strip along the United States frontier (Tijuana, Juárez, and Nuevo Laredo), and in the irrigated valleys of the northwest (Hermosillo), where commercial agriculture is practiced. The oil centres of Tampico, Poza Rica, Minatitlán, and Coatzacoalcos; the large Yucatán metropolis of Mérida; and the tourist and commercial centre of Acapulco have developed independently. Since the 1950s, the great urban agglomerations of Mexico City, Guadalajara, Monterrey, León, and Puebla have accelerated their growth, as have other cities located in the central and lowland regions, in the northeast, and in Veracruz state.

Many cities in the northwest and in the border zone—among them Mexicali, Nogales, Obregón, Juárez, Piedras Negras, Tijuana, and Los Mochis—have developed since the 1950s and are new in appearance. This is also true of certain industrial centres, such as Sahagún, Poza Rica, Coatzacoalcos, Reinosa, and Monclova, whose growth has been stimulated by heavy government investment.

In contrast to the new cities, the colonial towns and cities contrive to preserve their original central plaza, from which radiates a network of streets. In the early 1970s these older towns and cities experienced enormous difficulties in expanding. Many of them, such as Guanajuato, Zacatecas, Pachuca, and Taxco, are old mining cities in a state of decay, which have retained only their cultural, commercial, or administrative importance.

The large cities

By 1970 the greater Mexico City area contained 18 percent of the total national population. It was the largest industrial centre, producing about 45 percent of the total value of gross national production in 1965 and was also the largest consumption centre in the country. An enormous, cosmopolitan city, Mexico City is the true economic, political, and cultural heart of the republic. Guadalajara in the west, with its 1,500,000 inhabitants, is a flourishing state capital experiencing rapid industrialization. Monterrey (population 1,200,000) is a steel centre and the financial capital of the north.

In the large cities, skyscrapers and deluxe housing developments are found beside poor and sordid slums. Rapid expansion and industrial concentration have caused serious problems of water and power supply, as well as traffic, housing, and cost-of-living problems; there is, moreover, a diminishing amount of "green space" for gardens and parks. There is also a problem of pollution, increased in Mexico City by the lack of rivers and by the partial or total drying up of old lakes, such as Texcoco, Zumpango, Xaltocan, Chalco, and Xochimilco.

PEOPLE AND POPULATION

Groups historically associated with contemporary Mexico. Spanish is the official national language and is spoken either as a first or second language by almost the entire population. Mexican Spanish is greatly influenced by the country's indigenous languages, which are spoken by about 8 percent of the population. About 1,000,000 people speak only their indigenous tongues. The highest concentrations of Indian language speakers are found in the states of Oaxaca, Puebla, Yucatán, Veracruz, Chiapas, México, Hidalgo, and Guerrero. The main Indian language families are Uto-Aztecan, Mayan, Hokaltecan, and Oto-Manguean.

Mestizos compose about 55 percent of the population. About 29 percent of the population is Indian and 15 percent is of European ancestry; of the latter, two-thirds are native-born and the remainder are aliens from Spain, the United States, China, France, and Guatemala.

About 96 percent of the population is Roman Catholic.

The constitution of 1917 established the separation of church and state, and the church lands are controlled by the government. The Protestant population numbers about 879,000 and there are about 49,000 Jews.

Mexico, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census
Zones (*zonas*)				
Centro				
Federal district				
Distrito Federal	579	1,499	4,871,000	6,874,000
States				
Aguascalientes	2,158	5,589	243,000	338,000
Guanajuato	11,810	30,589	1,735,000	2,270,000
Hidalgo	8,103	20,987	995,000	1,194,000
Jalisco	30,941	80,137	2,443,000	3,297,000
México	8,286	21,461	1,898,000	3,833,000
Michoacán	23,114	59,864	1,852,000	2,324,000
Morelos	1,908	4,941	386,000	616,000
Puebla	13,096	33,919	1,974,000	2,508,000
Querétaro	4,544	11,769	355,000	486,000
Tlaxcala	1,511	3,914	347,000	421,000
Golfo de México				
States				
Campeche	21,666	56,114	168,000	252,000
Tabasco	9,522	24,661	496,000	768,000
Veracruz	28,114	72,815	2,728,000	3,815,000
Yucatán	16,749	43,379	614,000	758,000
Territory				
Quintana Roo	16,228	42,030	50,000	88,000
Norte				
States				
Chihuahua	95,400	247,087	1,227,000	1,613,000
Coahuila	58,522	151,571	908,000	1,115,000
Durango	46,196	119,648	761,000	939,000
Nuevo León	24,925	64,555	1,079,000	1,695,000
San Luis Potosí	24,266	62,848	1,048,000	1,282,000
Tamaulipas	30,822	79,829	1,024,000	1,457,000
Zacatecas	28,973	75,040	818,000	951,000
Pacifico Norte				
States				
Baja California	27,071	70,113	520,000	870,000
Nayarit	10,664	27,621	390,000	544,000
Sinaloa	22,429	58,092	838,000	1,267,000
Sonora	71,403	184,934	783,000	1,099,000
Territory				
Baja California, Sur de	28,447	73,677	82,000	128,000
Pacifico Sur				
States				
Chiapas	28,528	73,887	1,211,000	1,569,000
Colima	2,106	5,455	164,000	241,000
Guerrero	24,631	63,794	1,187,000	1,597,000
Oaxaca	36,820	95,364	1,727,000	2,172,000
Total Mexico	759,530*	1,967,183	34,923,000†‡	48,382,000
	761,600§	1,972,547§		

*Converted area figures do not add to total given because of rounding.
†*De jure.* ‡Figures do not add to total given because of rounding.
§Includes 2,071 sq mi (5,364 sq km) of uninhabited offshore islands.
Source: Official government figures.

Demography. *Fertility and mortality rates.* Throughout the 20th century, Mexico's fertility rate has been one of the highest in the world. It has maintained itself at a constant level—approximately 45 births per 1,000 population. Yet fairly important changes in fertility rates among women of different age groups have occurred. The fertility rate of young women has dropped, whereas that for women over 35 has increased. The fertility of women living in urban centres is less than that of women who live in rural areas. There are, however, no indications of a decrease in the general fertility rates in the cities, despite the fact that the rate is decreasing among young women.

The rate of growth has produced a rejuvenation of the population. The proportion of persons under 14 rose from 41 percent in 1930 to 46 percent in 1970; this was the result of the maintenance of high fertility rates and of a greater decrease in the mortality rates of infants and young people than in those for adults and the aged.

The mortality rate

There has been a notable decrease in the mortality rate since the 1930s. From a gross mortality rate of 27 deaths per 1,000 population in 1930, the rate dropped to 9.9 per 1,000 in 1970. Average life expectancy correspondingly rose from approximately 37 years in 1930 to 61 years for men and 64 years for women in 1965–70. The drop in the mortality rate was due partly to better economic and social conditions in general but especially to advances in medicine.

Infant mortality declined about 69 percent between 1930 and 1965. Deaths resulting from infectious and parasitic diseases have been reduced, although these causes of death are still the most important. To an increasing extent, the lengthening of life will be subject to the control of "degenerative" diseases.

Migration. International immigration is relatively unimportant, as less than one-half of 1 percent of the population in 1970 was foreign-born. Of the 191,000 persons not born in Mexico, 35 percent were concentrated in the Federal District and 38 percent in the states of Baja California, Chihuahua, Jalisco, Nuevo León, and Tamaulipas—*i.e.*, in the states closest to the United States.

Most Mexicans who emigrate go to the United States. Between 1900 and 1950 approximately 1,000,000 Mexicans migrated north. From 1960 to 1970 an estimated 30,000 to 40,000 emigrants entered the United States.

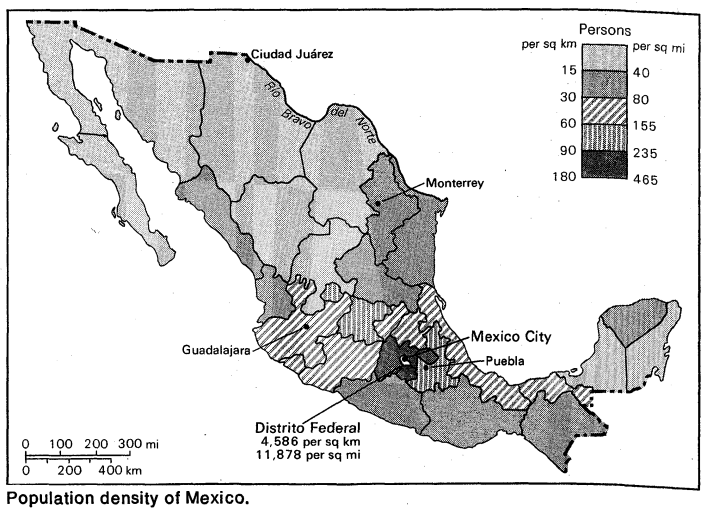
Internal migration between 1940 and 1970 was marked by three major characteristics—an increase in the number of persons residing in a state different from that of their birth; a heavy concentration of the migrant population in specific areas; and—from 1960 to 1970—a stabilization of the proportion of migrants to the total population. For every 1,000 inhabitants, 105 were residing outside their state of birth in 1930, 128 in 1950, 149 in 1960, and 145 in 1970. In 1970, there were, altogether, 7,000,000 persons living outside their state of origin.

In 1970, 74 percent of the migrant population was concentrated in the Federal District and in the states of Baja California, Tamaulipas, Veracruz, México, Nuevo León, Chihuahua, and Jalisco. The metropolitan area of Mexico City alone absorbed over 47 percent of the migrants. The migrants are predominantly young persons; among them, the proportion of women to men is slightly greater.

The attraction of the cities

Population distribution. The accelerated growth rate in the urban centres is a result of both migration and of the birth rate among the population. The average annual rate of population increase in communities of fewer than 2,500 persons is less than 2 percent annually, while in communities of 2,500 or more persons, the rate of increase had risen to more than 5 percent by 1970.

In 1970, more than 42 percent of the population lived in communities of less than 2,500 persons, 23 percent in those of 2,500 to 20,000; almost 10 percent in those of 20,000 to 75,000; 14 percent in those of 75,000 to 500,000; and 11 percent in cities of 500,000 persons or more.



Population density of Mexico.

Demographic trends. A decrease in fertility rates is foreseen in response to changing fertility patterns, higher levels of education, raised income levels, and an increase in the age of matrimony. Estimates have been made of a birth rate of 41 births per 1,000 population by 1980, with a decline to 33 per 1,000 by 2000. Estimates indicate a

life expectancy at birth of approximately 72 years of age by 2000. This would mean a population of approximately 135,000,000 in the year 2000. (P.G.C.)

THE NATIONAL ECONOMY

Nationalization of the economy

Mexico's most remarkable economic achievement has been the sharp reduction in the degree of foreign control exercised over key sectors of the economy, while maintaining overall national growth. The state has assumed control of telegraphic and electrical generation and distribution systems and has reduced the foreign interest in mining and in important manufacturing fields. State initiative has also resulted in the establishment of a motor-vehicle manufacturing industry based on local components. At the same time Mexico has recorded the most sustained pace of economic growth in Latin America; the gross national product (GNP) at 1950 prices more than doubled between 1950 and 1965, and in the period between 1965 and 1970 the real increase in GNP averaged 7 percent a year.

Such rapid development inevitably put a strain on the balance of payments. The heavy intake of foreign capital meant that the amount spent on servicing the public debt has been proportionately higher than in any other Latin American country except Argentina. Massive imports of raw materials and machinery, in spite of extensive import substitution, more than doubled the trade deficit between 1960 and 1970. The peso, however, has retained its value on world markets since 1954 and is used by the International Monetary Fund as one of the world's reserve currencies. Moreover, in spite of its economic nationalism, close ties with the United States still enable Mexico to attract public and private investment from abroad. The trading dependence of Mexico on the United States has, in fact, hardly declined, for the gradual diversification of export clients has been overshadowed by the massive annual influx of United States tourists.

Mexico has a wide variety of natural resources and has an enormous power potential in its petroleum, hydroelectric, and uranium reserves. It supplies almost one-fifth of the world's silver and is the world's largest producer of fluorite (the principal source of fluorine, used as a flux in metallurgy); it is the third largest producer of bismuth (a brittle metallic element used in the manufacture of fusible alloys and in medicine); the third largest of sulfur; the fourth largest of graphite; and the fifth largest of mercury. It is also an important producer of gold, lead, copper, zinc, manganese, cadmium (a white ductile metallic element resembling tin), and antimony (a brittle, lustrous, white metallic element used chiefly in alloys and in medicine). Its coal reserves, together with those of Colombia, are the only deposits in Latin America directly convertible into coke. Combined with its large iron deposits, Mexico has built up the second biggest steel industry in Latin America (after Brazil). The country has the largest petroleum and basic petrochemical industry in Latin America, as well as the highest glass output. It provides half the world's supply of hormones and is the world's largest exporter of natural honey.

The extent and distribution of resources. *Mineral resources.* The state of Chihuahua is responsible for approximately half the nation's lead and zinc production, about one-third of its gold, and most of its uranium. Sonora has dominated the copper industry because of its deposits at Cananea and La Caridad. In the early 1970s, known iron ore deposits were estimated at about 600,000,000 tons or more; they included the Las Truchas deposits on the Michoacán and Guerrero borders. Coal reserves are estimated at 2,200,000,000 tons in the state of Coahuila.

Silver and gold are largely produced as a by-product of lead, zinc, and copper. Other metals to be found include manganese, molybdenum (a metallic element used as an alloy with iron in making high-speed cutting tools), tungsten (a relatively uncommon metallic element used in some alloys), antimony, tin, bismuth, and cadmium. Sulfur, the major nonmetallic ore, is located in the Isthmus of Tehuantepec, although a small amount of sulfur of volcanic origin is obtainable in the state of San Luis Potosí. Fluorite is a major mining product, as is barite (the principal ore of the metallic element barium).

The petroleum industry is oriented toward supplying domestic fuel needs. About 90 percent of industry's energy needs are supplied by petroleum. Exploration and drilling are geared to guaranteeing supplies for 20 years in advance; the main hopes of exceeding this limit come from the offshore fields in the Gulf of Mexico off Tampico, where reserves are estimated to exceed 10,000,000,000 barrels. Two-thirds of the nation's crude oil comes from the state of Veracruz and the rest from Tabasco, which also produces approximately 45 percent of the country's natural gas.

Biological resources. Tropical hardwoods are an important biological resource. They include mahogany, cedar, logwood, primavera (white mahogany, used for making furniture), ceiba (kapok), and fustic (a tree of the mulberry family that yields a yellow dye). Hardwoods are exploited chiefly in the southeastern region of the Isthmus of Tehuantepec and on the Yucatán Peninsula. The sapodilla tree yields chicle, the chief ingredient of chewing gum.

The main fish reserves are found in the Gulf of California and the Gulf of Mexico. They include shrimp, oysters, tuna, sardines, and red snapper.

Hydroelectric resources. Water resources are unevenly distributed, but the Sierra Madre ranges give rise to sizeable rivers. Before the nationalization of the electrical industry in 1960, thermal (fuel-burning) plants supplied 57 percent of the total capacity. Since then, there has been a major shift to developing the country's hydroelectric potential. This has led to the construction of the Infiernillo project on the Michoacán–Guerrero border and of the Malpaso project in the state of Tabasco. The latter is the largest single source of power in Mexico. Power for the northern border is derived from geothermal sources, which tap live steam from beneath the Earth's crust. A geothermic plant is in operation near Mexicali.

Sources of national income. *Agriculture, forestry, and fishing.* The dependence of national growth on the agricultural sector has been considerably reduced since World War II. In 1970 agriculture, forestry, and fishing together provided only about 12 percent of the gross domestic product. The agricultural sector is the major source of employment and involves approximately 40 percent of the active population. The growth of agricultural output has been slow but remarkable in the context of Central and South America; livestock rearing, in particular, has shown steady growth.

The most dispersed agriculture is that of the scattered Indian communities, which maintain only minimal contact with the money economy. Nevertheless, they provide an important quantity of maize and in some areas are self-sufficient in many basic commodities.

Agricultural cooperatives

The ejidos are communal farms, created after 1910, that form the basis of the land-redistribution program; these communities have supplied labour to the more organized agricultural sectors.

Plantation agriculture produces the main export commodities, such as cotton, coffee, sugar, and henequen. Cattle rearing is also practiced on a large scale.

Because of fluctuations in world-commodity prices, the government has taken a large stake in the sugar industry and has assumed control of henequen processing. The production of coffee has grown steadily, but cotton has fallen drastically since the late 1960s. Tobacco production has held firm. The export of fruit and vegetables to the United States was becoming significant in the early 1970s.

Fishing cooperatives dominate the fishing industry, especially on the west coast. Mexico is the most important supplier of prawns (shrimplike crustaceans) to the United States market. Forestry is affected by ejido rights, which have limited expansion of the commercial forested area to 12,000,000 acres. The main product is hardwood timber.

Mining. Mining output has increased since the mid-1960s because of special incentives, the extension of ma-

jority ownership rule to all mining companies, and the upturn in world markets. Coal and iron ore cannot be exported, but mining still makes an important impact on exports. The rapid growth of manufacturing diverts approximately 60 percent of production away from the export markets, and local needs influence government policy.

Manufacturing. The manufacturing of chemicals, especially of basic petrochemicals, was the fastest growing industry in the early 1970s. The iron and steel industry is dominated by three basic producers of ingots and rolled products, and by one producer of tubing; the government has an interest in all but one of these companies. There is one basic aluminum producer, and a strong mechanical engineering sector has grown up around the automobile industry.

The glass industry is well developed. There are almost 50 basic producers of paper and about 20 producers of pulp, but considerable quantities of pulp and newsprint have to be imported. The textile industry has traditionally been a major source of manufacturing employment. In spite of some progress in the wool and cotton industry, locally produced man-made fibres are gaining in importance. A substantial industry in consumer durable goods (such as cars, radios, etc.) has grown up; production of television sets, for instance, amounted to more than 400,000 in 1969.

Public works accounted for half the construction activity, which represented more than 4 percent of the gross domestic product in 1970. The country was self-sufficient in cement in the early 1970s.

Industry is concentrated in the Mexico City area and to a lesser extent in the Monterrey and Guadalajara areas. Satellite industrial zones have been established around these centres; the largest of these are in Puebla, Querétaro, Cuernavaca, and Toluca.

Energy. The state petroleum monopoly, Petroleos Mexicanos (Pemex), is the largest Latin-American oil company. Productivity, however, is not high, as it employs only slightly more workers than the Italian oil company, Ente Nazionale Idrocarburi (ENI). No crude oil is exported.

The installed capacity of the electricity network run by the Comisión Federal de Electricidad (Federal Commission of Electricity) was 7,495,000 kilowatts in 1970. Electricity is also produced privately by industrial and mining concerns.

Financial services. The dual banking structure consists of government financial institutions and a number of private merchant, deposit, and mortgage banks—all closely linked with industrial and commercial groups. The stock exchange plays only a minor role in the provision of new capital, as most of the activity is in government bonds or banking securities.

The Banco de México (Bank of Mexico) follows the general monetary and financial policy laid down by the secretary of the treasury. It engages in open-market operations, sells gold to the public, fixes reserve requirements, enforces credit controls, and has the sole right of note issue. There are about 30 other state financial institutions, the most important of which, the Nacional Financiera (the National Merchant Bank), is one of the main channels of foreign funds into infrastructure projects such as road construction and water supply.

The Nacional Financiera has a direct stake in a number of industrial concerns, as does the smaller holding company, the Sociedad Mexicana de Crédito Industrial (Mexican Society of Industrial Credit). The other government banking institutions each specialize in one sphere of activity—such as providing credit to cooperatives, rural communities, public works, or foreign trade.

The private banking system is composed of deposit, merchant, and mortgage banks. The merchant banks overtook the deposit banks in the size of their combined resources in 1965 and have had the competitive edge since then. The secretary of the treasury controls prime interest rates, but both the cost and availability of credit vary according to the relationship of the lender to the borrower, since financial power is very concentrated. Three merchant banks control up to 40 percent of the lending power of this type of institution, and another seven merchant banks control about 33 percent. The leading merchant banks are linked with the major deposit banks, where resources are also concentrated—two banks control over one-half of the total of deposited bank funds and four others control one-third. The result of this concentration has been to restrict the volume and increase the cost of credit to companies outside these groups, particularly to companies involved in agriculture, low-cost housing, and similar nonmanufacturing activities. Investment in some sectors, such as agriculture, is therefore dependent on government institutions.

Foreign trade. Industrialization has reduced the need for imported consumer goods, but the need for capital equipment and production goods has led to a rapid rate of import growth. Exports, however, increased only slightly from 1965 to 1970, with the result that the trade gap has widened. The impact of the trade deficit has been moderated by the increase in tourism and expenditure by United States citizens in the northern border zone, which produces a gross foreign-exchange income greater than that earned by merchandise exports. At the same time, however, there has been a rapid rate of growth in travel abroad by Mexicans, as well as in profits sent abroad as a result of increased foreign investment.

The balance of trade

The main destinations for Mexico's major exports of cotton, sugar, coffee, and sulfur are the United States, Japan, Switzerland, and West Germany. The chief import sources are the United States, West Germany, Japan, the United Kingdom, and France. Major imports include machinery, motor vehicles, electrical equipment, chemical products, and iron and steel.

Management of the economy. *The private sector and the role of government.* The state controls electricity generation and distribution; petroleum production, refining, and distribution; the telegraphic and railway systems; and one of the two international airlines. The government also has exclusive rights to operate in basic petrochemicals and dominates the fertilizer market. It controls the Nacional Financiera and the Sociedad Mexicana de Crédito Industrial, as well as the major producers of steel, tubing, copper, iron ore, and motor vehicles.

The state affects the operation of private industry by various means, both formal and informal, including the system of import licenses, the production quotas and requirements for use of local materials applied to some industries, and the control of the retail prices of some consumer products. The private sector is also legally debarred from certain activities reserved for the state, including the manufacturing of basic petrochemicals.

An industrial incentive scheme operates through tax incentives and other concessions, and a special industrialization program is also in operation along the northern boundary with the United States. There is no national development plan, although a public-investment program and a plan for agricultural development were under consideration in the early 1970s.

The government also defines the spheres of activity of the foreign investor. Foreign capital is barred completely from a number of industries, including insurance, petroleum, and forestry, and is permitted only on a minority basis in activities such as mining, transport, radio and television broadcasting, fishing, advertising and publicity, soft drinks, rubber products, fertilizers, iron and steel, cement, glass, pulp, aluminum, and banking. Moreover, no incentives are granted to companies with a majority of foreign ownership; the government also has discretionary power to introduce minimum local ownership requirements.

Taxation. Income taxes provide about 47 percent of the government's revenue. Others include import and export duties and taxes on profits, industry, trade, the exploitation of natural resources, and public services.

Trade unions and employer associations. The right to strike is guaranteed, but only strikes that follow prescribed procedures are protected by law. Industrial peace is largely maintained through the close relations of the unions with the ruling Partido Revolucionario Institucion-

al (PRI). The Confederación de Trabajadores Mexicanos (Confederation of Mexican Workers), with an estimated membership of 2,000,000 and a wide net of affiliated regional unions, is the most important union.

Approximately half of the nonagricultural labour force, including civil servants, belongs to a union. Unions have generally succeeded in establishing closed shops in most major industries; dismissals are resisted by unions and are penalized by high severance payments. The national minimum wage level is adjusted every two years. The 1970 labour law laid emphasis on fringe benefits and on programs requiring that workers participate in the profits. Workers not covered by collective contracts are guaranteed the national minimum wage, but the system is widely disregarded in rural districts where it is difficult to enforce.

Industrial boards

Every individual employer or company is required to join an industrial board; these are grouped under either the Confederación de Camáras Industriales (Confederation of Industrial Boards) or the Confederación de Camáras Nacionales de Comercio (National Confederation of Industrial Boards). These and the private association of employers, the Confederación Patronal (Employers' Confederation), provide the focus for lobbying and public relations activities.

Contemporary economic policies. Economic policy is aimed at the achievement of economic growth and the dissemination of economic benefits to the public. Priority has been given to agricultural development and the elimination of the pockets of poverty existing in the countryside. Social security and public-health programs are to be improved and an attempt made to reduce the housing shortage. New industrial investment is to be directed outside the central area. This policy requires the strengthening of budgetary resources and the external-payments position. Taxes on luxury items have been increased, and the tax structure is oriented toward the provision of investment incentives, especially for the production of exports. Foreign trade is to be promoted by bodies that will coordinate and finance exports and smooth out the problems at ports. Development of tourism will continue, in order to moderate the outflow of capital resulting from the high trade deficit and foreign investment.

Problems and prospects. About one-third of the population have a substandard income and live in one-room dwellings. Most of the wealth is concentrated in the Mexico City area or in the major provincial capitals. With a high rate of population growth, prosperity cannot depend on traditional activities such as agriculture. Industrial growth, however, has passed the first stage of import substitution, so that growth now must come from the scale of operations made possible by the size of the home market. This, in turn, may lead to a greater export capability.

Inflation has not been a problem, but the close links of the financial and trading system with the United States makes the system vulnerable to any downturn in the economic fortunes of its northern neighbour. The most critical short term problem is likely to be the deficit in the balance of payments. Lack of controls on capital repatriation and on the sending of remittances abroad will continue to be required so as to attract foreign private capital investment, but curbs may be imposed on imports and on travel abroad by Mexican residents. (E.I.U.)

TRANSPORTATION

Overall patterns of transport. In 1965, over 90 percent of Mexico's passenger traffic was carried by automobiles, while the railways transported almost 50 percent of the total freight tonnage. The central and northern zones, where economic activity and population are most concentrated, have the most developed transport network.

The three transportation regions

There are three major transportation regions. In the northeast, Saltillo, Monterrey, and Ciudad Victoria are the principal convergence points for railway, highway, and air transport. Connections are made to the United States by way of Piedras Negras, Nuevo Laredo, Reinosa, and Matamoros or via Tampico on the Gulf of Mexico; numerous routes lead to the rest of the country. In the north central region, the main focusses of transportation are Chihuahua and Torreón. These are highway and rail connections to the northern border by way of Ojinaga and Ciudad Juárez and to the northwest by means of the Chihuahua-Pacific Railroad, which ends at the port of Topolobampo. In the northwest, the road and rail centres are Guaymas, Hermosillo, and Nogales. The Baja California Peninsula is linked with the coastal states of Sonora and Sinaloa by means of highways, railways, and trans-gulf ferries. A highway running down the entire length of the peninsula was under construction in the early 1970s.

In the south and southeast, the main focuses of transportation are the cities of Tapachula, Oaxaca, Acapulco, and Salina Cruz in the Pacific coastal area, and Veracruz, Villahermosa, and Mérida in the area of the Gulf of Mexico. Connections are made to Guatemala by way of El Talismán, Ciudad Hidalgo, and Ciudad Cuauhtémoc in Chiapas.

Component systems. *The highway network.* In 1970, highways totalled more than 46,000 miles. Tollways and federal highways—which extend throughout the country—represent almost half the total road mileage and form the trunk lines of the system. The remainder of the system consists of a growing network of rural and state roads; these promote the development of new regions and feed the principal highway and railway lines. The 1,250-mile (2,000-kilometre) Pan-American Highway runs from Ciudad Cuauhtémoc, on the border with Guatemala, to Ciudad Juárez, on the border with the United States, passing through Mexico City.

The railways. There are 15,000 miles of railway lines. The system was constructed in the late 19th century and rehabilitated after 1910. Since the 1950s, new railway construction has taken place in order to help service national markets. In addition to the lines that lead to the United States border and those that connect the central areas with the Gulf and Pacific ports, railways connecting the central plateau with Yucatán and Baja California have been completed. In 1961 the Chihuahua-Pacific line, between Ojinaga on the northern frontier and Topolobampo on the Gulf of California, was opened.

Maritime transport. Approximately 50 percent of all foreign commerce is carried by sea. There are 14 ports on the Gulf of Mexico and 16 on the Pacific. Major ports include Tampico, Veracruz, Salina Cruz, Acapulco, Ensenada, Guaymas, Manzanillo, and Coatzacoalcos. The merchant marine is relatively small, and port installations remained insufficient in the early 1970s. There is a considerable amount of coastal shipping.

Air services. In 1921 the first concession was granted for regular passenger, mail, and air-freight services between Mexico City and other Mexican cities and ports. By the early 1970s there were over 60 airports, not including small runways and rural airfields. Of these airports, 23 were equipped for international service. Scheduled flights connect Mexico with cities in North, Central, and South America, Europe, and Asia.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. *Constitutional framework.* The 1917 constitution of the United Mexican States (Estados Unidos Mexicanos) declares that Mexico is a representative, democratic, and federal republic. Executive, legislative, and judicial powers are divided, and provision is made for the direct election of the president and the legislators. In addition to guarantees of personal and civil liberties, the constitution establishes principles of an economic and political nature.

Constitutional guarantees

Supreme executive authority is vested in the president, who is elected to serve one six-year term. His power of appointment extends to the members of his cabinet and the attorney general, and to the territorial and federal district governors, diplomatic officials, the higher officers of the army, and ministers of the higher courts of justice. As the leader of the dominant political party, the president can effect constitutional amendments that then require approval by a two-thirds vote of a quorum in Congress and the support of a majority of the state legislatures. The chief executive may also establish basic rules

or *reglamentos*, which give effect to general statutes and have the force of law. The president has extensive control over foreign affairs. (P.G.C.)

A bicameral Congress is composed of the Chamber of Deputies and the Senate. Two senators from each state and from the Federal District are elected for six-year terms, concurrent with that of the president. The 178 deputies are elected every three years, one deputy representing 200,000 people. Each state has at least two deputies and each territory at least one. Congressmen may not be re-elected for the immediately succeeding term. Each house is the final judge of the election of its own members, an arrangement that has permitted the majority party to declare its own candidates elected, regardless of ballot results.

A permanent committee of 15 deputies and 14 senators serves during the congressional recess and approves certain presidential requests, the most important of which may be suspension of guarantees in times of emergency.

The legislative branch has declined in importance as compared with the executive. Annual changes in committee membership, low pay, infrequent meetings, and short sessions have discouraged the full development of congressional powers. The fact that the majority party leader in each house outranks the congressional president of each house (who is selected at monthly intervals) further restricts congressional autonomy. Potentially, however, the Chamber of Deputies could exercise great power over the executive through its functions of verifying the presidential electoral returns, passing the budget, and initiating tax bills. (M.B.T.)

Local and municipal government. Mexico is politically divided into 29 states, two territories, and one federal district, which is the seat of the federal government. The territorial division and political and administrative organization of the states are based on the concept of *municipio libre* (literally "free municipality). Each state is an autonomous unit governed by a constitution modelled on the national document. State governors are elected by popular vote every six years; deputies of the unicameral state legislatures are elected every three years and senators every four years. Each state may levy taxes and legislate under specified conditions; all powers not delegated to the federal government are reserved for the states.

The federal union of states

The federal district is governed by a *jefe* (chief), or *regente* (regent), who is appointed by the president of the country. The two territories of Quintana Roo in the Yucatán and Sur de Baja California are headed by governors, also appointed by the president. An amendment was approved in 1971 for the reinstitution of *municipio libre* in these entities.

The political process. The history of 20th-century Mexico has been eventful and rich in de facto and legal political movements. With the formation in 1929 of the Partido de la Revolución Mexicana (National Revolutionary Party), now the Partido Revolucionario Institucional (PRI, or Institutional Revolutionary Party), politics became more strongly and more significantly institutionalized.

All Mexicans over the age of 18 have the right to vote. The dominant feature of elections is the strength of the PRI. Since its foundation as the Partido Nacional Revolucionario, it has not lost a presidential, gubernatorial, or senatorial election. Presidential elections, which are the model for elections on other levels, have reaffirmed the dominant character of the party in power and have minimized the opposition's possibilities. Extreme violence has not occurred since the murder of President-elect Alvaro Obregón in 1928.

The PRI is the expression of the interests of the country's middle classes, with respect to its nominees; its managing cadres; its support of a policy of state intervention in economic, political, and cultural life; and its reconciliation of different social sectors.

In the early 1970s there were three other legally constituted national parties. The Partido Acción Nacional (PAN, National Action Party) is the most successful opposition party. It presents an image as the defender of popular causes but adopts an eminently individualistic policy toward matters of education and property. The Partido Popular Socialista (PPS, Socialist People's Party) is only minimally important; it is preoccupied with the demands of labour and has a Marxist orientation. The military is supported by the elitist Partido Auténtico de la Revolución Mexicana (PARM, Authentic Party of the Mexican Revolution).

Justice. The administration of justice, at the federal level, is the responsibility of the Supreme Court of Justice and of the circuit and district courts. In each state or federal unit, there is a Superior Court of Justice, as well as civil, criminal, and mixed peace courts (courts with civil and minor criminal jurisdiction).

The attorney offices of the Ministerio Público (Public Ministry) prosecute criminal cases in their respective jurisdictions: there is an attorney general for the republic and one attorney for each state and territory and for the federal district.

The federal constitution protects the accused and guarantees free trial and due process. No one may be imprisoned for a debt of a purely civil nature, and cruel or unusual punishments are forbidden. No criminal trial may go through more than two levels of appeal. Laws or acts of authority that violate constitutionally guaranteed rights of the individual may be contested before the federal courts by means of the Juicio de Amparo (Protection Proceeding); Amparo is also used in ruling on conflicting state and federal statutes.

The Juicio de Amparo

In its relations with the executive branch of government, the federal judiciary is independent and has the power to restrain the president. The Supreme Court, however, largely follows the policies of the executive and, in effect, serves to give it greater stability.

The armed forces. The Mexican Army does not play a preponderant role in the political life of the country. Since the 1920s, measures have been taken to achieve civil control of the military, through the agency of the PRI and the power of the president, who appoints the higher military officers.

The regular armed services are composed of almost 68,000 members, of whom 54,000 are in the army, 7,600 in the navy, 6,000 in the air force, and 900 in the marines. The navy is equipped with frigates, minesweepers and other craft, and the air force has jet-trainer fighter bombers among its aircraft. Given the absence of external conflicts, the army's function consists exclusively of maintaining internal order. Conscripts—18-year-olds who serve a compulsory year in the military—and the Rural Defense Corps (consisting of some 80,000 cavalry and 40,000 infantry) are also considered integral parts of the army.

Administration. *Educational services.* Education is free and compulsory. It provides for two years of pre-primary and six years of primary schooling and for three years of secondary training. On the primary level, three-fourths of the schools are public and are located in rural areas; the remainder are located in urban areas. Most of the rural establishments, however, provide only three, instead of six, years of primary instruction. The private primary schools are generally directed by religious orders.

Post-primary education is provided by professional, normal (teacher-training), secondary, preparatory, and subprofessional schools.

There are more than 35 universities. The two large institutions of higher education, both in Mexico City, are the Universidad Nacional Autónoma de México (Autonomous National University of Mexico) and the Instituto Politécnico Nacional (National Polytechnic Institute).

Health and welfare services. Social aid is furnished through three main institutions—the Secretariat of Health and Aid, the Mexican Institute of Social Insurance (IMSS), and the Institute of Social Services and Security of the Workers of the State (ISSSTE). The secretariat provides for more than 2,700 medical centres. The IMSS provides medical services, workmen's compensation, and retirement benefits to labourers and employees in the private and mixed-enterprise sector; it obtains funds through proportional assessments on their salaries.

Governmental social agencies

Most of the beneficiaries form part of the urban population and include the insured, the pensioned, and their families. The ISSSTE performs the same functions as the IMSS but provides benefits only to the bureaucracy.

Housing. The problem of housing is one of the most serious facing the country, particularly in the cities. In addition to a chronic housing shortage, a large proportion of the existing units are substandard or have insufficient water supplies, poor drainage, or other defects. By 1964 new housing was being built at a rate of 12,000 units per year. In the early 1970s, official policy was to construct large, multifamily units, which were to be sold as condominiums; it was estimated that it would be necessary to build some 300,000 housing units annually to satisfy the existing demand.

Police services. There are two federal police forces—the Judicial Police of the Federal District and Territories and the Federal Judicial Police.

The local police of Mexico City consist of grenadiers (riot police), mounted, transport, motorized and traffic police.

Social conditions. *Wages and cost of living.* The range of wages is extremely wide; it runs the gamut from high salaries earned by senior government officials and managers of private companies at one end of the scale, down to the precarious subsistence levels of the peons in the rural areas, at the other. The wages of workers and employees depend upon their qualifications and length of service, but there are appreciable differences according to the type and size of firm.

The minimum wage

Since 1934, there has been a law requiring owners to pay a minimum wage that is fixed every two years in accordance with the cost of living of each region. This law is not, however, fully observed. Workers are legally empowered to ask for biennial raises, but the increase in cost of living continues to parallel the raises granted.

Health conditions. Health conditions have markedly improved because of investment both in social assistance services and health services.

The greatest incidence of illness is caused by diseases of intestinal origin, either parasitic or infectious. As a result of bad diet, there is a high incidence of such ailments as rickets, vitamin deficiency, and anemia (deficiency of red blood cells). There is also a high incidence of diseases that are difficult to control, such as tonsillitis, influenza, and respiratory diseases.

Economic and social divisions. In the early 1970s, 40 percent of the economically active population were employed in agriculture, 23 percent in industry, and 32 percent in services.

There are three major agricultural groups—the *ejidatarios,* or communal farmers, the proprietors, and the agricultural workers. The *ejidatarios* are poor and are on about the same economic level as the least of the proprietors, those who own the *minifundios.* The small and medium-sized landowners form the rural middle class, whereas the large landowners form an agricultural elite. Possessing about 84 percent of all privately owned land, this elite receives close to half of the net national agricultural income and has managed to elude reforms aiming at redistribution of land. The rural proletariat is made up of 3,500,000 landless peasants who labour for landowners and communal farmers. They are employed during sowing and harvesting periods for an extremely low wage.

Blue-collar and white-collar workers form the bulk of the urban middle class. These two categories are subdivided by levels of income that are proportional to training and education. The urban higher income group is comprised of public officials as well as affluent groups of Mexican and foreign origin.

CULTURAL LIFE AND INSTITUTIONS

The cultural environment. A tendency toward cosmopolitanism, which paradoxically blends with a tendency toward a nationalistic culture, is perhaps the most characteristic aspect of Mexico's cultural environment. The old obsession for remaining in contact with "universal" culture persists, yet at the same time interest in Mexico's own cultural heritage continues.

The state of the arts. Culturally, the Mexican is concerned with his role in the wider, international community. In the field of philosophy, this concern was initiated by Samuel Ramos' speculations on man and culture in Mexico. Literature no longer deals with the revolution. Although the cares and sufferings of the Mexican peasant continue to be the subject of such writers as Juan Rulfo, there is now a universal literature with characters and themes that are universally applicable. Contemporary literature ranges from the fantasies of Juan José Arreola to the novels of Carlos Fuentes, portraying Mexican cosmopolites, to the poetry of Octavio Paz.

In music, Carlos Chávez, without renouncing the heritage of the past, has transformed it to create a music belonging to no specific country—a trend which has been amplified by those composers who have followed the Chávez tradition. Something similar has also taken place in the contemporary theatre, through the work of such figures as Rodolfo Usigli, who has taken material for his dramas from Mexican history.

Painting

The Mexican Muralist School of Painting, which sought its inspiration in the Combative period of the Mexican Revolution, has been followed by contemporary artists who have sought to express themes that transcend the national. Rufino Tamayo, who crystallized this trend and demonstrated its possibilities, was followed by Carlos Mérida and Juan Soriano. Painters such as José Luis Cuevas have sought only to express themselves, regardless of whether their artistic work might or might not be viewed as meeting any particular standard or as belonging to any particular school. The national art is no longer simply folkloric or popular, although popular art may be the point of departure for an apparently abstract art, which is also truly national.

Cultural institutions. The Instituto Nacional de Bellas Artes (National Fine Arts Institute) is devoted to the stimulation of creativity in the fine arts and to the dissemination of artistic work. Theatre, literature, dance, the plastic arts, and music are all objects of its attention. The Orquesta Sinfónica de México (National Symphony Orchestra) has been given notable government support, as have various ballet groups such as the Ballet Folklórico de México (Folkloric Ballet) and the Ballet Moderno y Clásico (Modern and Classical Ballet).

National University carries on extensive artistic activities, both within the University and extramurally. Theatre, dance, films, music, the humanities, and the sciences receive great stimulus in faculties, schools, centres and institutes. Original, classical, and contemporary works are presented to audiences through the use of the media and through university facilities.

Folklore and popular culture are promoted under the auspices of the Secretariat of Public Education. The Instituto Nacional de Antropología e Historia (Institute of Anthropology and History) includes the Museo de Antropología (Museum of Anthropology), the Museo de las Culturas (Cultural Museum), the Museo de Artes Populares (Museum of Popular Art), and the Instituto Nacional de Bellas Artes (National Fine Arts Institute). The Instituto Indigenista (Native Institute) encourages native crafts.

Other important cultural institutions include the Colegio de México (the College of Mexico), the Colegio Nacional (the National College), and the Seminario de Cultura (Cultural Seminar); there are also academies of the arts, languages, and history.

Press and broadcasting. A variety of cultural publications is issued by the Secretariat of Public Education. The National University publishes through its General Directorate the works of its professors, researchers, and artists. Among publishing firms, the Fondo de Cultura Económica (Foundation of Economic Culture) and the Editorial Siglo XXI (21st Century Publishers) produce works that reach every Spanish-speaking country.

Daily newspapers

In the late 1960s, there were more than 230 daily newspapers, of which 18 were located in Mexico City, the rest being spread throughout the country. They operate under the constitutional guarantee of freedom of the press, but conservative social pressures and a desire to make Mexi-

co attractive to tourists have resulted in low-key criticism of the government. The Mexico City morning daily *Excelsior* is often considered as the nation's best newspaper; moderately conservative, it is circulated throughout Mexico and has a wide impact on public opinion. Other influential dailies include *El Universal, Novedades, Diario de la Tarde,* and the English-language *News.* The morning daily *La Prensa* is the leading liberal newspaper; *El Nacional* supports government policies.

Radio plays an important role in linking rural areas to the cities. In the late 1960s there were almost 400 standard-band radio stations in operation throughout the country, as well as shortwave and FM transmitters. Two large radio networks owned about 130 stations, but most provincial stations were independent. Most broadcasts are in Spanish, with some in English, French, Italian, and German.

Mexico City acquired the first licensed television station in Latin America in 1950. In the early 1970s there were three commercial television stations in the capital; colour transmission was introduced in 1967. Latin America's first full-time educational television station began transmission in 1959. Television is also broadcast from 18 provincial cities and is received by a total of 38 channels.

PROSPECTS FOR THE FUTURE

In the early 1970s—more than 60 years after the revolution of 1910—Mexico had achieved many of the Revolution's objectives. Political advances included the ban on re-election of the governors and the president. Labour's position had improved markedly through constitutional provisions, and agricultural reform had been instituted. The nationalization of natural resources promised continued economic progress, and the secularization of public education reinforced a pattern of continuing social change.

Mexico finds itself, however, confronting the same problems as other nations engaged in industrialization. Economic problems include the necessities of establishing competitive prices, broadening domestic and foreign markets, and of lowering interest rates on loans. In the social area, the dissolution of family ties and the high crime rate present grave difficulties. To these problems must be added the serious demographic explosion, which will probably cause the population to double in 20 years. Mexico is also plagued by the lack of adequate technology, the restiveness of a politically aware youth, and particularly by the difficult situation in which the agrarian population finds itself placed.

BIBLIOGRAPHY

General works: DANIEL COSIO VILLEGAS, *Ensayos y notas* (1966), is a collection of essays on Mexican problems written by a notable historian. An important discussion of Mexico's industrialization is presented in SANFORD A. MOSK, *Industrial Revolution in Mexico* (1950). JORGE L. TAMAYO, *Geografía general de México* (1949), is an extensive geographic work; and FRANK TANNENBAUM, *Mexico: The Struggle for Peace and Bread* (1950), is important for its valuable observations.

Specific works: Natural resources and physical geography are discussed in ANGEL BASSOLS BATALLA, *Recursos naturales: Climas, agua, suelos,* 2nd ed. (1969); and P.E. JAMES, *Latin America,* 4th ed. (1969). H.G. GIERLOFF-EMDEN, *Mexico: Eine Landeskunde* (1970), is an excellent work on the geography of Mexico. MAXIMILIEN SORRE, *Le Méxique* (1928), is a classic work of great importance. For demographic information, see RAUL BENITEZ ZENTENO *et al., Dinámica de la población de México* (1970); ERIC WOLF, *Sons of the Shaking Earth* (1959); R. DE ZAYAS ENRIQUEZ, *Los Estados Unidos Mexicanos, sus condiciones naturales y sus elementos de prosperidad* (1893); and ALEXANDER VON HUMBOLDT, *Ensayo político sobre el reino de la Nueva-España,* 4 vol. (1822). Administration and social conditions are discussed in PABLO GONZALEZ CASANOVA, *La democracia en México* (1965); E.R. RUIZ, *Mexico: The Challenge of Poverty and Illiteracy* (1963); and JORGE ALBERTO LAZAYA, *El Ejército mexicano (1911–1965)* (1970). Cultural life and institutions are presented in JUSTINO FERNANDEZ, *Arte mexicano* (1958); JOSE MORENO VILLA, *Lo Mexicano en las artes plásticas* (1948); LEOPOLDO ZEA, *América como conciencia* (1953); and FONDO DE CULTURA ECONOMICA, MEXICO, *América en la historia* (1957).

(P.G.C.)

Mexico, Gulf of

The Gulf of Mexico is a partially landlocked body of water on the southeastern periphery of the North American continent. It is connected to the Atlantic Ocean by the Straits of Florida, running between the peninsula of Florida and the island of Cuba, and to the Caribbean Sea by the Yucatán Channel, which runs between the Yucatán Peninsula and Cuba. Both these channels are about 100 miles wide. It covers an area of more than 500,000 square miles (1,300,000 square kilometres). To the northwest, north, and northeast the Gulf is bounded by the southern coast of the United States, while to the west, south, and southeast it is bounded by the east coast of Mexico.

Area and boundaries

The various rivers flowing into the Gulf drain a total area of about 1,250,000 square miles. The Gulf, with a climate that varies from the tropical to the subtropical, is traversed by currents that become the Gulf Stream, entering through the Yucatán Channel and flowing in a clockwise direction to exit through the Straits of Florida. From June to October the Gulf is liable to become a spawning ground for hurricanes.

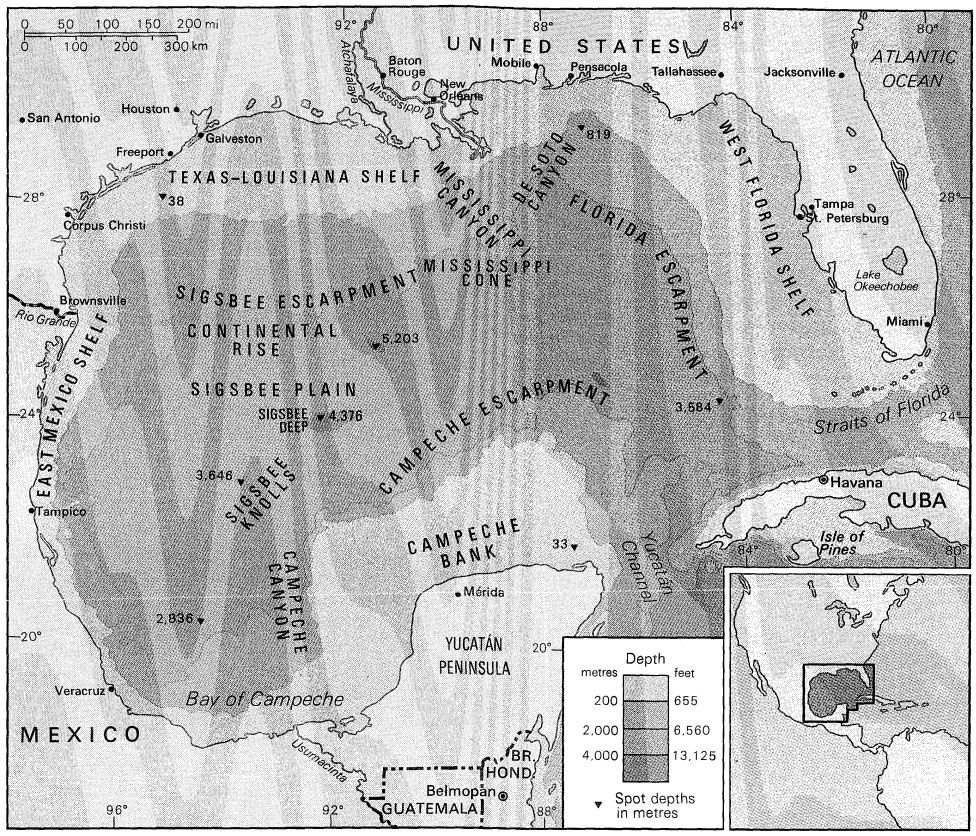
The Gulf has considerable economic resources, which include oil and gas, sulfur, and large quantities of fish; its coastal waters are also much used for recreation. To an increasing extent, however, the Gulf is becoming a receptacle for waste of all kinds, including pesticides and other pollutants, issuing particularly from the Mississippi. (For associated physical features, see CARIBBEAN SEA; MISSISSIPPI RIVER; and GULF STREAM.)

Physiography and geological history. The Gulf consists of several ecological and geological provinces, chief of which are the coastal zone, the continental shelf, the continental slope, and the abyssal plain. The coastal zone consists of tidal marshes, sandy beaches, mangrove-covered areas, and many bays, estuaries, and lagoons. The continental shelf forms an almost continuous terrace around the margin of the gulf; its width varies from a maximum of more than 200 miles (320 kilometres) to a minimum of about 25 miles (40 kilometres). Off the west coast of Florida as well as off the Yucatán Peninsula, the continental shelf consists of a broad area composed primarily of carbonate material (formed of salt or other chemical compounds). The remainder of the shelf consists of sand, silt, and clay sediments. On the shelf and on the slope that dips downward to the abyssal plain, buried salt domes occur at various depths; important oil and gas deposits are associated with them. The abyssal plain, which forms the floor of the Gulf, consists of a large triangular area near the centre, bounded by abrupt fault scarps toward Florida and Yucatán and by more gentle slopes to the north and west. The basin is unusually flat, having a gradient of only about one foot in every 8,000 feet. The deepest point is Sigsbee Deep, which is 17,070 feet (5,203 metres) below sea level. From the floor of the basin rise the Sigsbee Knolls, some of which attain heights of 1,300 feet; these are surface expressions of the buried salt domes.

Only since about 1950 have comprehensive and systematic oceanographic studies of the Gulf been begun. Much remains to be done before most of the scientific questions about it can be answered. There is, for example, considerable scientific speculation as to the geologic origin of the Gulf. Some contend it is a foundered continental crust that has been flooded by the ocean; others believe it to be an ocean basin that has been subjected to rifting. Proponents of the theory that the continents were once joined but drifted apart contend that, like the Mediterranean, the Gulf is an ancient sea that was in existence even when the various continental masses of the Earth formed a single landmass. Unfortunately, the scientific drilling of the sea floor of the Gulf during recent expeditions was not deep enough to obtain the definitive evidence needed to decide the question.

Oceanographic studies of area

Hydrology. *Currents.* The Gulf Stream is the principal current moving oceanic waters into the Gulf. It enters through the Yucatán Channel, the floor of which forms a sill at about 5,000 to 6,000 feet beneath the surface, and flows out via the Straits of Florida. Meandering masses of



The Gulf of Mexico.

water, known as loop currents, break off from the mainstream and move clockwise into the northeastern part of the Gulf. Both seasonal and annual variations occur in these loop currents. A less well-defined pattern exists in the western Gulf. Here, the currents are relatively weak, varying appreciably in intensity with season and location. There is extreme variability in both current direction and speed on the continental shelf and in the coastal waters of the Gulf, where currents are subjected not only to seasonal and annual variations caused by major circulation patterns but also by changes in the prevailing wind direction.

Salinity and temperature. The salinity of the Gulf is also subject to wide variations. In the open Gulf the salinity is comparable to that of the Atlantic Ocean, with about 36 parts of salt for every 1,000 parts of seawater by weight. This proportion varies markedly during the year, however, in coastal waters, particularly near the outflow of the broad delta region of the Mississippi. During periods when the volume of the Mississippi's flow is greatest, salinities as low as 14 to 20 parts of salt per 1,000 of water occur as far as 20 to 30 miles offshore.

Sea-surface temperatures in February vary between 64° F (18° C) in the northern Gulf and 76° F (24° C) off the Yucatán coast. In the summer, surface temperatures of about 90° F (32° C) have been measured, but the usual variation is between 65° to 75° F (18° to 24° C). Bottom-water temperatures of about 43° F (6° C) have been recorded near the northern part of the Yucatán Channel. The thickness of the isothermal layer (a surface layer of water of constant temperature) varies from a few feet to more than 500 feet, depending on seasonal and local conditions as well as on location. The tidal range is small, averaging less than two feet in most places; in general, only one high water and one low water occur during the lunar day.

Climate. The climate varies from tropical to subtropical. The hurricane season officially starts June 1 and ends November 1. Meteorological and oceanographic conditions are conducive during most of this period to the formation of hurricanes anywhere in the Gulf. Hurricanes spawned in the Caribbean and South Atlantic may also move through the Gulf at this time.

Fish and wildlife. Red snappers, flounder, shrimps, mullet, oysters, and crabs provide the basis for commercial fishing in the Gulf; tarpon and king mackerel are caught for sport. The shores of the Gulf are a major wintering area for waterfowl and whooping cranes, and substantial colonies of seabirds are located on the offshore islands. There is a marked absence of marine mammals.

Resources and their exploitation. Man makes extensive use of the Gulf as a major source of food, energy, and raw materials; he also uses it for recreation and, to an ever increasing extent, for waste disposal. The Gulf yields a rich harvest of fish and shellfish, which are sold for human consumption. In addition, a large tonnage of the fish caught are used to provide fish protein concentrate for raising cattle and poultry. About 1,400,000,000 pounds of fin fish were landed in 1970, and about 230,000,000 pounds of shrimp. The total value of these landings was about $167,000,000.

The Gulf is an important source of oil and gas, which is obtained from wells drilled on the continental shelf in water depths of as much as 500 feet and at distances of up to almost 100 miles offshore from Louisiana. Recent annual production of oil has exceeded 1,000,000 barrels per day. Oil reserves on the continental shelf alone are estimated to amount to about 2,400,000,000 barrels and gas reserves to about 34,000,000,000,000 cubic feet (960,000,000,000 cubic metres). Sulfur is also extracted from wells drilled on the continental shelf off Louisiana, with recent annual production amounting to several million tons. As much as 21,000,000 tons of oyster shells are obtained annually from the shallow waters of the Texas Gulf Coastal Plain, as well as from bays and estuaries. These are used in the chemical industry as a source of calcium carbonate and also provide material for building roads.

The coastal waters of the Gulf are used extensively for sport fishing, boating, sailing, swimming, and scuba diving. The estimated annual economic value of the recreational activities is more than $1,000,000,000.

Future prospects. The Gulf of Mexico is used with increasing frequency as a means of disposing of wastes—

Pollution of the waters

both directly by the discharge along the coast of sewage and industrial wastes and indirectly by the dissemination of pollutants carried by river waters draining into the Gulf. The Gulf thus becomes the final resting place of tremendous amounts of chemicals from pesticides, herbicides, and fertilizers used in agriculture in much of the United States. The introduction into the Gulf of large amounts of phosphates occasionally leads to the occurrence of red tides, so called partly because of the mass mortality caused among fish at various places along the shore from the tip of Florida to southwestern Texas and partly because of the large masses of red algae (Rhodophyta) that the phosphates cause to flourish. The true capacity of the Gulf as a receptacle for the wastes of a major portion of the United States remains to be determined, but it is certainly not infinite.

BIBLIOGRAPHY. More detailed information may be found in J.W. ANTOINE and J.C. GILMORE, "Geology of the Gulf of Mexico," *Ocean Ind.*, 5:34–38 (1970); P.S. GALTSOFF (ed.), *Gulf of Mexico: Its Origin, Waters, and Marine Life* (1954); R.A. GEYER, "A Bibliography on the Gulf of Mexico," *Texas. J. Sci.*, 2:44–93 (1950), and "The Occurrence of Pronounced Salinity Variations in Louisiana Coastal Waters," *J. Mar. Res.*, 9:100–110 (1950); D.F. LEIPPER, "A Sequence of Current Patterns in the Gulf of Mexico," *J. Geophys. Res.*, 75:637:657 (1970); and ROBERT B. ORTON, *The Climate of Texas and the Adjacent Gulf Waters* (1964).

(R.A.G.)

Mexico, History of

Before the arrival of Europeans in the Americas, Mexico was the home of a number of advanced cultures (see MESO-AMERICAN CIVILIZATION, HISTORY OF). From the time of the Spanish conquest (1519–21) until it achieved independence in the early 19th century, it was part of the viceroyalty of New Spain, which, until *c.* 1700 embraced, in addition to Mexico, the Philippine Islands, the Spanish Caribbean islands, and Central America down to Panama. Its ill-defined northern frontiers included what are now Florida and parts of Georgia in the southeastern United States and much of the U.S. Southwest, including California. These sprawling, loosely connected jurisdictions were governed by the Habsburg monarchs of Spain through a series of interlocking and overlapping bureaucracies. In Spain itself the Council of the Indies administered the overseas realms in the name of the king, whose prime emissary was the viceroy in Mexico City, who headed numerous subordinate hierarchies. The basic jurisdictional unit was a county-sized area ruled by a so-called alcalde mayor, or corregidor, as the viceroy's local representative.

I. Colonial period after 1701

A fundamental shift in the governance of New Spain occurred as a result of the War of the Spanish Succession (1701–13), when Bourbons replaced Habsburgs on the Spanish throne. Ruling by divine right, the Bourbon kings were enlightened despots whose major interests lay in increasing the economic returns from the Spanish Empire; they introduced many French practices and ideas into overseas administration.

Among the notable administrative reforms undertaken by Charles III in 1784 was the creation of 18 intendancies. In place of the numerous *alcaldías mayores* and *corregimientos*, which were abolished, *partidos* of about the same size were created within each intendancy. Headed by the intendancy of Mexico, each intendant was given considerable autonomy in increasing economic production within his sphere, developing useful arts and sciences, and bettering education and social conditions, all of the latter less for altruistic than for economic reasons.

Fed by currents of rationalism from England and Europe, the Enlightenment in Spain and Mexico spurred the spread of new scientific knowledge and, especially, its application to mining and agriculture. Political liberalism became a factor when the American and French revolutions implied the denial of the divine right of kings. A later development was the proliferation of the military establishment under the impact of British and Russian pressures on the frontiers of New Spain. Having strung a series of mission-forts across northern Mexico, authorities in Madrid and Mexico augmented the few regular Spanish troops that could be spared from the Peninsula by fostering a local militia with special exemptions (*fueros*) granted to Creole (Mexican-born) officers. Thus, an explosive combination of new ideas, guns, and administrative confusion between the old Habsburg and the new Bourbon administrative system all made almost simultaneous appearances.

The turmoil of Napoleonic Europe was the immediate background of the move for Mexican independence. Napoleon Bonaparte occupied Spain in 1808, imprisoned King Ferdinand VII, and attempted unsuccessfully to impose his brother Joseph Bonaparte as monarch. Rebelling, the Spanish resurrected their long-defunct *Cortes* (representative assembly) to govern as regent in the absence of the legitimate king, and, with representation from the overseas realms, the *Cortes* in 1812 promulgated a liberal constitution in the king's name. The document provided for a constitutional monarch, popular suffrage, a representative government, and other features taken from the French and U.S. constitutions.

Continental events had a factionalizing effect in Mexico as contradictory commands were received from Spain. The viceregal establishment put down sporadic rebellions by those who professed loyalty to the imprisoned king but who demanded some form of self-government.

Factionalism and rebellion

The most important local revolt was sparked by Miguel Hidalgo y Costilla, a parish priest in Dolores, Michoacán, who had been the centre of a group of "enlightened" reformers. On September 16, 1810, Hidalgo issued the "Grito de Dolores" ("Cry of Dolores"), calling for the end of rule by Spanish peninsulars (*gachupínes*), for equality of races and for redistribution of land. Mexican Independence Day commemorates this event.

Warning that the Spaniards would deliver Mexico to the godless French, Hidalgo exhorted his followers to fight and die for the Mexican Virgin, Our Lady of Guadalupe. When Hidalgo left his tiny village, the tens of thousands who flocked to his banner followed him into Guanajuato, a major colonial mining centre, peopled by Spaniards and Creoles. There the leading citizens barricaded themselves in a warehouse that was captured by Hidalgo on September 28. Most of the Creole elite were massacred, and the town was pillaged.

The Guanajuato massacre swung moderate and undecided support behind the Viceroy's efforts to crush the Hidalgo rebellion, lest a full-scale caste war ensue. Royalist forces defeated Hidalgo at the Bridge of Calderón on January 18, 1811. On March 19 he was captured along with other major insurgent leaders. On July 31 Hidalgo was executed, ending the first of the political civil wars that were to wrack Mexico for three-quarters of a century.

The Hidalgo cause was taken up by his associate José María Morelos y Pavón, another parish priest, who won control of substantial sections of southern Mexico. The constituent congresses, which Morelos called at Chilpancingo in 1813, issued at Apatzingán in 1814 formal declarations of independence and drafted republican constitutions for the areas under his military control.

At about the same time Napoleonic troops were withdrawing from Spain, and in 1814 Ferdinand VII returned from involuntary exile. One of his first acts was to nullify Spain's 1812 constitution drawn up by the *Cortes*. The departure of the French freed Spanish troops needed to crush the Morelos revolution. Captured and defrocked, Morelos was shot as a heretic and revolutionary on December 22, 1815. Scattered but dwindling guerrilla bands kept alive the populist, republican, nationalist tradition of Hidalgo and Morelos.

Mexican independence came about almost by accident when constitutionalists in Spain led a rebellion that, in 1820, forced Ferdinand VII to reinstate the liberal constitution of 1812. Conservatives in Mexico, alarmed that anticlerical liberals would threaten their religious, economic, and social privileges, saw independence from Spain as a method of sparing New Spain from such

changes. They found a spokesman and able leader in Agustín Itúrbide, a first-generation Creole and the son of a moderately wealthy Spaniard. Itúrbide, who had served as a loyal royalist officer against Hidalgo and others, had been given command of royal troops with which he was to snuff out remnants of the republican movement, then headed by the future president Vicente Guerrero in southern Mexico.

But, while ostensibly fighting Guerrero, Itúrbide was in fact negotiating with him to join a new independence movement. In 1821 they issued the so-called Plan of Iguala, a conservative document the 23 articles of which declared that the Mexican nation was to be independent, that its religion was to be Roman Catholicism, and that its inhabitants were to be united, without distinction between Mexican and European. It stipulated further that Mexico would become a constitutional monarchy under Ferdinand VII, that he or some Spanish prince would occupy the throne in Mexico City, and that an interim junta would draw up regulations for the election of deputies to a congress, which would write a constitution for the monarchy.

The Plan of Iguala

United as the Army of the Three Guarantees (independence, union, preservation of Roman Catholicism), the combined troops of Itúrbide and Guerrero controlled most of Mexico by the time Juan O'Donojú, recently appointed Spanish captain general, arrived in the viceregal capital. Without money, provisions, or troops, O'Donojú felt himself compelled to sign the Treaty of Córdoba on August 24, 1821, thus officially ending New Spain's dependence on Old Spain. The convention provided that the Mexican nation, thenceforth to be styled the Mexican Empire, was to be recognized as independent. Under the treaty, the new Empire's congress was to elect an emperor if no suitable European prince could be found.

II. Mexico from independence to revolution (1821–1910)

INDEPENDENCE

Although Spain at first disavowed O'Donojú's actions as unauthorized, the date now recognized as that of separation of New Spain from Old Spain is in fact August 24, 1821.

The Mexican Empire, 1821–23. The first Mexican Empire spanned only a short transitional period from colony to republic. Independence had been the point on which republicans and conservatives alike could agree. The new constitutional monarchy shielded the conservatives from a Spanish government that had suppressed religious orders, tried clerical cases in civil courts, and decreed that the church could not acquire real property. Such a government was also palatable to the insurgents, whose major objective was independence.

Itúrbide became president of a council of regents, which convoked a Sovereign Constituent Congress. Deputies to the congress represented the intendancies, the name they retained during the empire. When representatives from the Central American intendancies decided that their areas did not wish to remain part of the Mexican Empire, they were allowed to withdraw and to organize their own independent governments.

On the evening of May 18, 1822, military groups proclaimed Itúrbide Emperor Agustín I, and on the next day a majority in congress ratified the "people's choice" and recommended that the monarchy be hereditary, not elective. Agustín I was crowned in a long ceremony on Sunday, July 21. The empire was recognized by the United States on December 12, 1822, when the Mexican minister was officially received in Washington. But even then Agustín's power and prestige were already ebbing, and conflict soon developed between the military hero Itúrbide and the primarily civilian congress. On October 31, 1822, the Emperor dismissed congress and ruled through an appointed 45-man junta. The act, which was condemned by many as arbitrary, provided discontented military men with a pretext to revolt. Among their leaders was Gen. Antonio López de Santa Anna, whom Agustín had first promoted for services in the Army of the Three Guarantees and later insulted in public. In Veracruz, Santa Anna proclaimed a Mexican Republic on December 2, 1823, and was supported by old guerrillas. Other military men had similar plans. Agustín was forced to reconvene congress and to abdicate. In 1824 he returned from European exile but was arrested and shot. This first epoch of independent Mexican national life thus foreshadowed many problems of the succeeding republic.

The early republic. When a republican constitution was adopted in 1824, the Mexican people had had little or no previous experience in self-government. Their economy was precarious; mining, a mainstay in colonial times, had declined, and widespread anti-Spanish feelings had caused an exodus of Spaniards, depleting both the nation's capital reserves and its pool of trained people. Political instability made borrowing abroad expensive, and nearly all public revenues had to come from customs receipts, which were pledged well in advance. As Mexico's national debt mounted, so did its problems. A vicious, seemingly unbreakable cycle marked the first half century of Mexican national life. Whenever public monies were insufficient to pay the army, its officers revolted, captured the government, and negotiated international loans. The high interest payments on such loans reduced available funds for education and other social and cultural improvements, which many Mexican leaders thought were urgent requirements.

Troubles of the new republic

The constitution of 1824 provided for a federal republic, consisting of 19 states, four territories, and a federal district; it also set a number of goals, toward which the nation subsequently inched its way. Although the monarchical idea survived for nearly half a century more, Mexico remained a democratic republic. The constitution abolished slavery, which was never a major issue, and enfranchised all inhabitants; Indians thus lost their special colonial status as wards of the government. Restrictive state legislation, however, excluded the great mass of peasantry from the political process; power rested in the hands of a small Creole and Mestizo (persons of mixed Spanish and Indian ancestry) elite.

Under various labels, two factions contended for control. Generally conservative, the Centralists favoured a strong central government in the viceregal tradition, a paid national army, Roman Catholicism as the exclusive religion, and minimal regional and state autonomy; they supported a union of church and state. Opposed to them were the Federalists, who favoured limited central government, local militia, and nearly autonomous states; they tended to be anticlerical and opposed the continuance of colonial *fueros*, which gave special status to ecclesiastics and the military and exempted them from various civil obligations.

The pendulum of power swung back and forth between the two groups. In 1824 Manuel Félix Fernández Guadalupe Victoria, a Federalist and a leader in the independence movement, was elected president. Centralists replaced Federalists in 1828. A Federalist revolt in 1829 put Vicente Guerrero in the presidential chair, but he was soon overthrown by the Centralists, who held power until 1832. In 1833 another change placed Federalists in power until 1836, when Centralists again regained control and held it for nearly a decade.

The age of Santa Anna: Texas and the Mexican War. After the downfall of Agustín Itúrbide, Mexican politics revolved for some time about the enigmatic personality of Antonio López de Santa Anna, a charismatic general with seemingly few fixed ideological or political beliefs.

Allied with the Federalists, Santa Anna was first chosen president in 1833, but, rather than serve, he placed the liberal vice president Valentín Gómez Farías at the head of the government until Farías and his group in 1834 attacked the privileges of the clergy. Then Santa Anna assumed his presidential post and nullified the anticlerical legislation.

Santa Anna was president when difficulties over Texas first began to mount. Under favourable terms, some 30,000 U.S. immigrants had populated that previously desolate area. Fearful that their growing numbers posed a threat, the Mexican government in 1830 closed the bor-

Problems over Texas

der to further immigration and imposed on the Texans oppressive restrictions that contravened the Mexican constitution. In March 1836 Texas declared itself an independent republic, and Santa Anna, the leading military figure of the day, gathered an army to crush the revolt. He met with initial success when he trapped a small Texas garrison at the Alamo and totally eliminated it, but he was defeated and captured by Texas forces in April 1836 and subsequently freed. Though Mexico made no further efforts to reconquer Texas, it refused to recognize its independence.

At that time a doctrine now known as Manifest Destiny was at its height in the United States. It expressed a belief that it was the destiny of the United States to occupy all the North American continent and perhaps all of Mexico. From Spain, in 1819, the United States had obtained Florida and New Spain's southeastern territory to the Mississippi; next it annexed the Republic of Texas in 1845, a move that caused the Mexican government to break off diplomatic relations. Santa Anna was overthrown for his apparent willingness to negotiate with the United States. The Centralists installed Mariano Paredes y Arrillaga.

Although the United States claimed that the southern boundary of Texas was the Rio Bravo del Norte, the fact is that before annexation Texans had not settled below the Nueces River. Shortly after his election in March 1845, United States Pres. James K. Polk tried to secure an agreement on the Rio Grande boundary and to purchase California, but the new Paredes government refused to discuss either matter. Polk then ordered United States troops to occupy the disputed territory between the rivers. When Mexican and United States patrols clashed in April 1846, Polk asserted that American blood had been shed on American soil—an outrage that required action. Less warlike politicians, such as the Illinois Congressman Abraham Lincoln, to no avail submitted resolutions asking Polk to point out the precise location of this outrage. Polk's congressional majority formally declared war on Mexico in April.

War with the United States

Without major difficulty, United States troops captured New Mexico and what the Mexicans called Upper California (now the state of California). Gen. Zachary Taylor led the main United States force to quick victories in northeastern Mexico. At that juncture the Paredes government was overthrown, and Santa Anna re-emerged as president in September 1846. Almost immediately, Santa Anna mobilized Mexican forces and marched northward, boasting that the superior numbers and courage of his men meant that he would sign a peace treaty in Washington. Although Taylor and Santa Anna fought a close battle at Buena Vista, Santa Anna was beaten and forced to retreat on February 23, 1847. Both sides sustained heavy losses.

A change in U.S. strategy left Taylor holding ground in northern Mexico; it was decided that Mexico could be beaten only by capturing Mexico City, via Veracruz. Gen. Winfield Scott was given command of that expedition. On April 18, 1847, he defeated Santa Anna in the critical battle at Cerro Gordo. Though Mexican resistance continued to be formidable, Scott captured Mexico City on September 14, 1847. Santa Anna went into voluntary exile, while a new Mexican government negotiated peace.

Dated February 2, 1848, the Treaty of Guadalupe Hidalgo terminated the war. Under its terms Mexico ceded all territory north of an irregular line of the Rio Grande and the Gila River across the Colorado to the Pacific. The United States paid Mexico $15,000,000 and assumed $3,000,000 in claims held by United States citizens against Mexico. At the time, the large area involved was lightly populated and economically unexploited.

After the war, Santa Anna figured in one more major episode before the political scene changed. In 1853 conservatives seized power and invited him to become dictator. Among other things, on December 16, 1853, Santa Anna decreed that the dictatorship should be prolonged indefinitely and that he should be addressed as "His Most Serene Highness." To raise funds for an expanded army, he sold the Mesilla Valley in southern Arizona to the United States for $10,000,000; this Gadsden Purchase, as it is now called, was the last significant boundary change of the Mexican Republic.

An era was ending. The death of Lucas Alamán on June 2, 1853, deprived the conservatives of their most important intellectual leader. On September 11 his distinguished colleague, José Mariáno y Mendivil Tornel, also died. These men, with other like-minded Mexicans, had sought to give their country a national culture and an economic independence that would free the nation from colonialism but retain the Spanish religious heritage. Though they played subordinate administrative roles, their collective impact on Mexico may have been more important than the antics of Santa Anna.

The reform. Since independence a new generation of Mexicans had been born; appalled at the easy victory the United States had won, the more thoughtful among them felt that Mexico's survival as an independent nation depended on fundamental reform. Among the new faces was Benito Juárez, a Zapotec Indian educated as a middle class liberal, who had moved to New Orleans and had discussed and planned Mexico's future with fellow expatriates. With no military force to implement their plans, they bided time until their opportunity came, in 1854, when Juan Álvarez, a surviving hero of independence, and Ignacio Comonfort, a political moderate, proclaimed a liberal rebellion against Santa Anna and forced him out of the presidency. (Santa Anna went to Colombia in exile in 1855; almost unknown, he returned to Mexico City before his death on June 21, 1867. Congress refused to pension his widow or to restore posthumously his military honours and titles.)

Neither Álvarez nor Comonfort, who followed him in the presidency, had any clearly defined program. The role of the returned expatriates was to act as a brain trust to carry out *La Reforma* ("The Reform"). Its aims were to abolish remnants of colonialism by removing special ecclesiastical and military privileges; to separate church and state by secularizing education, marriages, and burials; to reduce the economic power of the church by forcing it to sell its properties; to foster an economic development that envisaged Mexico as a country of yeoman farmers and small industrialists; and, above all, to establish a single standard of legal justice, based on laws passed by democratically elected representative assemblies.

Juárez was made minister of justice. Among his first reforms was the so-called Ley Juárez (November 23, 1855), which abolished *fueros* and the use of special military and ecclesiastical courts in civil cases. The minister of finance, Miguel Lerdo de Tejada, sponsored Ley Lerdo (June 25, 1856), which restricted the right of ecclesiastical corporations to own lands by decreeing that church lands not directly used for religious purposes must be sold and lands held in common by Indian communities (ejidos) be distributed to individual villagers.

Constitution of 1857

The reformers called a convention to revise the 1824 constitution, several provisions of which had been partially suspended or changed. A new constitution, which provided a legal base for the reform, was promulgated on February 12, 1857, but did not become effective until the following September 16. It prohibited slavery and abridgements of freedom of speech or press; it abolished special courts and prohibited civil and ecclesiastical corporations from owning property, except buildings in use; it eliminated monopolies; it prescribed that Mexico was to be a representative, democratic, republican nation; and it defined the states and their responsibilities. This constitution, which remained in force until it was modified in 1917, increased the power of the central executive.

Neither the religious nor the military accepted the 1857 constitution, and both inveighed against the reform, calling for retention of "religion and *fueros*." The church excommunicated all civil officials who swore to support the constitution. When civil war erupted, Comonfort went into exile after his efforts at compromise had failed; Juárez automatically succeeded him as constitutional president. The conservatives captured Mexico City and

set up a competing regime. Juárez and his government moved to Veracruz, where they controlled the customs receipts.

Foreign powers became involved in the Mexican struggle. On April 6, 1859, the United States recognized the Juárez government; Pres. James Buchanan permitted war material to be shipped to Juárez' forces. Americans were encouraged to serve the liberal cause as volunteers; Spain and France generally favoured the conservatives, as did Great Britain.

Juárez in July 1859 issued a series of decrees: all church property except buildings used for worship was to be confiscated without compensation; all marriages apart from civil were declared annulled; the formal separation of church and state was proclaimed; cemeteries were declared public property, and burial fees were abolished. Moneys from the sale of confiscated church property, though less than anticipated, speeded the end of the civil war. On December 22, 1860, the liberals won a critical battle, and, when the conservative president, Miguel Miramón fled, the conservative cause collapsed. The victorious liberal army of about 25,000 men entered Mexico City on New Year's Day, 1861. On his return (January 11) Juárez was greeted by an enthusiastic populace who welcomed the end of the long and devastating civil war and the re-establishment of government under the constitution of 1857.

French intervention. Exiled Mexican conservatives, who continued to intrigue, enlisted the help of a powerful ally, Napoleon III, the ruler of France who wanted to create a Latin league that would include the Mediterranean lands and the former possessions of Spain and Portugal in the New World as well. (The term Latin America dates from this time and concept.) With its strategic position and its economic potential, Mexico seemed especially attractive to the Napoleonic imperial scheme. A French bastion in Mexico would check the Manifest Destiny of the United States and provide a base from which Central and South American protectorates could be added. The fact that the United States was engaged in a civil war (1861–65) was a determining element. In 1861 Napoleon III was willing to believe the Mexican conservatives' assertions that the masses of Mexican people would support his intervention to restore religion and set up a monarchy.

Ambitions of Napoleon III

In Maximilian of Habsburg and his wife, Charlotte (Carlota), daughter of the king of Belgium, Napoleon found eager aspirants for the Mexican throne. With Napoleon's promise of support, Maximilian agreed to become emperor of Mexico if a plebescite showed that the Mexican people wanted him. Reluctantly, however, he renounced any future claim to Habsburg succession.

Meanwhile, France, Spain, and Great Britain prepared to send a punitive expedition to collect debts owed them in Mexico. Juárez in 1861 had suspended for two years all payments on a foreign debt of $80,000,000, a sum that excluded some questionable French bonds amounting to $15,000,000. The intervention was spearheaded by Spain, the forces of which landed at Veracruz on December 14, 1861 and were followed soon after by French and British contingents. When the allies fell into dispute over the $15,000,000 French claim for payment of the bonds, both Spain and Great Britain disengaged from the joint venture.

The French expeditionary force began its march toward Mexico City. When the Mexican army made a stand at Puebla on May 5, 1862, the French retreated to await reinforcements. Napoleon dispatched 30,000 more troops under the command of the French general Élie-Frédéric Forey. The Mexicans could not withstand French might, and on June 10, 1863, Forey rode as conqueror into Mexico City. The French rapidly secured much of central Mexico, forcing Juárez and his government to keep constantly on the move in the north.

French occupation forces set up an Assembly of Notables and a council of regents to govern Mexico. The assembly, as expected, offered the throne to Maximilian in the name of the Mexican people. Assured of Napoleon's continued military support and the economic backing of the British, Maximilian and Charlotte arrived in Veracruz on May 28, 1864, having passed through Rome to confer with the Pope before they embarked. On June 12 Emperor Maximilian was welcomed in Mexico City.

Emperor Maximilian

Maximilian attempted to follow a policy of national conciliation, hoping to unite Mexican factions and interests. But he proved too much a Habsburg to be an effective tool of Napoleon's schemes, too much a liberal to please the conservatives who had engineered his coronation, and too tainted by conservative sponsorship to win republican support.

Maximilian was perhaps less naive than he has been pictured. Not fully convinced that a majority of Mexicans welcomed him as emperor nor sure that he should place full reliance on French troops, he tried to create a Mexican rural guard and a separate imperial army around a nucleus of Austrian and Belgian volunteers. His proposal to reduce the number of regular Mexican army generals to 18 brought early disillusionment to the army, a mainstay of conservatism, which had expected Maximilian to be a puppet. The Mexican officers became further irritated when, with Maximilian's approval, French officers outranked them. At this time the Habsburg emperor and the French commander, Achille François Bazaine, were following conflicting army policies, Napoleon having ordered Bazaine to Mexicanize the army to cut time and costs of occupation.

At the same time, clerical groups, eager to reap rewards for their efforts on the Emperor's behalf, pressed him to reverse the reform. A papal nuncio from Rome arrived with a message asking that Maximilian revoke the anticlerical laws, establish Roman Catholicism as the exclusive religion, restore the religious orders, remove the church from its dependence on civil authorities, turn education over to ecclesiastics, and return properties confiscated and sold by the republicans. Replying that he, not outsiders, would decide such matters, Maximilian issued decrees establishing religious toleration, with Catholicism favoured but still dependent on the state. He confirmed that the previous sales of church property under the Reform laws were legal and that revenues the church had received from property Juárez had nationalized were to be ceded to the state. Thus, Maximilian's conservative support further dwindled because the clergy and their followers felt betrayed.

In September 1864 Maximilian took what amounted to a guided tour of the cities that supported his empire. The warm welcomes he received led him to conclude that a majority of Mexicans wanted peace and justice, which the activities of the republican guerrillas threatened. He therefore decreed on November 4 that, thenceforth, republicans would be considered bandits and brigands, subject to extreme penalties; this negated Maximilian's attempts to woo their supporters by inviting them into his council of state.

In 1865 French troops chased Juárez to, but not over, the United States border. Believing that the Mexican president had left national territory and that republicanism had therefore collapsed, Maximilian on October 2 issued a strong decree, ordering that all guerrillas be court-martialled and shot within 24 hours; the same penalties were to apply to persons who hid, aided, or otherwise helped them. Juárez had earlier issued a parallel decree (January 1862) against those who aided the French interventionists and imperialists.

French troops, though they were effective in keeping republicans on the outskirts of major productive areas, were also costly. Napoleon's Mexican adventure came under heavy press and parliamentary fire in France as costs in men and francs mounted without economic or political advantages to compensate for the expenditures. There was criticism from overseas as well; with its civil war ended, the United States turned to the international scene. William H. Seward, the United States secretary of state, brought mounting diplomatic pressure on Napoleon to withdraw French troops; in February 1866 Napoleon agreed to clear foreign troops from Mexico by November 1867. A United States request to the Austrian government to stop enrollment of volunteers for the

Withdrawal of foreign troops

Mexican imperial army also brought an affirmative response.

In liquidating his Mexican venture, Napoleon said that, since Maximilian had not carried out his part of the pact to bring peace and orderly government to Mexico, the French were relieved of their obligations for military and financial support. Stunned, Maximilian sought to have the decision reversed. The empress Charlotte tried without success to persuade the French ruler to honour his solemn pledge. From Paris she travelled to Rome to plead with the Pope; there she went mad and was taken to Belgium, where she lingered insane until her death in 1927. Bazaine, ordered to withdraw all French troops immediately, auctioned off military material not worth shipping to France—horses, saddles, etc.—and destroyed large supplies of powder and projectiles rather than turn them over to Maximilian. In February 1867 Bazaine left Mexico City, and by March 12 all 28,690 men had embarked. Napoleon made Bazaine the scapegoat of the Mexican fiasco, blaming its high costs and lack of results on his mismanagement, and implied personal dishonesty.

As French power withdrew, republican forces reconquered Mexico, and Maximilian was left with only a regiment of Austrian hussars, a battalion of infantry, and a small army of relatively untrained Mexican draftees. After first considering abdication, Maximilian decided to defend his imperial status and his honour as a Habsburg by making as strong a stand as possible, though planning, in the event of defeat, to negotiate an honourable exile. With these resolves he concentrated most of his troops—9,000 men—at Querétaro, a city loyal to the imperial cause. On May 5, 1867, the republican forces laid siege, initially with 32,000 men, later with an additional 10,000. By May 14 the starving imperialist force, reduced to about 5,000, had decided to withdraw and take a stand in the mountains. A disaffected imperial officer, in return for a promise that Maximilian be spared, placed republican soldiers, whom he passed off as relief troops, at strategic places in Querétaro. The siege ended the next day when Maximilian and his generals surrendered.

Under Juárez' decrees of 1862, Maximilian and his two leading generals were court-martialled and sentenced to death by firing squad. President Juárez refused to be swayed by the petitions for mercy that poured in from foreign governments. He wanted to demonstrate that Mexico could act independently; that, as the reform contended, all men were equal under law; and that foreign monarchical adventures in Mexico were futile. Maximilian and the generals were executed on June 19, 1867, terminating a bizarre interlude of Mexican history.

Emperor Maximilian left few permanent legacies. He did beautify and rehabilitate Mexico City, and his castle at Chapultepec Park became a national historical museum overlooking the broad avenue he planned—ironically named Paseo de la Reforma.

The restored republic. At least 50,000 persons lost their lives in the struggles between republic and empire, and when Juárez re-entered Mexico City on July 15, 1867, his immediate task was to abate the rancors of civil war. The vindicated Juárez regime took few major reprisals—principal imperialists were fined, some were imprisoned for short terms, and a few were exiled. One of Juárez' first acts was to start rebuilding the shattered economy. In an era of goodwill engendered by the sympathy and aid the United States had extended to the Mexican cause, the claims of the two countries against each other were settled by peaceful arbitration. Diplomatic relations were gradually re-established with Europe; new treaties were aimed at furthering trade.

Rebuilding the republic

In December 1867 Juárez was re-elected president. Apart from trying to foster political tranquillity, his main aims were to improve public education and to put the economy on a sound footing. To avoid the problems of dogma and religious opinion that had divided Mexicans, Juárez entrusted the development of a national educational system to Gabino Barreda, a follower of the French thinker Auguste Comte, who had said that the human mind and society passed through three successive stages—religious, metaphysical, and positive. Known as Positivists, Barreda and his followers contended that the reform, by displacing the church and militarism, had done away with the earlier two stages and that Mexico was in the third, or positivist, stage. The public-education law for the Federal District, which was to serve as the national model, stressed the secular state as the inculcator of scientific ethical norms, with "Liberty, Order, and Progress" as the means, base, and product of the system.

The chief architect of economic rehabilitation was Matías Romero, who had been Juárez' minister in Washington, and who believed that Mexico's development was dependent on three basic elements: immigration, communication networks, and the exploitation of natural resources. In 1867 and 1868 the government renewed concessions to British capitalists for the completion of the Veracruz–Mexico City railway and issued concessions for others; it authorized the opening of new roads and the extension of the telegraph system. Work was begun on reforming the tax systems and tariff schedules.

The re-election of Juárez in 1871 was contested more heatedly than that of 1867 had been. Thereafter, despite formidable opposition in congress, tariff reform was approved, as was Mexico's adoption of the metric system, which ended the chaotic colonial system of weights and measures.

After a short illness, Juárez died suddenly on July 18, 1872, his death closing one era and opening another. Behind him lay Mexico's long colonial history and its partial survivals through the early 19th century. The notion of a Mexican monarchy had been forever buried with Maximilian. Under Juárez, Mexicans had begun to modernize the economy and some of the social institutions, to expand rail, road, and telegraph networks, and to develop secular education.

Juárez' death also brought temporary political peace. Without incident, Sebastián Lerdo de Tejada, the president of the supreme court, and next in line of succession, was sworn in as acting president on July 19. Congress immediately began to lavish posthumous honours on Juárez, who by his innate abilities and great strength of character had led his people through unprecedented travail. He remains a major figure in the history of Mexico.

Temporary political peace

A national election placed Lerdo in the presidential chair in his own right on November 16, 1872. The course Juárez had charted remained unchanged. On New Year's Day, 1873, the Veracruz–Mexico City railway was inaugurated. Other works were commissioned. Congress and the executive branch of government continued to dole out railway and telegraph concessions. To safeguard the nation against future bloodshed, congress on May 31 added specific reform laws to the constitution of 1857—church and state were explicitly declared independent of one another; freedom of religion was proclaimed; church acquisition of real estate was abolished; religious oaths were banned in civil courts; forced labour was forbidden; and liberty of man in respect to labour, education, and religion was declared inviolable. The degree to which the reform had triumphed was evidenced by the fact that no national movement developed against these additions to the organic laws. The Lerdo government in 1874 renewed diplomatic relations with France, Spain, and Prussia.

The age of Porfirio Díaz. From 1876 until a political revolution unseated him in 1910, the personality of Porfirio Díaz dominated the history of his country. Like Juárez, Díaz was a poor Indian from Oaxaca but of Mixtec rather then Zapotec heritage. Educated locally, he had chosen a military career and had become an outstanding general in the republican cause against the French intervention and empire. Although he vied for the presidency against Juárez in 1867 and again in 1871, their ideological differences were not great. When in 1875 Lerdo ran for re-election, Díaz led a successful revolt and assumed the presidency in November 1876.

Díaz took the blueprint for Mexico's future that Juárez and Lerdo had elaborated and implemented it. While clearing out pockets of political resistance, he turned the presidency over to Manuel González, a companion in arms. Díaz won the election again in 1884 and was regu-

larly returned to that office through 1908. During his long regime he scrupulously kept democratic and constitutional forms intact, partly in the conviction that it was the president's duty to train the unready Mexican people to use them properly.

The *científicos*

Around him Díaz gathered many intellectuals (the *científicos*), who stressed the need for rational planning and development. The emphasis was on economic development to assure social progress. How such development was to be achieved was translated into one of Díaz' political slogans, "pan o palo" ("bread or the club"), meaning that acquiescence to official policies would insure livelihood, even wealth, but failure to agree would bring sure reprisals—harassment, imprisonment, death. Another slogan was "few politics, much administration." Liberty was dropped from the earlier positivist triad of liberty, order, and progress. It was the price the Mexican people were expected to pay for the benefits the *científicos* policies would provide.

Capital, though badly needed, could not be attracted until Mexico had tidied up its international and national fiscal affairs. Mexican finances were placed on a solid base, and a stable currency was established. With guarantees of political and social tranquility, foreign investment was encouraged and obtained. European and United States funds built some 15,000 miles of railways, provided electricity and streetcars for the cities, created industrial complexes, rehabilitated port facilities, and developed the mining of industrial metals. Early petroleum concessions to foreigners laid the groundwork for serious problems later, when world navies shifted from coal to oil and when automobiles were mass produced.

A complex mechanism in which all major and most minor decisions rested in the hands of the president evolved during the first two decades of the Díaz regime. The success of the practice rested on self-interest; Díaz made it worthwhile for everyone to support the system. For the most part, the small body of intellectuals was absorbed into the expanding bureaucracy or the subsidized press. The army and the church were made handmaidens of the regime, rather than its adversaries; generals were encouraged to become entrepreneurs. While retaining the reform laws on the books, Díaz was purposely conciliatory toward the church and allowed it to regain some of its former economic power without letting it develop significant political influence.

The regime also perfected instruments of repression and control, though for many years it used them sparingly. Forced military conscription lay in the hands of the president's appointees. Troublesome elements such as the Yaqui Indians were inducted en masse and shipped as cheap labour to the sisal plantations of Yucatan or the tobacco fields of Oaxaca. Governors were personal appointees of the president. Their actions were monitored by *jefes políticos* ("political chiefs"), who reported directly to the president and on his authorization intervened in municipal and state affairs. An elite constabulary, the *rurales*, like the Texas Rangers and Canadian Mounted police, created a myth of ubiquity that eliminated the brigandage and banditry characteristic of earlier 19th-century Mexican countrysides.

In an era in which material success was highly regarded, Díaz' accomplishments were praised, but his popularity began to decline, perhaps about 1895. Prosperity had been pre-empted by a relatively small group, many of whom were foreigners. The Mexican economic and social elite self-consciously aped European (especially French) modes of dress, education, and even language. Between the affluent and the growing urban and rural proletariat there was an expanding middle class body. Meanwhile, a new generation, which could recall none of the chaos of pre-Porfirio days, began to question the system. As criticism increased, so did repression. It was from this interaction that the Mexican Revolution resulted.

PRECURSORS OF THE REVOLUTION

From about 1900 a majority of Mexicans began to question the nation's apathetic acceptance of the Porfirian peace. The earliest and most vocal critics were Mexican radical groups, perhaps the most important of which called itself Regeneration. Its members were anarchists who adapted their dogmas to the Mexican scene. While always small in numbers and often ineffective in actions, this group had great influence. Many of the reforms and programs it advocated were embodied in the Mexican constitution of 1917.

The Regeneration group

The leader of the Regeneration group was Ricardo Flores Magón, who had been born in Oaxaca of an Indian mother and a Mestizo father and had been sent for further education to Mexico City, where he had turned to idealistic student activism. For leading a small demonstration against the re-election of Porfirio Díaz in 1892, he was jailed for the first of many times. The group's movement took form in 1900, when Camilo Arriaga, a well-to-do engineer in San Luis Potosí, organized first a club and then a small party to restore the liberalism of Juárez. Arriaga called a national meeting of liberal clubs in 1901, and a short time later most of the small band were jailed, and their newspaper, *Regeneration*, which Flores Magón edited, was suppressed. After they had served their prison sentences the young radicals fled to the United States and Canada, settling for a while in St. Louis, Missouri, where they formally organized the Mexican Liberal Party. Although its name suggests the earlier reform, in fact it was anarcho-syndicalist, dedicated to the overthrow of the Mexican government and the total renovation of Mexican society.

In 1906 the Regeneration group published a comprehensive program in the form of a manifesto that had wide, if clandestine, circulation in Mexico. It advocated a one-term presidency, guarantees of civil liberties, breaking the hold of the Catholic Church, vast expansion of free public education, and land reform. It asked that Mexican citizenship be a prerequisite to property ownership and that unused land be distributed to the landless. The manifesto proposed confiscation of the wealth that Díaz and the *científicos* had illegally acquired, the abolition of child labour, guaranteed minimum wages, and bettered working conditions for workers. In muted tones it criticized capitalism as a system of exploitation. The Regeneration group drew its main lines of thought from Mikhail Bakunin, a Russian revolutionary writer who believed that the power of any institution, including government that exercised controls over individuals, should be reduced.

Many charges by the Regeneration and similar groups were borne out when the Díaz troops, in bloody fashion, broke strikes in the textile region of Veracruz and the mining regions of the north. Several of these strikes had been fomented by Regeneration organizers at the same time that muckrakers in the United States were exposing the evils of corrupt government, big business, and other aspects of life in the United States and abroad. Possibly to refute their unfavourable reports, Díaz gave an interview in 1908 to a United States reporter that became a milestone in pre-revolutionary history. To blunt charges of one-man rule, Díaz very carefully but clearly said that in his view the time had come for Mexico to advance toward democracy, that he would welcome an opposition party in the 1910 presidential election, that he would be most happy to sustain and guide the opposition party, and that to inaugurate a democratic government in Mexico he would forget himself. This fell like a bombshell in Mexico, where most readers failed to note that he had not specified a time. It was widely believed that this implied he would not run for the presidency in 1910. Mexican newspapers not only reprinted the interview but also began openly to discuss Mexican political questions, long forbidden by the regime. Pamphlets and then books added to the political rumblings.

Political opposition to Díaz

Two main opposition groups soon emerged. One backed Gen. Bernardo Reyes as vice presidential candidate over Díaz's handpicked candidate. As an efficient and stringent governor, Reyes had transformed the state of Nuevo León into a prosperous, progressive model; he forthrightly opposed *científico* theories and practices and had initiated a state workman's compensation law.

The other main party, anti-re-electionist, had been cre-

ated largely through the efforts of Francisco I. Madero, then a political unknown, whose efforts elevated him to the highest place in the revolutionary pantheon as the "Apostle of Democracy." Born into one of the ten richest families in Mexico, whose agricultural enterprises spread over much of northern Mexico, Madero was educated in the United States and France. In his own right he became an enlightened entrepreneur and amassed a considerable personal fortune.

Madero became concerned with how badly Mexico was faring in comparison to the rest of the world. Mexican food prices were rising, and rural and urban standards of living were dropping. He attributed Mexico's retrograde position to the prolonged political dictatorship. He helped journalists to expose these matters and provided considerable early financial support to the Regeneration group, but he disassociated from them after about 1907, when it became clearer that they intended to destroy, not reform, the system.

Madero, setting about to organize a national party to compete in the 1910 elections, published *La Sucesión Presidencial en 1910* ("The Presidential Succession in 1910") as a campaign document, two-thirds of which dealt with the history of Mexico and the corrupting influences of absolute power and the rest with his program to revive the democratic usages that had atrophied for so long. The initial aim was to win the vice presidency.

Despite harassment, Madero carried on a vigorous and wearying campaign in the summer of 1909. During an interview with Díaz, he was surprised by the dictator's advanced old age and his remoteness from current issues. Díaz' condition and the enthusiasm of Madero's crowds led the Anti-Re-electionists to modify their goals. Rather than accept Díaz as the inevitable president in 1910, they decided to try for the presidency themselves.

On June 14 Madero was arrested and jailed and thus became the martyr and victim of the system he was trying peacefully to change. Since it was perfectly clear that Díaz was not going to permit free and honest elections, Madero and his followers decided that the only hope of improving Mexico was through armed revolt. On October 4, 1909, the Chamber of Deputies, which had assembled as the electoral college, declared that Díaz had been re-elected. On October 5 Madero managed to escape from San Luis Potosí jail. He arrived on October 7 in San Antonio, Texas, where with aides he prepared and issued, as of the day of his escape, the Plan of San Luis Potosí, which proclaimed the principles of "effective suffrage, no re-election." Madero declared that Díaz was illegally president of Mexico. Designating Sunday, November 20, as the day when citizens should take up arms against the government of Porfirio Díaz, he promised that a successful revolution would institute political reforms.

The Plan of San Luis Potosí

But on November 20, the official birthday of the Mexican Revolution, no mass uprisings took place. Small bands of guerrillas, most of them in northern Mexico, kept the rebellion alive while Madero used his family fortune to supply them with arms from Texas. When the guerrilla bands combined, they began to defeat Federalist forces, who held most of the strategic rail lines, especially those emanating from Ciudad Juárez, on the United States border, where the Federalists troops had consolidated. Until the revolutionists laid siege to that city, no more than 2,500 armed men were engaged in the Madero revolution.

III. Mexico since 1910

THE MEXICAN REVOLUTION AND ITS AFTERMATH, 1910–40

When armed outbreaks initiated the Mexican Revolution in 1910, their immediate object was to replace Díaz as president. What began as a relatively simple political movement broadened into a major economic and social upheaval that gave Mexico many unique characteristics. During the long struggle, the Mexican people developed a sense of identity and purpose, unmatched by any Latin American republic. Many reforms had been established by 1940, when the goals of the revolution were institutionalized as guidelines for future Mexican policies.

The military revolution. On February 14, 1911, Madero crossed into Mexico near Ciudad Juárez to head his forces. There they learned how debilitated the Díaz army had become; led by aged generals, the Federalist troops lacked discipline, cohesion, unity of command, and effectiveness. Under these circumstances, the revolution gained ground and momentum. On May 10 the Federal commander surrendered at Ciudad Juárez. That marked the beginning of the end. An agreement negotiated with the Díaz regime provided that Díaz would resign, that an interim president would call general elections, and that revolutionary forces would be discharged. On May 25 Díaz resigned and sailed for Paris, a bewildered former hero. Several revolutionary bands, including that of Emiliano Zapata, resisted being disarmed by the Federal army.

In the heated political atmosphere, congress on November 2 declared that Madero had won the presidency. The new government was able to withstand constant attacks from the right and left for only 15 months. A series of unsuccessful counterrevolutionary revolts culminated in a successful plot in February 1913. Following ten days of cannonading the heart of Mexico City, with high civilian casualties, Victoriano Huerta, commandant of government forces, betrayed Madero and arranged for his arrest and that of his vice president so that Huerta could assume the presidency himself. These plans were made with the knowledge, if not the connivance, of the United States ambassador, Henry Lane Wilson, who was thus defying Pres. William H. Taft's instructions to remain aloof from domestic Mexican politics.

Shortly thereafter, presumably on Huerta's orders, Madero was shot while being transferred from one prison to another. His death rekindled revolutionary fires. In northern Mexico, Venustiano Carranza, refusing to recognize Huerta as president, demanded that the office be elective, as specified in the constitution. Former chieftains such as Pancho Villa made loose alliances with Carranza. The revived revolution took on highly local and regional aspects.

The new president of the United States, Woodrow Wilson, was determined to oust Huerta and, on flimsy pretexts, landed United States troops at Veracruz and occupied it. All of the revolutionary leaders except Villa rejected this external intervention in a national struggle. The combined revolutionary forces unseated Huerta in 1914 but then split over who was to exercise presidential power. Zapata in Morelos and Villa in the north joined to fight the revolutionary groups under Carranza, the most important of which was headed by Gen. Alvaro Obregón, a former schoolteacher. Obregón won a decisive victory over Villa at Celaya in April 1915 but failed to bring the civil war to an immediate end.

Split among revolutionary groups

The Constitution of 1917. With most of central and southern Mexico under constitutionalist control, Carranza in 1916 convoked a constituent congress in Querétaro to revise and update the constitution of 1857. In the course of fighting, the economic and social demands of the radical precursors had become common slogans as contending revolutionary bands bid for popular support. The constitution of 1917 incorporated the aspirations of those groups involved in the revolution. A major concern of Zapata's followers was land reform. Carranza's advisors had prepared a draft document that included not only most of the revolutionary decrees about land reform but also the most advanced European social legislation.

The constitution of 1917 specifically incorporated the major features of the 1824 and 1857 charters regarding territorial organization, civil liberties, democratic forms, anticlerical and antimonopoly clauses, etc. To avoid abuses such as those that had occurred under Díaz, the office of vicepresident was suppressed, as was the system of *jefes políticos*. The constitution completely reversed the concept widely held in Mexico and elsewhere that government should take only a limited, passive role. Its philosophy was that the national government had an obligation to take an active role in promoting the social, economic, and cultural well-being of its citizens. Article 3

sketched a vast plan of secular, free, compulsory public education. Article 14 reaffirmed the sanctity of private property and contracts, but article 27 interjected concepts of social utility and national benefit to limit the untrammelled use of private property. The most important new concepts came in articles 27 and 123. The former reasserted national ownership of subsoil resources and outlined alternative land-reform and agrarian programs. The latter, the magna carta of labour and social welfare, was set apart to highlight its importance; in addition to guaranteeing minimum wages and the right to organize and strike, it gave labour social status and destroyed the concept of it as an economic commodity to be bought at the lowest rates to maximize profits. Article 123 also outlined a comprehensive system of social security, including public-health and welfare programs. Reflecting the nationalistic feelings of the revolutionaries, foreigners and foreign interests were placed under limitations. Article 33, for instance, permitted the president to deport foreigners at will.

The constitution of 1917, promulgated on February 5, thus set the goals toward which presidents were to work. As expected, Carranza was elected president and given de jure recognition by the United States. When Zapata was betrayed and killed in 1919, the last organized opposition to the Carranza-Obregón reorganization dissolved. Pancho Villa retired from active campaigning after his raids across the border, especially one in Columbus, New Mexico (March 9, 1916), had failed to embroil the United States in conflict with Carranza. Villa was ambushed and shot in 1923.

The northern dynasty: Obregón and Calles. When Carranza failed to move toward immediate social reforms, General Obregón enlisted two other powerful north Mexican chieftans, Plutarco Elías Calles and Adolfo de La Huerta, to join him in an almost bloodless coup; together they formed the northern dynasty. Carranza was killed as he fled from Mexico City, and Obregón took office as president December 1, 1920. The dynasty agreed that peace was needed to rehabilitate Mexico from the devastations of nearly a decade of civil upheaval. By bribery and minimum force, Obregón placated the horde of revolutionary generals and colonels.

Overthrow of Carranza

Under Obregón a start was made toward fulfilling the ideals set forth in the constitution. Administrative machinery was set up to distribute land to the landless and to restore communal holdings (ejidos) to villages. The government supported the Regional Confederation of Mexican Labour (CROM). José Vasconcelos, who was named minister of education, was to implement the program of rural education. He sponsored a cultural program that brought Mexico worldwide fame and importance. Radical mural painters such as Diego Rivera, José Clemente Orozco, and David Alfaro Siqueiros, who were commissioned to portray Mexican and especially revolutionary history on public buildings, exalted the Indian past, as did creative writers.

At the end of his term, Obregón stepped aside for Calles, after La Huerta's unsuccessful revolt in 1924 had eliminated him from the coalition. Calles' presidency followed the same general lines as had Obregón's. Land distribution was stepped up, an irrigation program was begun, and in 1925 renewed pressure was put on the petroleum companies to exchange for leases the titles they had obtained from Díaz or La Huerta. Problems with the Church developed when Calles instituted vigorously anticlerical measures; in retaliation, the church suspended all religious ceremonies and approved and possibly sponsored a rebellion in western Mexico known as the Cristeros. Mediation of the church–state controversy was unofficially accomplished by Dwight W. Morrow, the United States ambassador to Mexico, whose sympathetic and skillful diplomacy also eased tensions between the two countries.

In 1928 the presidential term was extended from four to six years, and the doctrine of "no re-election" was modified to mean "no successive re-election." Obregón was the successful presidential candidate, but, before he could take office, he was assassinated by a religious fanatic. With Calles unable to succeed himself, a peculiar, nearly unique Mexican single-party system came into being.

Single-party system

Calles formed the National Revolutionary Party, a coalition of regional and local military bosses and labour and peasant leaders. To safeguard the gains of the revolution, he excluded the church and other possible reactionary elements. With Calles as chief, the official party governed in the name of the revolution. A congress, drawn from party ranks, named successive, short-term presidents to fill out the Obregón term.

In the period 1928–34, a worldwide depression and increasing personal vested interests caused many of the older, now conservative revolutionaries, including Calles, to go slowly in implementing the reform mandates of the constitution. The ruling clique continued to be militantly anticlerical, but government support was withdrawn from the Regional Confederation of Mexican Labour, which disintegrated. On the positive side, the Calles years saw the beginnings of an irrigation and road-building program.

Resurgence under Cárdenas. Within the revolutionary family, Gen. Lázaro Cárdenas was a respected if not outstanding revolutionary. Having quietly and faithfully worked his way up the ladder of politico-military power during the Obregón and Calles years, he seemed a safe candidate in the 1934 elections. He was also acceptable to a powerful group within the party, who drafted a six-year plan incorporating extended revolutionary reforms. Although his election was a foregone conclusion, Cárdenas astonished everyone by waging a vigorous campaign, visiting even remote hamlets to bring the six-year plan to the people.

After his election Cárdenas strengthened the labour movement and took other steps that defied Calles. When the inevitable power test came, Cárdenas won, and Calles fled into temporary exile, after which Cárdenas renamed and reorganized the party on a functional national basis —the agrarians formed one sector; labour another; the military yet another; and bureaucrats, teachers, and miscellaneous groups made up the popular sector. The four sectors agreed to support the slate of candidates the party designated. The local, state, and national representatives made party policy and ratified the President's choice of candidates.

With massive popular support and with the power elites under control, Cárdenas tirelessly pushed toward revolutionary goals. He and his advisers elaborated the land-reform programs; using land expropriated from private owners, they created communal cooperatives and gave them ejido status. By the end of his term about 40 percent of the rural working force was under the ejido program. Cárdenas also nationalized railways and placed them under the management of labour.

Perhaps Cárdenas' single most spectacular action was the expropriation of foreign petroleum companies following a labour dispute in which unions demanded not only wage increases but also participation in what management considered its exclusive role. A series of court cases and special boards found in favour of the workers. When the companies refused to accept the decisions, declaring they could not and would not comply, Cárdenas on March 18, 1938, decreed expropriation of their holdings, thus nationalizing the petroleum industry. Petroleos Mexicanos (Pemex), then a small agency, was designated to administer the industry for the nation. The British government, whose nationals had a far larger stake than United States firms, immediately broke diplomatic relations. After a short delay the United States government, under Franklin D. Roosevelt, indicated that, if Mexico would make prompt and fair payments, he would not intervene diplomatically on behalf of the oil companies nor would he suspend purchases of Mexican silver, a main source of Mexican foreign exchange. This sympathetic reaction was based largely on the advice of the United States ambassador, Josephus Daniels, who as secretary of the navy had ordered the occupation of Veracruz in 1914 but who over the years had become a warm friend of the Mexican people.

Expropriation of foreign petroleum companies

Mexicans consider March 18, 1938, as the anniversary

of Mexican economic independence. Cárdenas, by taking this bold action, had not only restored an important resource to national patrimony but also showed that national honour and dignity could not be flouted by foreign entrepreneurs, however powerful. After engaging in propaganda campaigns, boycotts, blacklists, and other forms of economic warfare, the oil companies eventually were forced to settle their claims on essentially Mexican terms. Once the initial bungling was over, Pemex developed the capacity to fuel the industrial revolution that marked Mexico's next epoch.

Gen. Manuel Avila Camacho, whom Cárdenas supported, and Gen. Juan Andreu Almazán fought a close and bitter contest for the presidency in 1940. When Almazán lost, he sought United States backing for a revolution. But to emphasize the United States position toward Avila Camacho and Mexico, Roosevelt sent Vice President Henry A. Wallace to attend the inauguration. When Cárdenas stepped down from the presidency in November 1940, a major chapter of the Mexican Revolution had closed.

The Avila Camacho domestic policy was one of conciliation. While retaining earlier forms, he placated Catholic groups by announcing, "I am a believer." Under the doctrine of political unity and harmony, he brought back from exile various former presidents; with linked arms several of them, including Calles and Cárdenas, appeared on the National Palace balcony to greet a cheering crowd. Avila Camacho also mounted campaigns against illiteracy and other social deficiencies, but the overriding issues of the day were diplomatic and economic; the Cárdenas resurgence had increased Mexico's self-respect but had left its economy in a shambles.

WORLD WAR II (1941–45)

World War II brought profound changes to Mexico. Its basic economic structures were transformed, as to a lesser degree were its political, social, and cultural realms. Industrialization and urbanization were accelerated to meet wartime needs.

Industrialization and urbanization

The successes of Hitler's Fascist forces in Europe and the rising power of Japan in the Orient affected many of Mexico's national policies in 1939–40. While not yet an active belligerent, the United States saw itself as the arsenal of democracy and was gearing up its industrial plant for war. Many necessary raw materials came from Mexico.

For their mutual benefit, Mexico and the United States in November 1941 signed a General Agreement that liquidated most outstanding quarrels between the two countries. The old problem of United States agrarian claims was resolved; a reciprocal-trade treaty was outlined, and the Mexican peso was stabilized and supported to maintain a constant dollar ratio. The United States agreed to continue silver purchases at world prices and to provide long-term loans to buttress Mexico's economy. Separate agreements were reached on military aid, primarily to professionalize the Mexican army and its small air force; as a step in that direction the military sector was dropped from the official party, to eliminate the army as a bloc in politics, and officers and men were represented through the diversified popular sector.

The United States declared war on the Axis powers in December 1941. Mexico also became an active belligerent after Germany sank some of its tankers. A Mexican–North American joint defense committee planned cooperative operations to be carried out in case the Japanese attacked northern Mexico and the Pacific states. The former president Lázaro Cárdenas served on that committee and became secretary of defense when that post was created on January 23, 1944. A draft law requiring one year of military training provided trained reserves for the Mexican army, which was forced to modernize. A token Mexican air unit operated with the United States in the Philippines. But Mexico's major contribution to the war effort was to supply raw materials for the United States industrial plant. Mexico's railways had not been refurbished since the military phases of the revolution, and the frequency of wrecks highlighted the extent of deterioration under inept workers' administration. Joint planners rehabilitated the system, which remained nationalized under the direct management of the Mexican government.

With most of the free world producing war material, consumer goods were unavailable. To fill this vacuum, Mexican light industry developed, almost exclusively with Mexican capital. The new industrialists were nationalistic, patriotic, and able. They represented a new and growing interest in the old political equations.

As a recognized ally, Mexico took the lead in Latin America in preparing for the formation of the United Nations. It also supported the 1948 changes in the Organization of American States, the United Nation's regional body.

POSTWAR MEXICO (SINCE 1945)

From the end of World War II to 1970 Mexico's population about doubled; its cities continued to grow, and the nascent industrialism spawned by the war became a major element in the economy. The military increasingly faded into the background as arbiters of national policy, and Mexico had an unbroken line of civilian presidents, beginning with the election of Miguel Alemán in 1946. With him, the emphasis shifted from the Cárdenas approach of dividing Mexico's small agricultural land area among many hands to the development of new resources. Massive hydraulic projects were undertaken to furnish electric power, create high-yield irrigated lands, provide flood control, and become the nuclei of regional agricultural–industrial complexes. The nationalized oil industry changed from one aimed primarily at producing petroleum for export and processing abroad to a vast petrochemical enterprise geared to meet burgeoning domestic needs. Economic integration was accomplished by the extension of railroad, highway, and airline networks to nearly all regions of the republic. As the population increased, the domestic markets expanded.

Civilian presidents

Political changes. Postwar Mexico was marked by a continuity of basic policies unprecedented in Mexican history and by the peaceful constitutional transfer of presidential power from one civilian regime to the next. Presidents following Miguel Alemán were Adolfo Ruíz Cortines (1952–58), Adolfo López Mateos (1958–64), Gustavo Díaz Órdaz (1964–70), and Luis Echeverría Álvarez (1970–), all of whom had worked upward through party and bureaucratic ranks and most of whom had gone to the presidency from the ministry of interior. During the campaign periods before the July 1 elections, all followed the Cárdenas precedent of making extended national tours to confer with state and local leaders about their programs.

President Alemán was the major architect of new departures in the official party. Under him the name was changed from the Party of the Mexican Revolution to the Party of Institutionalized Revolution (PRI). This reflected Mexican thought that the task of the leaders was to make the institutions created by earlier revolutionaries produce the results they envisioned. Suffrage in Mexico was nearly doubled when Ruíz Cortines enfranchised women, who first voted in the national elections of 1958.

While opposition was not stifled, it simply had little to offer as an alternative to the well-oiled and responsive single party. For the same reason, the appeals of Soviet or Chinese Communism and Cuban Castroism elicited only token response. They were viewed as exotic foreign ideologies, suspect from the outset for being non-Mexican; most of their ultimate utopian goals had been stated in the 1917 Mexican constitution.

The economy. Mexico's economy combined private enterprise, state capitalism, and cooperative combinations of the two. Certain areas, such as mining, remained almost completely in the hands of the private sector, while such activities as transportation, communications, petrochemicals, and electric-power production were state owned and operated, usually by semiautonomous executive agencies. Private enterprise controlled most manufacturing, both heavy industry and consumer goods. A flexible government instrument, the Nacional Financiera, played an important role in furnishing capital for needed

Governmental and private ownership

Pres. Gustavo Díaz Ordaz of Mexico and Pres. Lyndon Johnson of the U.S. shake hands after completing legal steps necessary for relocating a section of the U.S.–Mexican border between El Paso, Texas, and Ciudad Juárez, Mexico, on October 28, 1967. Antonio Carrillo Flores (left), Mexican secretary for foreign relations, and Dean Rusk (right), U.S. secretary of state, look on.
UPI Compix

industries, either by purchasing shares in the enterprises or by directly launching the desired activity and then transferring it to private investors. A number of international corporations expanded into Mexico under the general doctrine that they would supply expertise for Mexican management and not more than half the capital. Thus, automobile-assembly plants and even mass merchandising, such as Sears, Roebuck stores, were joint foreign–Mexican ventures.

The industrialization process had not been fully completed by 1970. Among Latin American nations, however, Mexico was a close second to Brazil in gross national product and in per capita gross national product was outranked only by Argentina, Chile, Panama, and Venezuela. Comparison on an international scale with countries of comparable population, however, indicated that Mexico was still relatively underdeveloped.

Urbanization. The lure of better wages in such industrial centres as Mexico City, Puebla, Guadalajara, and Monterrey sped the growth of towns to cities and cities to regional metropolises and made the metropolitan area of Mexico City, with perhaps 4,000,000 people, among the world's largest. Rapid urbanization imposed serious strains on housing facilities, social services, and transportation systems. A subway system, inaugurated in 1969 and extended in 1970, partially alleviated the mass-transit problem in Mexico City. Industrialization and urbanization increased atmospheric pollution, which became especially severe in the Mexico City area, where surrounding mountains trapped the smog.

Sociocultural policies. National outlays for education began to grow in 1921, when Obregón created the National Ministry of Education; military expenditures were reduced as those for education rose. Obregón had to spend nearly half the federal budget to control the military, whereas López Mateos allocated only 6 percent to military needs. The major change came under Alemán, after Avila Camacho had reduced and professionalized the army; the per capita outlay in pesos of 1950 value stabilized at about 14.5. Educational outlay in the early years took a proportionately significant share of expenditures, but, as government revenues were small, the per capita outlay was also limited. The highest percentage outlay came under Cárdenas. Subsequent governments spent about 10 percent for education, but increased public revenues raised the per capita figure to nearly twice that expended for the military. Outlays for education, both percentage and per capita, began to outstrip the military under Ruíz Cortines.

Educational gains

With each new administration, gains were made in equipping the nation with sufficient classrooms and adequate teaching staffs at all levels, but by the 1970s much still remained to be done. Perhaps half the eligible elementary school children lacked facilities to attend school at all, and fewer than half of those in school went beyond the second grade; fewer than 20 percent of the eligible pupils completed the six grades of primary education. Yet, compared to 30 or 60 years earlier, the positive accomplishments were significant.

International relations. Although it outranked France in population and nearly equalled the United Kingdom in the 1970s, Mexico was still considered to be and seemingly was content to be a small world power. It faithfully discharged its duties in such multilateral organizations as the United Nations and the Organization of American States, where it pursued a course independent of the United States and where its leadership was respected.

By 1970 Mexico maintained diplomatic relations with more than 60 countries, including Cuba. Its single most important diplomatic partner was the United States, whose long, unfortified border it shared and with whom, after 1940, when an era of good feeling set in, it was able to remove all major and nearly all minor points of friction through amicable negotiation. Mutual respect has formed the basis for numerous joint United States–Mexican enterprises, especially along the north Mexican border, where irrigation projects have made the population the fastest growing in the Mexican Republic. Tourism ranked above petroleum and agriculture as a source of national revenue; most tourists come from the United States.

When Mexico played host to the Olympic Games in October 1968, persons from all over the world saw its natural and social wonders at firsthand or on television. In addition to the usual array of athletic contests, Mexico simultaneously sponsored "artistic and cultural Olympics" in which friendly rivalry rather than intense competition prevailed.

Mexico in the 1970s. In his inaugural address on December 1, 1970, Pres. Luis Echeverría indicated that his government would continue to strengthen Mexican democratic institutions, guided by the goals of the 1917 constitution. He noted that grave gaps and injustices remained despite the fact that the continuing revolution "assured individual liberty, internal peace, uninterrupted growth, and our capacity for self-determination." He predicted that Mexico's population would reach 70,000,000 by the end of the decade and 100,000,000 within

20 years. Advanced planning was imperative to provide these new Mexicans with jobs and opportunities when they needed them. He welcomed foreign capital that would complement, not displace, Mexican, and he announced, "Technological progress is today the best ally of the Mexican Revolution." Technology was expected to create a good portion of the 500,000 new jobs needed annually for the expanding labour force. The advancement of science, therefore, was set as the national goal to modernize the society and the economy.

President Echeverría repeated a theme often enunciated by his predecessors—that the "Mexican Revolution was not imported, nor have we tried to export it." He strongly affirmed that Mexico would continue to reject all interference in its internal affairs, "under whatever ideological label it may be disguised." He noted that "The Mexican people were never daunted by the grave challenges of history," and he called on his people "to reaffirm the staunchness of their principles; to live to the full the rich philosophy of the Constitution of 1917."

BIBLIOGRAPHY. ROBERT A. POTASH, "The Historiography of Mexico Since 1821," *Hispanic American Historical Review*, 40:383–424 (1960), introduces the vast scholarly and popular literature; Mexican history sections in the *Handbook of Latin American Studies* (annual), supplement and update Potash. Despite a slanted viewpoint, HENRY B. PARKES, *A History of Mexico*, 3rd ed. (1960), remains a standard single-volume summary. H.H. BANCROFT, *History of Mexico*, vol. 4–6 (*Works*, vol. 12–14, 1883–88), is a large compendium of facts, 1804–1887. Details on independence and the first empire are found in WILLIAM S. ROBERTSON, *Iturbide of Mexico* (1952). WILFRID H. CALLCOTT probed that era and the following in his *Church and State in Mexico, 1822–1857* (1926), and *Liberalism in Mexico, 1857–1929* (1931). A discussion of the evolution of the Reform appears in CHARLES A. HALE, *Mexican Liberalism in the Age of Mora, 1821–1853* (1968), while ROBERT A. POTASH, *El banco de avío de México: el fomento de la industria, 1821–1846* (1959), deals with contemporaneous economic developments. The Juárez era is covered by RALPH ROEDER, *Juárez and His Mexico*, 2 vol. (1947); and by WALTER V. SCHOLES, *Mexican Politics During the Juarez Regime, 1855–1872* (1957), both to be read in conjunction with E.C. CORTI, *Maximilian and Charlotte of Mexico*, 2 vol. (1928); the late Juárez, Porfirian, and early revolutionary years are being explored in depth in the massive volumes published by El Colegio de Mexico under the editorship of DANIEL COSIO VILLEGAS, *Historia moderna de México*, 8 vol. (1955–). Despite its title, HOWARD F. CLINE, *The United States and Mexico*, rev. ed. (1963), emphasizes Mexican events from 1900 to 1952, expanded in his *Mexico: Revolution to Evolution, 1940–1960* (1962). Achievements and gaps of the Revolution are recorded in JAMES W. WILKIE, *The Mexican Revolution: Federal Expenditure and Social Change Since 1910*, rev. ed. (1970). JAMES D. COCKCROFT, *Intellectual Precursors of the Mexican Revolution, 1900–1913* (1968), is revisionist of STANLEY R. ROSS, *Francisco I. Madero: Apostle of Mexican Democracy* (1955), a scholarly biography, and CHARLES CUMBERLAND, *Mexican Revolution: Genesis Under Madero* (1952).

(H.F.C.)

Mexico City

Mexico City (Ciudad de México), the capital of the United Mexican States, is one of the oldest metropolises in the Western Hemisphere and one of the world's fastest growing cities as well. Located in the southern part of the Valley of Mexico, it occupies an ancient plain that was once a lake and is surrounded by mountains. The city proper covers an area of 147 square miles (380 square kilometres), but the greater metropolitan area, where some 8,600,000 people lived in the early 1970s, stretches over 779 square miles (2,018 square kilometres). The centre of the powerful Aztec Empire of the 14th and 15th centuries AD, its name is derived from the Aztec words *mexliapan* ("lake of the moon") and *xitli* ("center"), which literally mean "city in the centre of the lake of the moon."

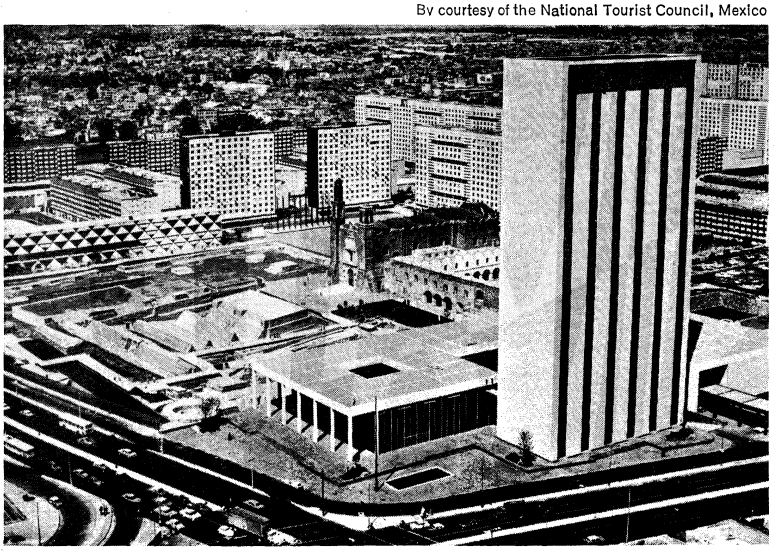
Mexico City is both the political and economic centre of Mexico and has remained a focus of Latin American culture since the 16th century. It is a sprawling megalopolis, the continuous rapid growth of which has brought in its train problems of poor housing, pollution, and social tension. At the same time, however, it is also a city of great charm that contains notable examples of colonial and modern architecture, palaces, and gardens.

History. *Origins.* The Aztec Indians arrived in the Valley of Mexico in about 1221. According to tradition, they found in 1325, on a small island in the middle of Lake Mexliapan (Lake Texcoco), the site that had been indicated by their god Huitzilopóchtli for locating a city. Known as Tenochtitlán (Place of the High Priest Tenoch), the city began to spread out over the island's marshes and swamps, and it soon became a place of fountains and gardens. It was divided into *calpullis*, or self-governed districts, each with communal lands and schools. There were also two zoos that belonged to the Great Tlatoani (emperor).

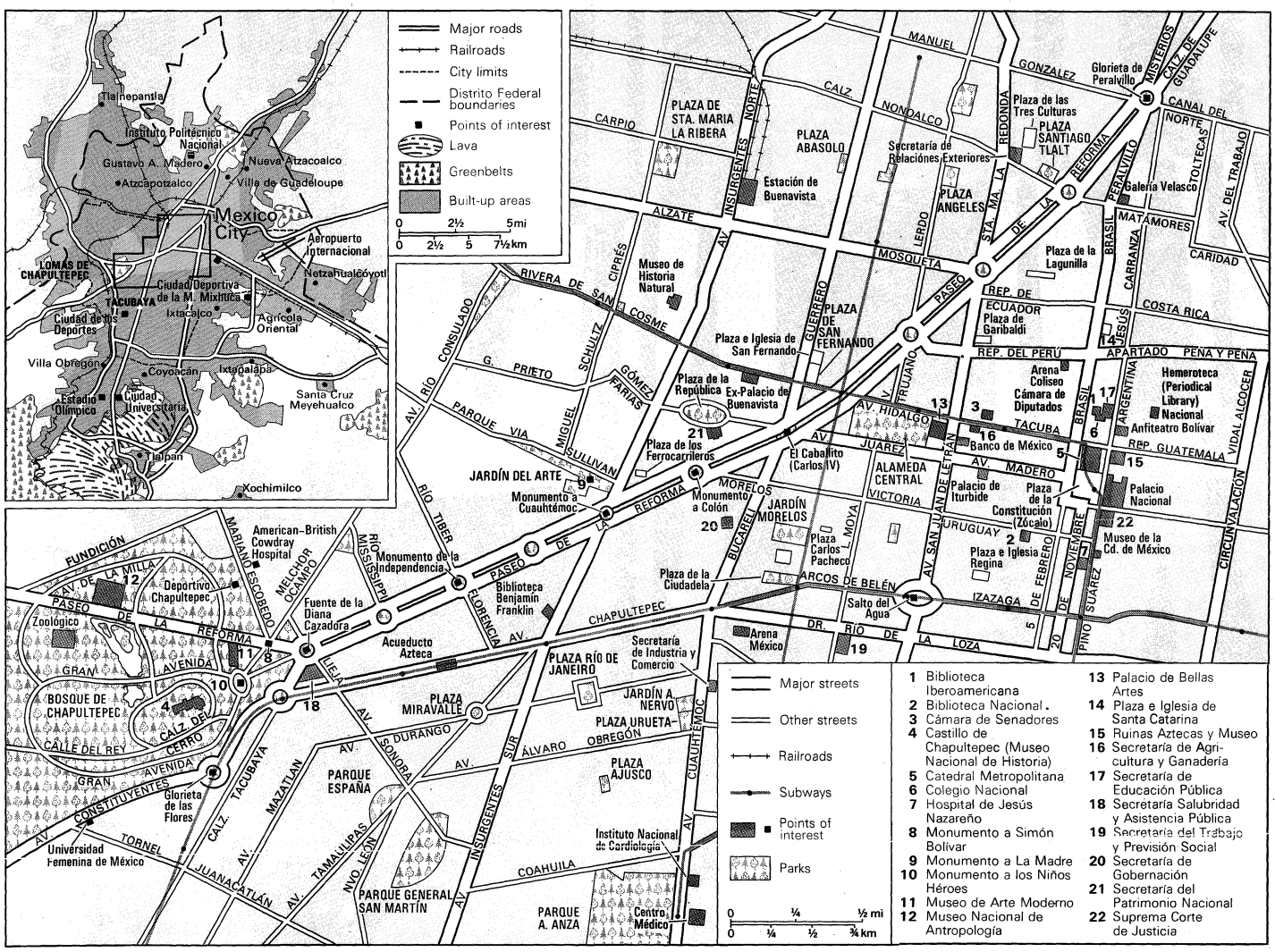
The city of Tenochtitlán

When the Spanish, led by the explorer Hernán Cortés, arrived in 1519 from the Old World, there were about 500,000 persons living in Tenochtitlán. The small islands in the lake had been united through land reclamation to form one large oval island that was connected to the mainland by three causeways—Tepeyac to the north, Ixtapalapa to the south, and Tacuba to the west. The three causeways converged on the ceremonial city centre, where the main temple and the palace were located. This centre, or plaza, now called the Zócalo, or officially

By courtesy of the National Tourist Council, Mexico



Foreign affairs building (foreground) adjoining the Plaza de las Tres Culturas (Plaza of the Three Cultures), a new commercial complex in northern Mexico City that incorporates reconstructed Aztec ruins, a 16th-century Spanish church, and modern apartment buildings.



Central Mexico City and (inset) its metropolitan area.

Plaza de la Constitución, remains the hub of Mexico City's life.

The viceregal city. Hernán Cortés, the conqueror of the Aztecs, razed Tenochtitlán and destroyed every element capable of reminding the Aztecs of their former institutions. To take advantage of the city's prestige as a former capital, however, he constructed a Spanish city on its ruins. He divided the central area among the conquerors and left the periphery to the defeated. The Aztecs intermingled with the Spaniards, making Mexico City the first mestizo city in history.

Mexico City received many privileges from the Spanish crown and in the 16th century became the most important capital in the Americas. The city was chartered, and its town council, or *cabildo*, was recognized in 1522; it was given leadership over the other *cabildos* of New Spain in 1535. Governed by colonial viceroys, it became the capital of Spanish America as far south as Panama.

Mexico City continued as a lakeside centre until constant floods necessitated the gradual filling of the lakes of the valley floor. During the 17th century it was essentially a well laid out assemblage of homes, public buildings, churches, and convents. In their execution of European designs, the Indians and the colonists created a Baroque architecture (*i.e.*, characterized by elaborate ornamentation) that employed *tezontle* (a light and porous volcanic rock) for facades. They used quarried stone for lintels. The 18th century was the golden age of architecture in New Spain, and after the Baroque manner reached its ultimate expression, Neoclassical ideas were introduced by Manuel Tolsá.

From independence to the revolution. After the fall of Emperor Agustín de Iturbide of Mexico, in 1824, the republican constitution created the Distrito Federal (Federal District) to centralize the new government in Mexico City. Invasions by the United States in the 1840s and by France in the 1860s as well as internal insurrections prevented a program of urban renewal. During the anticlerical reform movement of the 1850s, all the city's convents were either demolished or put to other uses. The city limits were expanded in 1865, when the Habsburg Archduke Maximilian—then emperor of Mexico—built the Paseo del Emperador (now Paseo de la Reforma) to connect his palace with the Castillo de Chapultepec (Castle of Chapultepec). At about this time the first *colonias*, or barrios (residential districts), appeared on the city's fringes to house the expanding population. During the rule of Porfirio Díaz, from 1876 to 1911, Mexico City was modernized on the lines of Baron Haussmann's Paris. European architectural styles were adopted in such buildings as the post office and the Palacio de Bellas Artes (Palace of Fine Arts). The drainage system was completed in 1907, and gas and electric lighting and streetcars were introduced.

The first urban expansion

The 20th century. During the civil wars of the Mexican Revolution, which lasted from 1910 to 1917, the city became a battlefield, for control of the nation without its capital was virtually meaningless, and disruption in rural areas resulted in an exodus from the countryside to the capital. In 1924, urban works were renewed with the construction of the 21-mile-long Avenida Insurgentes; French palaces were replaced with buildings designed in the style popular in the United States. In post-revolutionary Mexico City, however, luxurious districts, such as Lomas de Chapultepec (Chapultepec Hills), arose in sharp contrast to the working class *colonias* that were

inhabited by emigrant peasants. The first skyscrapers appeared in the 1930s and in the 1940s industrialization began to transform the city; large urban complexes were formed, and highways and a subway were constructed. The problem constituted by the belt of poverty that surrounded the city also grew more acute.

Rapid growth has continued, checked only by the encircling mountains. The metropolitan area spread beyond the original political limits of the Distrito Federal into the neighbouring State of Mexico. In effect, Mexico City in the 1970s was no longer a single city but an agglomeration of different and sometimes contrasting cities.

The contemporary city. *The city site.* Mexico City lies at an altitude of 7,350 feet (2,240 metres). It is walled in by the mountains of the Sierra de Pachuca, to the north; the Serranía del Ajusco, to the south; the Sierra de las Cruces, to the west; and the Sierra Nevada, with its two extinct volcanoes, Iztaccíhuatl (17,340 feet; 5,286 metres) and Popocatépetl (17,887 feet; 5,452 metres), to the east. In 1971 the city proper occupied 147 square miles. The 779-square-mile greater metropolitan area included all of the Distrito Federal—except for the rural area of Milpa Alta and its ten municipalities—and ten municipalities in the State of Mexico.

Although urban continuity is unbroken, it may broadly be said that the city proper extends to Tlatelolco—originally a twin city of Tenochtitlán—on the north, Avenida Río Churubusco on the south, Bosque de Chapultepec (Chapultepec Park) on the west, and the area of San Lázaro on the east. The present-day suburbs were once either isolated villages—such as Tlalpan, Coyoacán, and San Angel—or else cattle ranches and farms. The central section is located on a lacustrine plain the watery subsoil of which has caused the city to sink more than 20 feet since 1900.

The environment. The tropical, high-altitude climate lacks a winter season. The weather is cool and dry, except in the rainy season—from May through September. The annual median temperature is 64° F (18° C). The warmer months are July and August, when temperatures average about 70° F (21° C). Night frosts occur in December, and sudden temperature changes are usual.

Vegetation and animal life

The main agricultural product of the area is alfalfa, which is used as feed for livestock. Maize and beans are the staples of the human diet. A characteristic tree of the area is the *ahuehuete* (Montezuma cypress), which lives for as long as 2,000 years. City flowers include the *zempoazúchitl*, or flower of death (whose yellow colour was the colour of death for the Aztecs), and the poinsettia. The *charal* (a small edible fish), the *acocil* (a freshwater crustacean), and the *ajolote* (a vertebrate amphibian) are typical inhabitants of the valley's lakes.

Air pollution is a serious problem. The city's 800,000 vehicles, 4,500 industrial plants, and one oil refinery emit tons of contaminants daily. Deforestation and the drying up of the surrounding lakes' beds aggravate the problem of pollution by producing tons of dust laden with saltpetre (sodium nitrate).

The city plan and traditional neighbourhoods. At the centre of the original city of Tenochtitlán, the Spaniards created a planned town, placing the main public buildings around a square and utilizing former Indian sites. The great cathedral stands very nearly on the spot of the main temple of the Aztecs, and the viceroy's palace (now Palacio Nacional) was substituted for the palace of Montezuma. The conquered Aztec population was confined to areas in the north and east. The city gradually broadened and enveloped various new nuclei (*colonias*) from 1862 onward. Until Mexico City reached a population of more than 1,000,000, in 1930, its growth and development was haphazard and unplanned. Suppression of numerous municipalities and annexation of others in that year set the stage for a series of planning commissions to reorganize and reform the metropolis. Between 1930 and 1950, the city more than doubled its population.

Since the city was designed by Cortés in the 16th century, there has been no plan to regulate its growth. Although the area is overwhelmingly urban, there are places in the south that are agricultural or forested. The historic hub of the city is centered around the Plaza de la Constitución, where the Palacio Nacional (the original home of Cortés and the viceregal residence), the seat of the federal government, the Catedral Metropolitana, and the city hall are located.

Despite frantic destruction of the old parts of the city, plazas such as that of Santo Domingo preserve their original atmosphere. Numerous buildings of great artistic and historic value include the palaces of Iturbide, Buenavista, and Chapultepec, where more than 100 military rebellions have occurred.

In stark contrast to sites such as Netzahualcóyotl, where 90,000 persons live in conditions of extreme poverty, the area of Lomas de Chapultepec contains luxurious homes overlooking ravines. Huge mansions are integrated into the landscape of lava, rock, flowers, and water in the Jardines del Pedregal de San Angel. Xochimilco evokes a semblance of Tenochtitlán, while San Angel and Coyoacán resemble colonial Mexico.

Transportation. Mexico City is the hub of all national transportation lines. Four great highways lead from the capital to the United States; one connects the city to the Guatemalan frontier; four run to Atlantic ports; and five connect it with the Pacific coast. The national railways link the city with all of these points except Acapulco. The principal contact with other countries is by air, and more than 80 percent of the flights registered annually at the Aeropuerto Internacional are international.

Although foreign transport has improved markedly, overpopulation and excessive expansion have hampered the development of the domestic transport system. Traffic congestion is the rule in Mexico City; bus and taxi service is inadequate, and the bus lines are often symbolic targets of student protest. There is, however, a cheap service of collective taxis, known as *peseros*, along established routes, and the Metro, or subway, is highly efficient.

Population. In 1970 there were more than 2,907,000 persons living in the city proper and almost 8,600,000 in the greater metropolitan area. It is estimated that in addition to natural growth, 400,000 migrants from rural areas will become new residents each year. With a population density of 19,775 persons per square mile (7,650 persons per square kilometre), the greater metropolitan area of Mexico City holds almost 15 percent of the total population of the country; 62 percent are under 25 years of age and only 10 percent are over 50.

Most of the middle and high income groups live in newly developed suburbs, such as Ciudad Satélite (Satellite City), while low income families are segregated in the 450 slum districts called *ciudades perdidas*, or lost cities, in which many live in tin-roofed huts, and in the 5,000 *vecindades*, or neighbourhoods, of the capital.

Ethnically, the large majority of the population is mestizo (of mixed European and American Indian descent), professes Roman Catholicism, and lives in the poorer northern section. The middle income group is generally criollo (the term for Mexicans of European descent) or European. Its members live in the southern and western sections, speak English as a second tongue, and often send their children to European and U.S. schools.

The acute housing shortage

Housing. One of the most distressing problems facing Mexico City in the early 1970s was housing. All government efforts to resolve this problem had failed because of the population explosion and the high rate of internal migration. Hence, the Labour Law of 1970 required all companies that employ more than 100 workers to build housing for their employees. In 1970, 13,000 low-cost housing units were completed under that law, but there were 40,000 families waiting to occupy them. To contain the flood of migrants to the capital, the new city of Cuatitlán Izcalli was under construction north of the city in the early 1970s to house 1,600,000 persons. In addition, there was a continual program of rehabilitation of the *ciudades perdidas* and the *vecindades*.

Building types. Buildings include the cardboard and sheet metal huts of the "lost cities" and the deluxe mansions and modern condominiums of the well-to-do. A fascinating aspect is the contrast of old and new, of Spanish, Indian, and modern, and of homes with Moor-

ish-style interior patios and modern functional architecture.

The main architectural interest lies in the colonial buildings that largely date from the 18th century and that still stand in the city centre, as well as in San Angel and Coyoacán. These include the Catedral Metropolitana, the largest of the continent (begun in 1570; completed *c.* 1810), and other buildings around the Plaza de la Constitución, such as the municipal palace, or city hall, and the Palacio Nacional.

Economic life. The greater part of the nation's economic activity is concentrated in the capital, which generates 50 percent of all industrial production. It also accounts for 59 percent of the wages and 60 percent of the non-agricultural work force. Industry has located 31 percent of its plants in the city, where 40 percent of the gross national product is created.

Industry

The main industries deal with construction and the production of iron and steel, chemicals, plastics, cement, and yarns and textiles. More than 1,000,000 tourists are accommodated each year; 60 percent of the nation's commercial transactions are conducted in the capital; and there are more than 40 banking institutions (with 350 branch offices) in the metropolitan area. Finance, mortgage, and insurance companies have their home offices in Mexico City, which is also the seat of the nation's stock exchange. The Banco de México (Bank of Mexico) is the only institution that issues currency. Basically, the municipal budget depends on income, property, excise, and other taxes and tourism.

Political and governmental institutions. All the political parties have their headquarters in Mexico City, which is also the site of the Mexican Chamber of Deputies, the Senate, the Supreme Court, and the federal courts. The president of the republic administers the Distrito Federal through a regent (*jefe del departamento central*), who is appointed by him when the cabinet is named.

Public utilities. Water supply has been a constant problem for Mexico City. A Central Drainage Canal (El Gran Canal del Desagüe), which was built to move sewage waters, is inadequate; hence, three interceptor canals that will run the length of the city at depths of up to 720 feet were being built in the 1970s. Also, because local springs have run dry, drinking water must be supplied from increasingly remote areas.

The electric company (Compañía de Luz) operates 12 generating plants powered by hydraulic turbines and has a capacity of more than 647,000 kilowatts.

Health and safety. Mexico City has the oldest hospital in the Western Hemisphere—the Hospital de Jesús Nazareño, which was founded by Cortés in the early 16th century. Among the government-run institutions are the ISSSTE (Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado), which runs hospitals for government employees, in addition to social security hospitals, and the Centro Médico (Medical Centre), a conglomerate of specialized units. There are several private hospitals, including the American–British Cowdray Hospital. Although all of the barrios are serviced by general and maternity clinics and dispensaries, there is a shortage of available hospital beds.

The fire department, established in 1871, is outfitted with modern equipment and has demonstrated high levels of efficiency. The federal and city police number 20,000, an inadequate figure for a city of such dimensions.

Education. Although the government devotes 20 percent of its budget to education, the demand for buildings and personnel is still not met. In the early 1970s more than 90 percent of primary school age children attended classes; many, however, leave school at the secondary and university levels. There were more than 300 kindergartens, 2,000 primary schools, and almost 500 secondary schools in the city. The Universidad Nacional Autónoma de México (National Autonomous University of Mexico) is located south of the city on a campus known as Ciudad Universitaria, or University City. Its buildings are beautifully decorated with murals by the painter Diego Rivera and others. The enrollment in the early 1970s was over 100,000. The Instituto Politécnico Nacional (National Polytechnic Institute) has an enrollment of about 80,000 students. There are also several technical schools and private universities in the city. Classes are conducted in Spanish, but some institutions offer instruction in English during the summer sessions.

University City

Cultural life. The city retains its historic position as one of the greatest centres of Latin American culture. In addition to the research institutes of the National University, the most outstanding institutions of advanced learning are the Colegio de México and the Colegio Nacional. There are about 35 museums, including the magnificent Museo Nacional de Antropología, which opened in 1964 and contains collections of almost all the ancient cultures of Mexico.

The public libraries are poor, tradition being more in favour of vast private libraries. There are more than 50 active publishing houses and approximately 1,000 bookstores.

The media. Twenty newspapers are published daily, including *Excelsior*, considered one of the most important newspapers in the Spanish-speaking world. Several of the papers have nationwide circulation and are important factors in the formation of public opinion. There are numerous informative, literary, and other specialized magazines. Of Mexico's seven television channels, one is devoted exclusively to cultural broadcasts, and there are about 60 radio stations.

Recreation. Each neighbourhood has its own parks and gardens. Bosque de Chapultepec, the largest and most beautiful of these, encompasses almost 1,000 acres in the southwest. Its grounds contain museums, zoos, botanical gardens, an amusement park, lakes, and fountains. El Desierto de los Leones (Desert of the Lions), another park ten miles southwest of the city, occupies 4,695 acres of mountains, pine forests, springs, and aqueducts. The Alameda Central (Central Mall), west of the Zócalo, was created in 1592.

The most popular spectator sports are football (soccer) and bullfights. Soccer is played in the Azteca Stadium, in the stadium located in the Plaza Ciudad de los Deportes (Sports City), and in the Olympic Stadium of University City. There are two bull rings—México, which is the world's largest, and El Toreo. In addition, there are 50 sports centres and numerous swimming pools, as well as other facilities that were built to house the Olympic Games in 1968. The Ciudad Deportiva de la Magdalena Mixhuca (Magdalena Mixhuca Sports City) covers 565 acres and includes sports fields and a three-mile auto racing track. There are numerous cinemas and legitimate theatres, and free concerts and exhibits are held in the public parks.

BIBLIOGRAPHY. The best works on Mexico City, in English, are two minor 19th-century classics: FRANCES CALDERON DE LA BARCA, *Life in Mexico During a Residence of Two Years in That Country* (1843); and BRANTZ MAYER, *Mexico As It Was And As It Is* (1844). Numerous collections of photographs with brief comments include BOB SCHALKWIJK and J.M. COHEN, *Mexico City, México* (1967). SALVADOR NOVO, *Nueva Grandeza Mexicana* (1946; Eng. trans., *New Mexican Grandeur*, 1967), describes city life in the 1940s. The Spanish bibliography is, of course, extensive, although works on contemporary Mexico are scarce. The best summary is SALVADOR NOVO, *México* (1968). Historical works include HERNANDO CORTES, *Cartas de Relación* (Eng. trans. by FRANCIS A. MACNUTT, *Letters of Cortes*, 2nd ed., 2 vol., 1932); and BERNAL DIAZ DEL CASTILLO, *La Verdadera historia de la conquista de la Nueva España* (1632), on Tenochtitlán. On the colonial city, see FRANCISCO CERVANTES DE SALAZAR, *México en 1554* (1875; 3rd ed., 1964); and BERNARDO DE BALBUENA, *Grandeza Mexicana* (written 1604, first published 1829, reprinted 1941). On Mexico from the 17th to the 19th century, see LUIS GONZALEZ OBREGON, *México viejo*, new ed. (1900); JOSE M. MARROQUI, *La Ciudad de México*, 3 vol. (1901–03, reprinted 1969); and JESUS GALINDO Y VILLA, *Historia sumaria de la ciudad de México* (1925). The magazine *Artes de México* has published three volumes under the title *La ciudad de México* (1964), which have fine photographic illustrations. Regarding serious contemporary problems, the two volumes of *Memoria de la reunión para el estudio de los problemas del D.F.* (1971); and LUIS UNIKEL, *La dinámica del crecimiento de la ciudad de México* (1972), are indispensable.

(F.Be.)

Micas

The micas are an important group of common rock-forming minerals characterized by their highly perfect cleavage, or tendency to split along planes determined by crystal structure, yielding thin laminae. The name is from the Latin *mica* ("grain"). The principal members of the group are muscovite, biotite, and lepidolite. These are aluminosilicates with potassium, magnesium, iron, or lithium, and hydroxyl and fluorine. The micas as a group are highly variable in chemical composition, physical and optical properties, and mode of occurrence, but all are based on a relatively simple crystal-structure unit.

Importance and uses

Because some micas are poor electrical conductors, they are used in the construction of electrical equipment. Some micas are also poor heat conductors, transparent, and relatively resistant to water corrosion; these have been useful as stove windows and boiler gauge packings. Ground mica is commonly employed as a filler, lubricant, absorbant, dusting powder, and packing material. It is used in the manufacture of wallpaper, roofing paper, and paints. It is employed for special effects by decorators and toy manufacturers—for example, as "frosting" or Christmas tree "snow." Sheets of mica are used to make plates of precise thickness for use in the analysis of polarized light. Lepidolite is an important source of lithium and rubidium salts. Some of the synthetic micas are used as electrical insulators and in special delicate electronic instruments.

The common micas are rich in potassium and also contain some strontium and rubidium, which are essential elements involved in the age determination of minerals and rocks. The abundance of the elements involved in the decay of the isotope (isotopes are elements with the same atomic number but different atomic weights) potassium-40 to argon-40, and of rubidium-87 to strontium-87, in a mica can be measured with precision and the absolute age of the mineral calculated in millions of years. Metamorphism (alteration by heat and pressure) of the micas often affects the relative abundance of these elements, so that the micas usually record the last geological event or episode of metamorphism and not the date of their initial growth in the Earth (see further DATING, RELATIVE AND ABSOLUTE).

Occurrence of micas and their mining

The micas occur in all rock types—igneous (formed by the cooling and crystallization of molten silicate material), metamorphic, and sedimentary. They are absent in rocks so far recovered from the lunar surface, however. Members of the group are found under the entire range of pressures and temperatures that reflect geological conditions: thus, micas occur in lavas, schists, granites, marbles, soils and deep-sea sediments, mineralized veins, pegmatites (very coarse-grained igneous rocks), alteration zones, and in products of weathering of other minerals. Each particular mica type is limited in occurrence to a certain range of variables (such as composition, temperature, and pressure) of the geological environment, and some micas are found only in specific rock types. A characteristic occurrence is in granite (see Figure 1), in which biotite and muscovite usually form. Phlogopite is common both in ultrabasic rocks (those poor in silica) such as dunites and peridotites and in dolomitic marbles. Fine-grained white micas, or sericite, particularly muscovite, form in the early stages of metamorphism of sediments and are closely associated with biotite. During retrograde metamorphism (the occurrence of one episode of intense metamorphism followed by subsequent, less intense episodes) the aluminous silicates are altered to sericite; these silicate minerals include andalusite, kyanite, sillimanite, cordierite, and staurolite, as well as the feldspars. The illites, a group of clay minerals, are partly of mica and are a common constituent of soils and deep-sea sediments. Glauconite, another type of clay mineral, also occurs in marine deposits. Celadonite is an alteration product of some igneous rocks. Lepidolite is mainly found in veins and pegmatites, often in parallel growths with muscovite.

The mining of mica is difficult because the mica books (crystalline blocks that cleave or split readily into thin mica sheets) are readily damaged by rough handling. Methods of extraction are therefore primitive and necessarily wasteful; the veins and pegmatites are usually worked by quarrying, but underground excavation may be economical. The books are split and trimmed by hand; sorted according to size, transparency, colour, and freedom from spots or strains; and the scrap is ground. The production of mica sheets seldom exceeds 10 percent of the total mine output. Mining is conducted primarily in the United States (North Carolina, Idaho, South Dakota), India, the Republic of South Africa, the Soviet Union, Brazil, Argentina, Canada, and the Malagasy Republic.

This article treats the mineralogical properties, crystal structure, and chemical composition of the micas. It also includes discussion of the synthesis of micas and their stability limits in nature. For information on closely related mineral types see CLAY MINERALS and SILICATE MINERALS, and for further detail on the occurrence of micas in nature see IGNEOUS ROCKS; METAMORPHIC ROCKS; and SEDIMENTARY ROCKS. See also SHALES; MARINE SEDIMENTS; and GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES for related information.

PROPERTIES OF MICAS

Mineralogical characteristics. The different kinds of mica (see Table 1 for principal members) vary greatly in colour and are frequently zoned (*i.e.*, there are concentric colour bands). Muscovite is commonly colourless and may be gray, brown, light green, lavender, or ruby red. The biotites are commonly black, brown, green, yellow, and occasionally are colourless. High ferric iron content tends to make the crystals green, whereas high titanium oxide content gives a reddish brown colour. Intermediate proportions of titanium and ferric iron tend to produce colours of yellowish or greenish brown. Pink, violet, yellow, gray, red, brown, and green are the colours displayed by the lepidolites. Paragonite is colourless. The luster varies from pearly, as in muscovite, to submetallic, as in lepidomelane.

Table 1: Principal End-Members of the Mica Group*

	X	Y	Z	O_{10}	$(OH, F)_2$
Dioctahedral					
Muscovite	K	Al_2	$AlSi_3$	O_{10}	$(OH)_2$
Ferrimuscovite	K	Fe_2^{+3}	$AlSi_3$	O_{10}	$(OH)_2$
Paragonite	Na	Al_2	$AlSi_3$	O_{10}	$(OH)_2$
Margarite	Ca	Al_2	Al_2Si_2	O_{10}	$(OH)_2$
Celadonite	K	$MgFe^{+3}$	Si_4	O_{10}	$(OH)_2$
Al-celadonite (leucophyllite)	K	MgAl	Si_4	O_{10}	$(OH)_2$
Trioctahedral					
Phlogopite	K	Mg_3	$AlSi_3$	O_{10}	$(OH)_2$
Na-phlogopite	Na	Mg_3	$AlSi_3$	O_{10}	$(OH)_2$
Annite	K	Fe^{+2}	$AlSi_3$	O_{10}	$(OH)_2$
Clintonite	Ca	Mg_3	Al_2Si_2	O_{10}	$(OH)_2$
Eastonite	K_2	Mg_5Al	Al_3Si_5	O_{20}	$(OH)_4$
Iron-eastonite	K_2	$Fe_5^{+2}Al$	Al_3Si_5	O_{20}	$(OH)_4$
Polylithionite	K	Li_2Al	Si_4	O_{10}	F_2
Trilithionite	K	$Li_{1.5}Al_{1.5}$	$AlSi_3$	O_{10}	F_2
Intermediate					
Unnamed synthetic	K	$Mg_{2.5}$	Si_4	O_{10}	$(OH)_2$
Siderophyllite	K	$Fe_{1.66}^{+2}Al$	$Al_{1.33}Si_{2.67}$	O_{10}	$(OH)_2$
Lepidomelane	K	$Fe_{1.66}Fe^{+3}$	$Al_{1.33}Si_{2.67}$	O_{10}	$(OH)_2$

*General formula is: $XY_{2-3}Z_4O_{10}(OH, F)_2$.

Most occurrences of the micas yield fine-grained crystals; the largest crystals are found in veins and pegmatites. The largest book of mica discovered was 4.3 metres (14 feet) in diameter and 10 metres (30 feet) long. Paragonite is especially fine-grained, occurring in massive scaly aggregates. Where colourless micas occur in two size ranges, the smaller size range can usually, but not always, be attributed to paragonite and the larger to muscovite.

Form and cleavage

When well formed, the crystals are usually tabular with hexagonal or pseudo-hexagonal outline; hence, six-sided prisms are common. The perfect basal cleavage yields thin laminae that are usually elastic. Some varieties, margarite for example, have brittle laminae. The perfect bas-

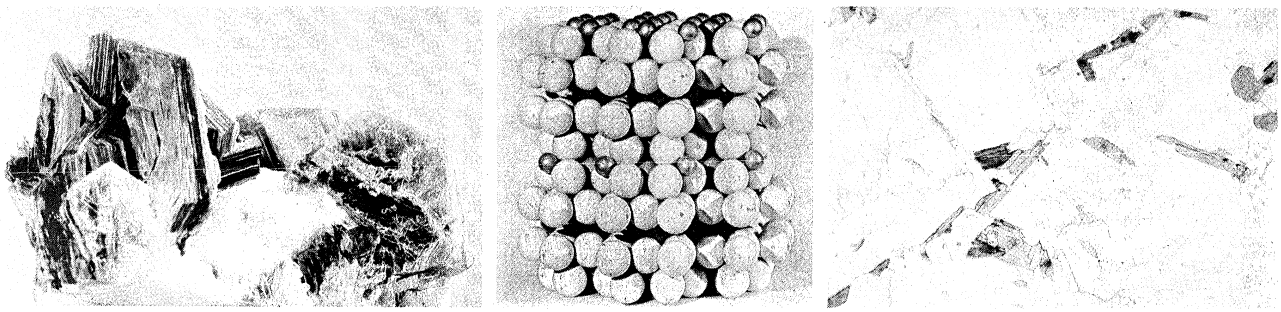


Figure 1: (Left) Single crystals of muscovite on feldspar. (Centre) Packing model of muscovite. (Right) Thin section of two-mica granite.
(Left) Joe J. Holdaway, (centre) B. M. Shaub, (right) Hatten S. Yoder, Jr.

al cleavage (perpendicular to the sides of the prism form) reflects the layered character of the internal crystal structure. The cleavage laminae are not equal in thickness throughout but may exhibit a series of steps. Each step is exactly equal to, or a multiple of, a crystal structure unit. These steps are also observed in the spiral growth of mica crystals.

When a cleavage flake is struck a sharp blow with a blunt needle point, a six-rayed star of cracks, or percussion figure, develops. A similar set of cracks develops 30 degrees from the percussion set when a blunt punch is pressed against crystal; these cracks, the pressure figure, coincide with secondary parting planes in the crystal. Such parting planes are inclined to the basal cleavage and often are mistaken for true crystal faces. The percussion figure is a useful aid in determining the orientation of the crystal symmetry in the absence of faces. Micas in the vicinity of a shear zone (zone of deformation in the Earth's crust) show kink banding. Kink bands are narrow, sharply bounded by plane surfaces inclined to the cleavage, and are parallel or subparallel. They indicate strong local deformation by translation gliding (slippage of crystal structure) whose precise nature is as yet undefined. Kink bands are displayed particularly well in micas in the sheared schists and cataclastic rocks (greatly deformed metamorphic rocks) of Scotland and Austria, and have been noted in mica-bearing rocks shocked by meteorite impact. Kink bands have been produced in laboratory studies of natural mica.

The hardness of the micas is anisotropic (that is, it varies with direction in the crystal) and ranges from 2.5 to 4.5 on the Mohs scale of mineralogical hardness (talc has a hardness of 1.0 and diamond a hardness of 10). The specific gravity ranges from 2.8 to 3.4, and the indexes of refraction (measure of light transmission) from 1.535 to 1.705. Bi-refringence (double refraction) is strong in the micas, meaning that the speed of light transmission varies greatly with crystal direction. The dark-coloured micas exhibit strong pleochroism; that is, different colours are exhibited in different directions. Inclusions are common in the micas and sometimes produce pleochroic halos (concentric colour bands). Dark spots in micas are often a result of radiation damage from included zircons, which contain radioactive chemical elements. The micas often enclose flattened crystals of garnet, tourmaline, quartz, and magnetite in thin plates between the sheets. Minute acicular (needlelike) inclusions arranged parallel to the rays of the percussion figure, particularly in phlogopite, give rise to asterism similar to that observed in rubies and sapphires.

Alteration and weathering

Muscovite and clintonite are relatively insoluble in acids, whereas biotite, lepidolite, and margarite are attacked by acids, leaving a residue of silica. Most micas yield water on ignition. The structure of muscovite can be completely destroyed by prolonged grinding in the laboratory and appears to be similarly affected to various degrees when the mineral is transported in streams and rivers. Alteration of muscovite usually results in a loss of alkalies and hydration, producing a mixed-layer mineral with montmorillonite (a clay mineral). Strong alteration may produce the clay minerals (*q.v.*) kaolinite or gibbsite with quartz. The farmer tends to restore most of the lost alkalies in the soil micas by using potash fertilizer; that potash is unavailable to the crop because it is taken up by the micas. The exchange of alkalies in micas is reversible, but some slight irreversible structural changes appear to occur. Muscovite, for example, can be converted to paragonite in solutions rich in sodium chloride, and paragonite can be converted to muscovite in solutions containing potassium chloride. The biotites on alteration are converted to chlorites. Potash feldspar, rutile, hematite, magnetite, or serpentine may also be produced. Weathering of biotite usually changes it to a soft brown or yellow mineral with a bronze luster, vermiculite.

Crystal structure. The fundamental unit of structure of the micas (see Figure 2) has been deduced by X-ray analysis of single crystals. It consists of two sheets of oxygen tetrahedra containing silicon and, usually, some aluminum. Each sheet consists of tetrahedra linked at each of three corners to form a hexagonal or pseudo-hexagonal array with all the bases parallel and all the vertexes pointing the same way. The two sheets are parallel and have opposing vertexes. These sheets are held together by magnesium or other ions (atoms that have gained or lost electrons and are thus charged), which are coordinated

From R.E. Grim, *Clay Mineralogy* (1953); McGraw-Hill Book Co., Inc.

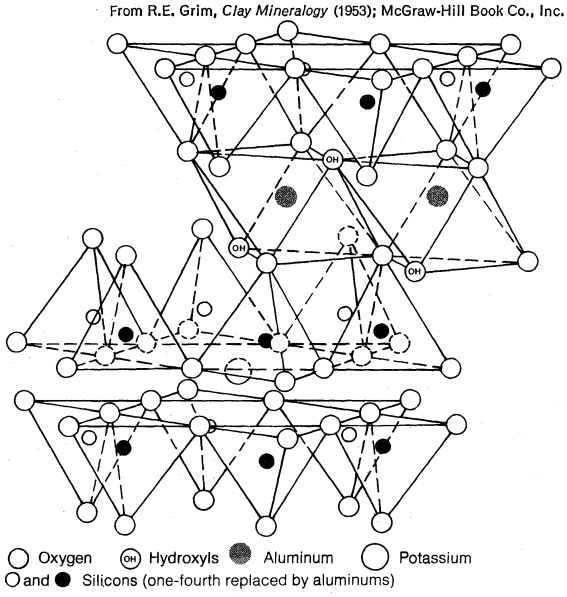


Figure 2: Structure of 1M muscovite (see text).

with two apical oxygens and one hydroxyl from one sheet and two apical oxygens and one hydroxyl from the opposite sheet, forming an octahedron. The hydroxyls occupy the centres of the hexagonal or pseudo-hexagonal arrays of apical oxygens. The structure is a succession of such double sheets with alkali atoms (*e.g.*, potassium) between them, the alkali atoms located between the hexagonal or pseudo-hexagonal arrays formed by the basal oxygens of the adjoining double sheets. The alkali atoms are thereby in twelvefold coordination, occupying the centre of a hexagonal prism formed by six basal oxygens from each of the adjoining sheets. The general structure may be

represented by the formula $XY_{2-3}\ Z_4O_{10}\ (OH,F)_2$, in which X in twelvefold coordination may be potassium, sodium, calcium, or barium; Y in sixfold coordination may be magnesium, ferric or ferrous iron, aluminum, lithium, manganese, and, rarely, chromium, vanadium, and titanium; and Z in fourfold coordination may be silicon and aluminum (see Table 1).

Polymorphism in micas

Minerals having the same chemical composition and different structures are called polymorphs. The simple structural unit described above has monoclinic symmetry; that is, it is related to three unequal crystal axes, two of which intersect obliquely and the third perpendicularly to the others. If successive units are stacked on top of each other with precisely the same orientation, the repeating unit of the mica is described as having a one-layer monoclinic structure (1M). If successive units are stacked with different orientations in a regular way, different structures (polymorphs) result. The symmetry of the surface of the unit restricts the number of ways in which succesive units may be stacked. The structures resulting from the various types of stacking may be described by the number of layers involved and the symmetry produced. The following structures have been observed in addition to the 1M: two-layer monoclinic, type one ($2M_1$); two-layer monoclinic, type two ($2M_2$); and two-layer orthorhombic (2O). Crystal symmetry in the orthorhombic system is related to three unequal and mutually perpendicular crystal axes: three-layer trigonal (3T)—a division of the hexagonal system in which crystal symmetry is related to three equal crystal axes that intersect at angles of 120 degrees and a fourth unequal axis that is perpendicular to these three; eight-layer triclinic (8TC)—crystal symmetry in the triclinic system is related to three unequal and mutually oblique crystal axes; twelve-layer monoclinic (12M); and eighteen-layer monoclinic (18M). Other more complex structures exist. A structure in which the units are randomly stacked also has been found. In addition, a six-layer hexagonal (6H) structure has been predicted but not found in nature. It should be emphasized that the structure of a ($2M_1$) mica, for example, is not exactly equivalent to two (1M) layers properly stacked. The internal strains and resulting distortions of the polyhedra produced by the various types of stacking yield new structures that, although very closely related, are not merely multiples of the simple structural unit as in polytypism. A single hand specimen may contain several polymorphs, but the biotites usually are (1M), the muscovites ($2M_1$), and the lepidolites ($2M_2$). Randomly stacked micas and mixed-layers with other layered minerals, such as montmorillonite, are common in sediments.

Chemical composition. The various species of the mica group may be described for the most part as consisting of variable proportions of ideal end-members (distinctive mineral species, between which gradual and continuous chemical variation occurs). Actual compositions are complex because of chemical substitutions. The group may be divided into those micas having two atoms in octahedral coordination (Y position), called the di-octahedral micas; those with three atoms in octahedral coordination, the tri-octahedral micas; and those containing between two and three atoms in octahedral coordination, the intermediate micas. The ideal end-members are listed in Table 1: the various atoms are arranged according to their structural position. Broadly speaking, the white micas, commonly referred to as sericite when fine grained, consist of muscovite, paragonite, or both. These form a limited series of solid solutions (a solid solution is a single crystalline phase that varies in composition between finite limits). Common glauconite consists of about two-thirds celadonite and one-third muscovite. It is not known whether glauconite is a true compound or a random mixed-layer structure, a structure consisting of discrete structural units of both muscovite and celadonite. Illite, a field designation for some sedimentary micas, is a mixed-layer mineral usually consisting of muscovite and montmorillonite. Common biotite is widely variable in composition and incorporates in solid solution various proportions of phlogopite, annite, eastonite, iron–eastonite, siderophyllite, and lepidomelane. The biotites do not form solid solutions with muscovite. Lepidolite consists mainly of polylithionite and trilithionite and forms limited solid solutions with muscovite and paragonite. There is no known complete series of solid solutions between a di-octahedral and a tri-octahedral mica. The rare micas include: a vanadium-rich mica (roscoelite), a chromium-rich mica (fuchsite), a manganese-bearing mica (manganophyllite), an aluminum-free mica (taeniolite), and a lithium–iron mica (zinnwaldite). Some mica samples yield more water on chemical analysis than that indicated by their formula. These are erroneously referred to as hydromica; in fact, they are mixtures or mixed-layered minerals of mica and a more hydrous mineral such as montmorillonite. A true hydromica, in which the hydronium ion, H_3O^+, replaces potassium, has not as yet been demonstrated in natural micas. The micas in some localities contain other volatile constituents such as ammonia and chloride in addition to hydroxl and fluoride.

Rare micas and hydromica

EXPERIMENTAL STUDY OF MICAS

Synthesis. Micas very closely approaching the properties of those found in nature have been synthesized in the laboratory; these include phlogopite, muscovite, paragonite, margarite, clintonite, annite, ferriannite, eastonite, siderophyllite, and celadonite. In addition, some end-member micas, possibly occurring as very minor solid solutions in other natural micas, have been made. Some of the fluorine end-members, such as fluorophlogopite, fluoromuscovite, fluoroparagonite, as well as polylithionite and trilithionite, have been synthesized. Na-phlogopite is another synthetic end-member. Other compounds having the mica structure, but not found in nature, have been prepared for special commercial applications. Examples of successful synthesis include fluorine (F) aluminosilicates that contain barium (Ba), magnesium (Mg), potassium (K), and nickel (Ni). The formulae of such compounds are: $BaMg_3Al_2Si_2O_{10}F_2$; $KNi_3AlSi_3O_{10}F_2$; and $KMg_2MnAlSi_3O_{10}F_2$. The hydrous micas are grown in steel pressure vessels capable of withstanding high water pressures, and the fluoromicas are grown from melts in platinum, graphite, or bonded silicon carbide crucibles at approximately atmospheric pressure. A few crystals in excess of 26 square centimetres (four square inches) and 0.3 centimetre (⅛ inch) in thickness have been produced from fluoride melts. Smaller synthetic fluoromica crystals have been bonded into large sheets and other forms.

Stability limits. The geological conditions under which a specific composition or polymorph of mica may be found are limited. Phlogopite ranges in occurrence from the relatively low temperature and pressure environment in which marbles are believed to have formed to the high temperature and pressure environment of some ultrabasic rocks. In fact, phlogopite will grow in hot silicate liquids (magma) and persist in the lava flows that appear at the surface of the Earth. Other micas such as lepidolite, however, are confined to the moderate temperatures and pressures of environment in which pegmatites are formed. The glauconites are confined almost exclusively to marine sediments and to low-temperature and low-pressure environments. These qualitative restrictions deduced from field observations do not fully describe the limits of stability of a mica because of the constraints imposed by the available compositions, temperatures, and pressures in the Earth. Quantitative measure of the limits of stability have been carried out in the laboratory, where composition, temperature, and pressure can be varied at will. Data are available on many of the pure end-member compositions listed in Table 1, as well as on compositions generated by mixtures of two or three end-members. The measured maximum stability limits in the pressure–temperature region studied for some micas are listed with their decomposition or breakdown products in Table 2. For reference, the pressure in the crust of the Earth increases with depth at an average rate of about one kilobar (the bar is a standard unit of atmospheric pressure; the pressure at sea level is 1,013 millibars) per 3.66 kilometres (273 bars per kilometre [434 atmo-

Environmental temperatures and pressures required for mica formation

Table 2: Maximum Thermal Stability of Synthetic End-Member Micas
(temperature in °C)

reaction	pressure in kilobars					
	0.5	1.0	2.0	5.0	10.0	20.0
Celadonite ⇄ ferriphlogopite + ferrisanidine + quartz + H_2O		405	420			
Paragonite ⇄ albite + corundum + H_2O	520	535	565	630	675	
Muscovite ⇄ sanidine + corundum + H_2O	635	655	690	710	740	
Annite + oxygen ⇄ kalsilite + leucite + fayalite + magnetite + H_2O	745	785	825			
Ferriannite ⇄ fayalite + magnetite + wüstite + liquid + H_2O		835	825			
Phlogopite ⇄ forsterite + liquid			1,220	1,265	1,305	1,340
Muscovite + quartz ⇄ sanidine + andalusite (or sillimanite) + H_2O		525	600	690		
Phlogopite + quartz ⇄ sanidine + enstatite + liquid + H_2O	820	820	820			

spheres per mile]). In general, the thermal stability of mica increases rapidly with pressure initially and then increases at a slow rate at pressures above about two kilobars. The stability of phlogopite at very high pressures is noteworthy; it is a mineral that is likely to be present in the Earth's mantle and is of particular interest because of its water and potassium content. The potassium content has great bearing on heat production in the mantle because it is a radioactive element, and water tends to lower the melting temperatures in the mantle and thus affect magma production.

The stability of a mica is also dependent on the presence of other minerals and the composition of the ambient fluid (mainly water). In Table 2 the results of studies on micas themselves, and when in the presence of quartz (SiO_2), are given. The muscovite + quartz reaction is important because it is prominent in the metamorphism of shales or clay-rich sediments. In addition, the stability of muscovite appears to be influenced by the potassium ion–hydrogen ion ratio in the fluid; and the magnesium–iron ratios, especially in biotite, are sensitive to the presence of sulfur. The presence of fluorine enlarges the thermal range of stability of the hydrous tri-octahedral micas.

BIBLIOGRAPHY. W.A. DEER, R.A. HOWIE, and J. ZUSSMAN, *Rock-Forming Minerals*, vol. 3, *Sheet Silicates*, pp. 1–102 (1962), contains discussion and detailed references on mineralogical properties, structure, polymorphism, chemical composition, occurrence, synthesis, and stability.

(H.S.Y.)

Michael VIII Palaeologus

Most notable as the ruler who restored the Byzantine Empire to the Greeks after 57 years of Latin occupation, Michael VIII Palaeologus is the founder of the Palaeologan dynasty, the last and the longest lived of the Byzantine imperial houses (1261–1453).

A scion of several former imperial families (Ducas, Angelus, Comnenus), Michael was born probably in 1224 at Nicaea in Asia Minor, one of the several successor states established by the Greeks after the conquest of Constantinople by the Fourth Crusade (1204). After a rather uneventful boyhood, seemingly marked primarily by fantasies of himself recovering Constantinople from the Latins, he spent much of his youth living in the imperial palaces at Nicaea and Nicomedia.

His remarkable resourcefulness and talent for intrigue were revealed early. At the age of 21, he was charged by the emperor John III Vatatzes of Nicaea with treasonous conduct against the state, a charge from which he extricated himself by the force of his wit. Later, on the death of the emperor Theodore II Lascaris in 1258, Michael was chosen regent for Theodore's six-year-old son, John Lascaris. Gradually usurping more and more authority,

Seizure of the throne

Michael seized the throne and early in 1259 was crowned emperor after shunting aside and blinding the rightful heir, his charge, John. Faced with rebellion by Lascarid supporters in Asia Minor, Michael succeeded, in the eyes of many Greeks, in legitimatizing his rule by retaking Constantinople from the Latins. Whether as the result of Michael's carefully planned ruse or of accident, or both, the great city fell to his general in July 1261. Although the Greeks generally were exultant, a few realized that the centre of gravity had shifted from Asia Minor to Europe. In the long run this concern with Europe was to prove fateful; for, it led to the neglect of the frontiers in the East and with that neglect eventually to the conquest and settlement of all of Asia Minor by the Turks.

From the first, Michael's hold on the throne was precarious, surrounded as it was on all sides by Latins desirous of restoring Latin rule. Especially active was Baldwin II of Courtenay, the last Latin emperor of Constantinople. In his manoeuvres to recover his throne from Michael, Baldwin finally entered into a diplomatic and matrimonial alliance with a man who was the West's ablest diplomat —in his machinations almost the equal of Michael himself—Charles of Anjou, brother of St. Louis of France. At papal invitation, Charles had advanced into southern Italy, expelled the last representatives of the imperial house of Hohenstaufen, Manfred and Conradin, and then from Palermo and Naples almost at once fixed his gaze across the Balkans onto Constantinople. To quote a chronicler, "he aspired to the monarchy of the world, hoping thereby to recreate the great empire of Julius Caesar by joining East and West."

Threatened by this grand design was not only Michael but the papacy as well; should Charles succeed in the conquest of Constantinople, the pope would become a mere puppet of Charles in Italy. Hence development of an alliance between Michael and the papacy was logical, and it took shape in two decades of intricate negotiations for religious union between Rome and Constantinople. In exchange for the papal promise to restrain Charles from attacking Constantinople, Michael promised to bring about religious union of the Greek Church with Rome. That promise provoked the violent opposition of most of Michael's own people, who opposed union on doctrinal grounds; specifically they objected to such parts of the Latin liturgy as the *Filioque* (procession of the Holy Spirit from the Son and the Father) and the use of the azyme (unleavened bread). Perhaps more important,

By courtesy of the Bayerisches Staatsbibliothek, Munich

Michael VIII Palaeologus, detail of a Greek manuscript, 14th century. In the Bavarian State Library, Munich (Codex Monacensis Graecus 442).

most of them refused to accept papal ecclesiastical supremacy, which they felt, however obscurely, would lead to restoration of Latin political domination and possibly even cultural assimilation to the Latins.

Union between Eastern and Latin churches

Despite all the obstacles, union was nevertheless finally pronounced at the second Council of Lyons in 1274. Though the Orthodox East was coerced into accepting union, immediately after Michael's death (1282) the Greek Church declared the union invalid and a fraud. The Greeks had objected to the council on the grounds that not all the Eastern patriarchs or their representatives had been present, that no discussion on problems separating the two churches had taken place, and that no subsequent council had declared that of Lyons ecumenical. Nevertheless, for political reasons, Michael had struggled to maintain the union. But when Charles of Anjou finally managed to enthrone his own candidate, Martin IV, as pope in 1281, Martin at once excommunicated Michael and at the same time pronounced Charles's projected expedition against Byzantium a "Holy Crusade" against the "schismatic" Greeks. Included in the vast network of alliances erected by Charles to conquer the Greek East were not only Sicily, parts of Italy, Greek Lascarid dissidents, various Slavs of the Balkans, Baldwin, France, and Venice but also the papacy. Venice's aim in particular was to recover the broad trading privileges it had exercised in the days of the Latin Empire and to oust its archfoe, the Genoese, from the lucrative Greek markets.

The diplomatic duel between Charles and Michael was intensified, with Charles striving unceasingly to prepare his troops and navy. He even launched an attack across the Adriatic on Berat (in modern Albania) under the French general Sully but was repulsed by Michael. What Michael had on his side—the result of his consummate diplomatic ability—was (for a time) the papal alliance, a secret agreement with the Hohenstaufen supporters in Sicily, the support of Genoa, and finally, most important, a secret alliance with the son-in-law of Manfred, King Peter III of Aragon. The denouement to this remarkable contest was the outbreak on March 30 or 31, 1282, of the famous Sicilian Vespers, the massacre of the French signalling the revolt against Charles. Byzantium was saved from a second occupation by the Latins.

Assessment of Michael's life

At his death, which occurred soon afterward (December 11, 1282), Michael thus left an intact empire to his son Andronicus II. But it cannot be denied that his policy of using ecclesiastical union to stave off Charles's attack on his capital and the deep opposition that policy provoked among the Byzantine population established a fateful precedent for later Byzantine history. Moreover, by focussing his attention too exclusively on Europe, his policy helped to lead to Ottoman occupation of all of Asia Minor and ultimately to the capture of Constantinople itself. Nevertheless, Michael's positive accomplishments cannot be overlooked. He gave Byzantium two centuries more of life, rebuilt the capital, and re-established the University of Constantinople. His sponsorship of a general revival of learning led to the important Byzantine "Renaissance" in the 14th and 15th centuries.

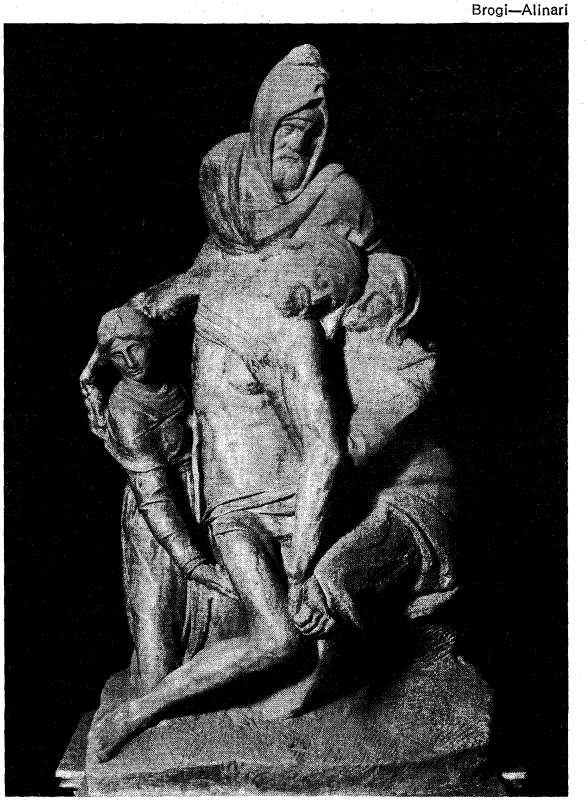
BIBLIOGRAPHY. For the most authoritative recent biography, see D.J. GEANAKOPLOS, *Emperor Michael Palaeologus and the West, 1258–1282: A Study in Byzantine-Latin Relations* (1959), with full bibliography. Also see C. CHAPMAN, *Michel Paléologue, restaurateur de l'empire byzantin* (1926), which contains extracts from Michael's so-called "autobiography." The most recent edition of the "autobiography" is by H. GREGOIRE, "Imperatoris Michaelis Palaeologi, 'De Vita Sua,'" *Byzantion*, 29–30: 447–476 (1959–60). For further references, see G. OSTROGORSKY, *Geschichte des Byzantinischen Staates* (1940, Eng. trans., *History of the Byzantine State*, 1956); and S. RUNCIMAN, *The Sicilian Vespers: A History of the Mediterranean World in the Later Thirteenth Century* (1958).

(D.Ge.)

Michelangelo

One of the major sculptors of all time, Michelangelo Buonarroti won incomparable recognition during the Renaissance for his paintings, architecture, and drawings. After the tardy appearance of his poetry in 1623, he slowly emerged as one of the most significant and personal poets of his age, his sonnets and madrigals appearing in almost all recent anthologies of Italian literature. So extraordinary was his influence on the history of art that the word Renaissance (Italian *rinascita*, "rebirth") was coined to characterize the saltation of the fine arts under Michelangelo and Raphael. According to the historian Giovanni Papini, the Reformation as well owes its very existence to Michelangelo. Papini reasons that Pope Julius II so increased the sale of indulgences to ensure the magnificence of his tomb, which Michelangelo was to execute, that Luther and others rebelled and split the church in two. Still a third definable age is occasionally credited to Michelangelo, who is often hailed as father of the Baroque for such tortuous and tense figures as the Haman of the Sistine ceiling. It is an overwhelming honour—if a debatable one—to be credited with the existence of the Renaissance, the Reformation, and the Baroque age, yet these claims at least suggest the influential role that Michelangelo played in the history of the fine arts.

Brogi—Alinari



"Pietà," marble sculpture by Michelangelo, *c.* 1550–56. In the Duomo, Florence.

Considering how few verifiable facts are known about William Shakespeare, who was born the year Michelangelo died, it is remarkable that the life of Michelangelo can be reconstructed almost month to month, if not week to week. This wealth of information derives largely from the 495 extant letters he wrote to others, the many written to him, the personal revelations in his poetry and artistic records, the biographies undertaken by his contemporaries (notably, Giorgio Vasari and Ascanio Condivi), conversations of his that were recorded by Donato Giannotti and Francisco de Hollanda, as well as from an abundance of direct and indirect testimony and comment by contemporaries who were aware that Michelangelo would outlive his age and be mirrored eventually in the glass of eternity.

CAREER

Early life and works. Michelangelo was born March 6, 1475, to Francesca and Lodovico di Lionardo Buonarroti Simoni in a small hill town near Arezzo called Caprese, now officially Caprese Michelangelo. Descend-

ed from two generations of Tuscan bankers, Lodovico was at the time nearing the end of his term as mayor of Caprese. Michelangelo's mother gave birth to five sons: her early death occurred December 6, 1481. All his life Michelangelo felt a strong bond with his father and brothers. The conscientious stream of letters to them and to his only nephew, Lionardo, from March 1497 to December 1563, pictures a Michelangelo unusually proud of his reported lineage from the counts of Canossa, eager to raise the status and fortunes of his family, interested in the health and welfare of each of the males who constituted that household, and hopeful for a brilliant survival of the Buonarroti Simoni line. The letters also reveal him as a son, brother, and uncle impatient with any relative's weaknesses or ingratitude, prodigal with complacent advice, and devoted to his kin in direct ratio to their willingness to stay away from him and let him work.

After Michelangelo's birth in 1475, the Buonarroti returned to Florence to share a crowded home in the Santa Croce neighbourhood with the family of Michelangelo's uncle Francesco. Michelangelo spent periods of his childhood on a family farm at Settignano, in the hills above Florence, where a wall drawing of his, a "Triton," can still be seen. On April 1, 1488, Michelangelo left the classroom of Francesco da Urbino to become an apprentice of the painter Domenico Ghirlandajo. His love of sculpture (he was to call it "the prime art") and a small stipend from the great Lorenzo de' Medici, the ruler of Florence, led to his leaving Ghirlandajo to study sculpture in the Medici garden (where Lorenzo had a collection of antique sculptures) under Bertoldo di Giovanni. This important step permitted him to know personally such Medici protégés as the philosopher Marsilio Ficino and the poet Politian, whose thought is reflected in his later art and poetry. At this time he executed his reliefs "Madonna of the Stairs" and the "Battle of the Centaurs." After this residence in the Medici household, he would be marked in Florence as a "Medici man," which, in the vicissitudes of Tuscan politics, could be either an advantage or a disadvantage. Indeed, after the death of Lorenzo in April 1492 and days before the expulsion of his son Piero in 1494 by the preacher and religious reformer Savonarola, Michelangelo prudently fled to Bologna. Housed by the nobleman Giovanni Francesco Aldrovandi in Bologna, he carved three small statues for the tomb of St. Domenic in the church bearing that saint's name. Having returned to Florence in 1495, he executed a sculptured "Sleeping Cupid" (now lost), making it appear as an antique. After it was resold as a work of antiquity to Cardinal Riario in Rome, Michelangelo offered to reimburse the money he had originally received for it. One important sculpture Michelangelo completed while still in Florence, probably before 1494, was a wooden "Crucifixion" (authenticated only in 1965), which he offered in gratitude to the church of S. Spirito for allowing him to study the anatomy of the corpses stored therein.

Apprenticeship with Ghirlandajo

Many reasons may have induced Michelangelo to leave Florence for Rome, where he arrived June 25, 1496: the discrediting of the Medici, Savonarola's burning of works of art, the invasion of Charles VIII of France, or perhaps, as Condivi suggests, his view of Rome as "a broad field in which a man may demonstrate his worth." From the first he was troubled by the bellicose and materialistic spirit of Rome, where, as he wrote, "Chalices become swords and Christ's blood is sold by the bucketful."

In Rome, as if mindful of one of Politian's Bacchic poems, he carved for the banker Jacopo Galli a marble "Bacchus," as pagan in spirit as the Greek fauns in the Capitoline Museum. In contrast, he turned in 1498 to the Christian themes that were to dominate his art henceforth; on a commission from Cardinal Jean Villiers de la Groslaye, he executed the marble "Pietà" that would later confront churchgoers entering St. Peter's basilica. Michelangelo explained that the youthful features of his Virgin were the result of her "never having entertained the slightest immodest thought which might have troubled her body." This "Pietà" was defaced in 1972 and subsequently restored.

Back in Florence in 1501, Michelangelo signed contracts for 15 figures for the Piccolomini altar in Siena, a commission only partially carried out, and for the marble "David," which he completed in April 1504. Immediately hailed by Florentines as "the Giant," it was viewed by the democratic Commune as an allegory of the moral strength of Florence. The Commune then requested a bronze statue of "David" to be commissioned as a diplomatic gesture to the Maréchal de Gié, the virtual ruler of France at the time. Completed with the help of Benedetto da Rovezzano in 1508, this work, last seen in the 16th century in the gardens of the Robertet Château in the Loire Valley, was apparently later moved to the Château de Villeroy near Mennecy, France. A marginal annotation penned alongside a sketch of Michelangelo's bronze "David"—"David with his sling and I with my bow"—permits the conclusion that the figures of David (two sculpted and two painted) in the artist's work represent Michelangelo's self-projections.

Late in August 1504 (possibly even in October) the secretary of the Commune, Piero Soderini, convinced Michelangelo to undertake a fresco of the "Battle of Cascina," fought in 1364 between Florentines and Pisans, for the Hall of the Five Hundred in the Palazzo Vecchio. This commission heightened the rivalry with his older competitor Leonardo da Vinci, who was painting the "Battle of Anghiari" on the wall opposite. This vast wall, later destroyed, illustrates Michelangelo's addiction to the tense, contorted male nude that reached its apogee in the vault of the Sistine Chapel and "The Last Judgment." His glorification of the male figure, in part a consequence of his long study of anatomy, he further justified by a Platonic assertion that it was the "veil" through which God revealed himself.

Mature life and works. Meanwhile Pope Julius II, aware of the absence of the young artist, recalled him to Rome, thereby initiating the temperamental, often bickering friendship that was to endure until the pontiff's death in February 1513. The Pope ordered Michelangelo to work on a vast sepulchre for him, containing over 40 lifesize figures, which was to be the eighth wonder of the world. The "tragedy of the Tomb," a tomb never completed, was to harass Michelangelo for years: problems of quarrying, inadequate assistants, niggardly financing, papal tantrums, and revisions of contract troubled the sculptor long after the Pope's death, when Julius' heirs threatened him with one lawsuit after another to bind him to complete the project. On April 17, 1506, Michelangelo fled Rome on horseback after an affront by the Pope. In Florence he entertained thoughts of accepting a commission from Bayazid, sultan of the Ottomans, to build a bridge between Constantinople and Pera. Piero Soderini urged him rather to effect a reconciliation with the Pope. So, on November 28, Michelangelo set off for Bologna to humiliate himself before the Pope, who was enjoying his military conquest of that city, and to agree to execute a bronze statue of Julius for the church of S. Petronio. This statue, installed in February 1508, was damaged by the vengeful Bolognese after restoration of the Bentivoglio regime and sent to Ferrara where Duke Alfonso d'Este had it melted down into a cannon in May 1512.

Recall to Rome by Pope Julius II

In May 1508 Michelangelo received the first payment from the Pope to begin the glorious expanse of the Sistine Chapel. Unhappy at being forced to paint rather than sculpt, he rigidly noted on the receipt that he, a sculptor, had received 500 ducats for the painting in the Sistine Chapel. A humorous, extended sonnet to his friend Giovanni da Pistoia at this time records the difficulties of this herculean task and complains that he "is not in the right place, nor is he a painter." (Many historians of his painting have commented, in fact, on its sculptural quality.) The panoramic vault, balancing pre-Christian prophets against the pagan sibyls, knitted together by nine histories and four spandrel (space between the exterior curve of an arch and an enclosing right angle) pictures from the Old Testament, fringed with the ancestors of Christ, and liberally decorated with nude youths that worried five popes, required his attention until October 31, 1512.

The Medici having returned to power in Florence in 1512, Michelangelo resided there intermittently between 1514 and 1534, after which date he lived in Rome until his death.

In May 1513 the sculptor was able to renegotiate the contract for Pope Julius' tomb with the pontiff's heirs and to work on the awesome "Moses" and two "Slaves" to adorn that sepulchre. In the meantime, two Medici popes, Leo X (the second son of Michelangelo's first patron, Lorenzo the Magnificent, who was elected pope in 1513) and Clement VII, were to keep him busy on their projects. In February 1518 Michelangelo designed a model for the facade of the Medici parish church of S. Lorenzo. Yet, after considerable work on the project, especially quarrying for marble at Carrara and Pietrasanta, he was transferred from it and in November 1520 signed a contract to plan a mortuary chapel in S. Lorenzo to contain tombs for the Medici family, starting with sepulchres for Giuliano, duke of Nemours, and Lorenzo, duke of Urbino, both recently deceased. From 1513 on, Michelangelo suffered under the contrary pressures of the Pope and the cardinal de'Medici, on the one hand, and the heirs of Julius II, on the other, each party demanding his undivided services. When this pressure began, Michelangelo wrote in a letter, "I am solicited so much that I cannot take time to eat."

Devotion to Dante

Michelangelo's great devotion to the Italian poet Dante, who had died in 1321, came to the fore in October 1519, when he approached the Pope with an unsuccessful offer to execute a tomb in Florence if the ashes of Dante could be returned from Ravenna. In a sonnet to the poet, Michaelangelo hailed him as a man "without peer." In another he wrote, "Would that I were he! Born to such a choice, I should give up the most felicitous state in the world for bitter exile, just to possess his virtue."

After Pope Leo's death in 1521, he was succeeded very briefly by Adrian VI, a patron with little interest in art. In November 1523 the artist's taskmaster Cardinal de' Medici became Pope Clement VII. The following year, he added to Michelangelo's chores in Florence the designing of the Laurentian Library. Even as Michelangelo worked on the New Sacristy and the Laurentian Library in 1526, he proceeded with the effigies of the two dukes. These two figures, detached from and set above their tombs in niches, dressed as placid captains of the church in diaphanous togas revealing their musculature, represented a new departure in mortuary design. Another departure was the idealized portrayal of the clean-shaven dukes, leading Florentines who had known them to complain that they did not resemble their bearded prototypes. Michelangelo, always interested in idealized representation, replied: "A thousand years from now no one could judge that they were otherwise." Four reclining marble statues of the times of day ("Dawn," "Day," "Dusk," "Night") represent the rapid passing of time, a constant motif in the artist's thought; in a prose poem, for example, Michelangelo regrets that the rapid course of night and day has brought Duke Giuliano to an early death.

The year 1529 brought a curious interruption to his busy schedule. Pessimistic about man in general, Michelangelo could only conclude that politics was an unattractive business to be shrugged off by an artist. He once wrote, "Artistic matters give me enough worries without politics." He was as emphatic in his condemnation of war: "He who wins by flight, let him quit the field." After the sack of Rome by the Spanish emperor in May 1527, the Pope and Emperor settled their differences and planned an attack against the Florentine Republic. In January 1529, Michelangelo was, despite himself, named magistrate of the Committee of Nine of the Florentine Militia and, in April, governor and procurator general of the city's fortifications. In July and August he visited the fortifications of Bologna, but in September he yielded to an impulse to "win by flight" and abandoned Florence for Ferrara and then Venice. In November he was declared a rebel by the City of Florence, only to return probably in December and continue fortifying the walls. In August 1530 the republic fell, and the Medici once more governed Florence. Michelangelo went into hiding and remained hidden until Pope Clement tacitly pardoned him for abandoning the Medici Chapel for the fortification of the city. While several of his friends went into exile in Rome and elsewhere, Michelangelo remained in Florence to complete the Medici Tombs (1534).

In 1531 Michelangelo's father, Lodovico, died, and his son wrote an affectionate eulogy in verse. The theme of the poem is that one is wrong to fear death, for the souls in heaven are emancipated, free of concerns and pressures, enviably exempt from hardships. Visiting Rome in July 1532, Michelangelo met a young nobleman, Tommaso Cavalieri, for whom he conceived a lasting Platonic love, a love he expressed in many sonnets and allegorized in many drawings (among them "Ganymede" [since lost], "Tityus" [Royal Collection, Windsor Castle], and "The Fall of Phaëton" [British Museum, London], according to Vasari). Their friendship endured until the artist's death, when the younger man sealed Michelangelo's eyes. Cavalieri and the widowed marchioness of Pescara, Vittoria Colonna, whom Michelangelo apparently met in 1536, were the two friends who penetrated most deeply Michelangelo's psychic sanctuary and gave the lie to his earlier (1509) claim: "I have no friends of any kind, nor do I wish any, for I have not so much time that I can afford to waste it." They form a contrast to the two young men, probably models, Gherardo Perini and Febo di Poggio, with whom Michelangelo probably had brief relationships of a homosexual nature.

On September 23, 1534, two days before the death of Clement VII, Michelangelo left Tuscany forever, returning to Rome, where the new pope, Paul III (Alessandro Farnese), appointed him official sculptor, painter, and architect of the Vatican. In May 1536 he was at work on "The Last Judgment" in the Sistine Chapel, a powerful fresco of the Day of Wrath inspired by the medieval hymn *Dies irae*, by his reading of Dante, and by his own fears of being judged unworthy at the Resurrection. His Christ-centred religiosity, often verging on mysticism, emerged both in verses composed concurrently with "The Last Judgment" and in the magnificent Christian poetic effusions of the 1550s, when the artist was "living on credit in the hostel of death" and fearful of salvation. His greatest anxiety, occasioned by the thought of the certainty of death, was his unworthiness to be judged on the Day of Wrath. His poems show him viewing death as a Gothic grim reaper, cutting down humanity like blades of hay. He was convinced that every passing day made a man less worthy of salvation: "he who lives least may hope more for pardon." From the unfurled banners of heaven at the top of the cosmic scene of "The Last Judgment" to the tumbling sinners condemned to hell at the bottom, the anger of the central figure of Christ the judge dominates the work. Whereas many theories of collective guilt have been elaborated to explain Christ's anger (for example, anger because of French military invasions into Italy), the Saviour's wrath and upraised arm are directed at individual sinners such as Michelangelo (clearly portrayed below the Saviour in the pelt of St. Bartholomew) who have disregarded Christian law. This personal meaning is borne out amply in the poetry of Michelangelo by such lines as "Hear not my fault with thy chaste ear, nor raise toward me thy righteous arm" and "Lord, in the last hours stretch toward me thy forgiving arms." "The Last Judgment" is clearly a testimonial to the artist's concern for his own soul.

Inspiration for "The Last Judgment"

A similar personal meaning is associated with his bust of "Brutus," which he offered in 1539 to his friend the historian Donato Giannotti. In a transcribed record of a conversation in Giannotti's home, Michelangelo speaks at length about the guilt of Brutus the tyrannicide. Michelangelo approved of Dante's placing Brutus in the lowest pit of hell (in the *Divine Comedy*), as slayer of a benefactor, and yet he felt obliged to admit to the Florentine exiles Giannotti and Luigi del Riccio that "he who kills a tyrant kills a beast." This final admission results from Michelangelo's disapproval of the Medici heirs then ruling Florence and of Duke Alessandro, murdered in 1537.

With "The Last Judgment" uncovered (October 31, All Saints Eve, 1541) and universally praised, Michelangelo turned for the last time to the tomb of Julius II. It was agreed with the Duke of Urbino that other masters would be brought into the project to relieve Michelangelo. In October 1542 the artist penned a plaintive letter about this project:

> I find myself having lost my youth and manhood tied down to this tomb, which I defended as much as possible with Popes Leo and Clement, and my excessive loyalty which no one has consented to acknowledge has ruined me. Thus my destiny wishes it.

The much reduced version of this "eighth wonder of the world," completed in February 1545, may be seen today in its modest chapel within the church of S. Pietro in Vincoli. Its chief attraction is the patriarchal "Moses," a prime display of *terribilità*, a quality Michelangelo acknowledged that his works possessed—and a quality that led the 19th-century English novelist William Makepeace Thackeray to remark that he would not wish to be left alone in a room with it. The curious horns that "Moses" wears are generally accepted as symbols of light rays derived from primitive statuary.

In November 1542 Michelangelo received payment to start the painting of the Cappella Paolina, a project completed in 1550. Two large frescoes depicting the "Conversion of St. Paul" and the "Crucifixion of St. Peter" were completed in 1545 and 1550. In July and August of 1544, Michelangelo was sufficiently ill that his friend Luigi del Riccio had him transferred to the Strozzi palace for care and recuperation, a transfer effected again in December 1545.

The bulk of Michelangelo's poetry having been written between 1530 and 1546, del Riccio collected the scattered poems and planned an edition of them. A rupture between the two friends (the cause is uncertain, but Michelangelo accused his friend of ingratitude) frustrated the project. The subsequent poems, the "trembling hand" verses, become more and more concerned with old age, approaching death, doubts, and renunciation. His fame as a poet gained ground when, in 1549, the Florentine humanist Benedetto Varchi published a textual exegesis of a sonnet embodying Michelangelo's Platonic view of art as something preordained by God and implanted within the marble or pigment to be "perceived" by the artist. Michelangelo wrote in return, "The sonnet comes from me all right, but the Commentary comes from heaven." The following year Giorgio Vasari published his biography of Michelangelo, and, in 1553, Condivi's life of the artist appeared.

Fame as a poet

In 1550 Michelangelo began work on his "Pietà" ("Deposition") to be placed in the Florence Cathedral. That it is less finished than other works does not lessen its emotional impact. For Michelangelo's theory of art as a preordained basic form allowed him to leave many of his sculptures consciously unfinished. To Michelangelo, the challenge offered to the artist is to capture the basic form that "grows" within the material, to "break the formula." The finishing process is a postlude and, indeed, could easily be entrusted to assistants. In the "Pietà," the group of Christ, Mary Magdalene, and Nicodemus is superbly done at focal points, with the finishing to have been undertaken by Tiberio Calcagni (who died in 1565 before doing so). The compassionate Nicodemus (a self-portrait of the artist) seems to be consoling the limp Christ with the words of Michelangelo's great hymn to the Redeemer, "Thou wast condemned to death that thou might with thy blood open the gates of heaven to man." Vasari wrote that Michelangelo intended this group to decorate his own tomb, even though he adds that Michelangelo at one point deliberately broke the stone "probably because it contained many flaws."

In March 1555, Pope Julius III died, to be succeeded by Marcellus II (Marcello Cervini), with whom Michelangelo had often disputed. The artist had ordered Cervini not to interfere with the activities in the Vatican workshop, to "keep the money coming in and keep it out of the hands of thieves, and as for the planning of the building, leave that to me!" This outburst is perhaps typical of Michelangelo's sometimes crotchety character, which created hostility among some of the assistants in the Vatican workshop. His rigidity is revealed in many self-assured phrases in his letters: "I am firm in my opinion," "I don't contradict myself," or "I'll abide with my original opinion." It was obvious that Marcellus II was going to side with the dissident workers who were loyal to the architect Sangallo. They had long complained that Michelangelo kept them ignorant of his plans. Fortunately for the artist, the pontiff's demise six weeks later turned the tide again in his favour. Marcellus' successor, the redoubtable Cardinal Carafa, who had been chiefly responsible for recent book burnings and the establishment of the *Index Librorum Prohibitorum*, was well disposed toward the octogenarian artist. As Pope Paul IV, he reconfirmed Michelangelo's title as architect of St. Peter's. Michelangelo was also protected by Bishop Beccadelli, a friend with whom he exchanged sonnets, against the workers' cabal.

In September 1556, Michelangelo's tendency toward panic flight before the threat of war overcame him one last time. With Rome threatened by the Spaniards under the Duke of Alba, Michelangelo took refuge at Spoleto. He whose art consistently disdained the portrayal of nature as landscape returned to Rome with the written declaration: "Less than half of me has returned to Rome. For truly one cannot find peace except in the woods."

In 1559, the Counter-Reformational zeal of Pope Paul led him to order Michelangelo's good friend the artist Daniele da Volterra to cover with loin cloths many of the nude figures in "The Last Judgment." Michelangelo's dismay could hardly have exceeded that of Volterra, who was given the mocking title of il Brachettone (the Breeches Maker) by the Roman populace. (It was Volterra, by the way, who executed the most accurate portrait bust of Michelangelo [in bronze] after casting a death mask of the deceased master.)

Late life and works. In his late years, Michelangelo's projects and commissions were mainly architectural, as though his tired arms needed respite from the hammer and chisel. These included the cornice of Palazzo Farnese (1547), the piazzale of the Capitoline Hill (1546), the designing of St. Peter's cupola (1557–58), the planning of the projected church of S. Giovanni dei Fiorentini (1559), designs and studies for the Porta Pia (1561), and the accommodation of the nave of the church of Sta. Maria degli Angeli into the precincts of the Baths of Diocletian (1561). Although familiar with the treatise on architecture by the Roman Vitruvius (1st century BC), Michelangelo typically declared to Pope Paul that he was no architect. Yet he did not hesitate to evaluate other architects (Bramante rated high, Antonio da Sangallo very low) and to boast of his plans for his S. Giovanni dei Fiorentini: "Neither Greeks nor Romans ever constructed such a thing among their temples." Criticized since mid-19th century for treating architectural details arbitrarily, for indifference to constructive efficiency, for the loss of conformity to constructive principle, Michelangelo has been praised by contemporary architects for having reacted to the ideal of static building during the High Renaissance.

Architectural projects

Yet, during his last days, he returned to "sweet sculpture," to the carving of the "Palestrina Pietà" (the authenticity of which is contested by only one historian), and the "Rondanini Pietà," on which the sculptor worked almost to his dying day. It was as though Michelangelo, fearful unto death of his unworthiness, had to carve one last time the pathos of the descent from the Cross and what he called the "mansuetude" (meekness or gentleness) of the Virgin Mary.

Michelangelo's Spartan habits and perhaps even his heavy work schedule toughened him for a long life in art. Of course he complained about his health, just as he sooner or later complained of much else in life. In a verbal self-portrait he claims to be plagued with diarrhea, coughs, catarrh, short-windedness, snoring, lumbago, hernia, exhaustion, dizziness, pain in his sides, weak teeth, and spotted eyes, weak hearing in one ear and a singing in the other—all this in addition to the painful gallstones for

which he sought relief from the waters of Viterbo. Yet his productivity is a testimony, if not to health, to his stamina and self-discipline. He who complained of his "imminent death" about 48 years before it took place managed to fulfill almost 89 years of life, the last of old Lodovico's sons to go. On February 14, 1564, he apparently suffered a stroke. Late in the afternoon of the 18th, Michelangelo died, Cardinal Salviati administering last rites and Daniele da Volterra, Tommaso Cavalieri, and a few other friends attending at the bedside. His last confession to Cardinal Salviati was worthy of the artist laureate of Christendom: "I regret that I have not done enough for the salvation of my soul and that I die just as I am beginning to learn the alphabet of my profession."

Although Michelangelo had been made an honorary citizen of Rome (in March 1546), he stated two days before death that he wished to be buried in Florence. Even after a state funeral in Rome's church of the SS. Apostoli, his nephew Lionardo managed to recover the body, have it baled in a packing box, and smuggle it to Florence. After more ambitious obsequies for their native son arranged by the Florentine Academy, Michelangelo's townsmen laid him to rest in Sta. Croce, where his sepulchre, designed by Vasari, may be seen next to the cenotaph of his lifelong hero Dante Alighieri.

CHARACTER AND ASSESSMENT

Michelangelo's poetry as a key to his character. The poetical works of Michelangelo, which have been correctly characterized as a "spiritual biography," consists of 343 sonnets, madrigals, *capitoli* (satirical pieces in tercet rhyme), quatrains, and fragments, unfinished pieces like much of his late art. Written without thought of publication, they provide a direct avenue to the artist's beliefs, thoughts, emotions, and moods. They were an intimate journal in which he could confide his thoughts on love, religion, friendship, politics, and even art, which he felt late in life had been a "false idol" and had pre-empted the years "granted him to meditate on God," as one sonnet complains. Some poems were shared with his few intimate friends, Luigi del Riccio, Tommaso Cavalieri, or Vittoria Colonna. A few served as verse epistles to his contemporaries: a gruff lecture to Pope Julius II, a request to Donato Giannotti to "adjust" a poem, a sonnet to Giorgio Vasari "to prove that he is in his second childhood," a gracious refusal to undertake a bust for Gandolfo Porrino, and so on. At one of the busiest times of his career, Michelangelo acceded to the request of Luigi del Riccio to compose 50 epitaphs on the death of the latter's adolescent nephew, epitaphs not always noble in quality and containing undertones homosexual in character. Never an activist in politics, he penned only a few poems that reflected either a direct or indirect condemnation of the Medici ravishers of the Florentine Republic.

Themes and moods of the poems

The lonely Michelangelo spent a great deal of time writing love poems. Just as symbols of love haunt his art (Venus and Cupid, Leda and the Swan, Samson and Dalilah, Tityus, Ganymede, for example), so the artist depicts himself as a figure assailed by love. He writes with consuming interest of the vagaries, suspense, power, renunciation, and even definition of love. He warns that "unbridled desire is lust and not love, and it kills the soul." Accordingly, his love poetry is usually spiritualized; but a few pieces wander away from the *amor onesto* ("true love"), and in his late poetry he regrets his indulgence in sensual pleasure. The poems fall into three classifications: generalities on love, poems to women (neither identified nor even authenticated, except for his chaste autumnal love, Vittoria Colonna), and poems to men (Cecchino Bracci named directly and Febo di Poggio indirectly).

The large corpus of confessional and mystical poetry of astonishing sincerity and revelation (discussed earlier) is valuable not only in itself but for the light it sheds on Michelangelo's art and the circumstances surrounding that art. The basic theme of the creation of day (light) and night (darkness) twice depicted on the Sistine ceiling is developed in several poems on the passage of time, especially in lines specifically alluding to the two creations on that vault:

> He who from nothing, and ere all things were, did time create, divided it in twain. To one he made the lofty sun a share, and to the other moonlight's dusky train.

Michelangelo continually deprecated his efforts as poet. He laughed off his poems as novelties and confections. To one poem he appended, "at Carnival time it seems legitimate for those not masked to commit some folly." Accompanying a sonnet sent to Vasari went the notation: "You will say that I am old and mad trying to make sonnets, but since many say I'm in my second babyhood, I've wished to play my part." Although he disclaimed its value, he was flattered and pleased at praise of his poetry. Yet he was hesitant to allow publication of them during his lifetime and frustrated the plan of Luigi del Riccio and Donato Giannotti to have them published. This heralded poet was to appear in print posthumously, when his great-nephew, the Florentine academician Michelangelo the Younger, brought out the first edition of the *Rime* in 1623. Unfortunately, this editor tampered with the Vatican and other texts and bowdlerized poems addressed to men, changing masculine pronouns to feminine. It was only in 1960 that a definitive edition was published.

First edition of the poems

A man of the Renaissance. Thanks to the long span of his lifetime and to the milieu in which that life was passed, Michelangelo was, despite himself, a man of the Renaissance, that Renaissance of which Vasari saw him the incarnation. As a product of the Humanistic age, Michelangelo's centrifugal interests brought him the breadth and universality to embrace and conciliate the polarities contained within the Italian Renaissance. Three such polarities seem particularly significant.

In the Middle Ages little art theory or criticism existed, leaving a gap between artistic self-consciousness and technical achievements. One of the chief characteristics of the High Renaissance was precisely this professional self-consciousness sensed by the artist, which led him inevitably to become both practitioner and theorist. In an age and country rich in treatises on art, Michelangelo himself planned a study to be entitled "Treatise on All Manners of Human Movements, Appearances, and Anatomy, with an Ingenious Theorick." Despite a reticence to talk about himself or even about art, he planned a treatise on painting as well. Although neither of these works of theory was ever written, Michelangelo left ample evidence of the nature of his speculation on art and artists in letters and conversations. Responding to the unremitting curiosity of his contemporaries, his many value judgments on artists were unusually clear and articulate. His immersion in art made it impossible for him to eschew the role common in the Renaissance of artist-theorist.

The greatest polarity to be reconciled in his lifetime resulted from the confrontation of pagan Humanism and Christian orthodoxy and authority. Here again, Michelangelo's thoughts and works served a unifying, conciliating purpose. His devotion to both Christianity and the Platonism of his friend Marsilio Ficino contributed to that accommodation. During his youth and manhood, creatures of Greco-Roman mythology, such as Cupid, Ganymede, Bacchus, Tityus, counterbalanced figures from the Christian theogony. The most conspicuous example of this accommodation is the juxtaposition of Old Testament prophets and pagan sibyls on the Sistine vault. His Neoplatonism was most noticeable in his poetry of the 1540s, the decade during which he executed the "Brutus." When the High Renaissance gave way to the Counter-Reformation, with Michelangelo being condemned at the Council of Trent by Gilio da Fabriano, the accommodation in which Michelangelo and the "Christian Humanists" had pioneered survived with some difficulty.

The third and last conciliation was the one that the artist effected among the three separate major arts. After his early, almost exclusive partiality to sculpture, Michelangelo was able to state in February 1546 that "everyone knows I am a sculptor, painter, and architect." It was not merely his increasing practice and success in all three arts, however, that led him to view them as parts of a

larger whole. It was rather the unifying embrace of Humanism, its unitarian drive, that led him to adopt as his personal emblem three circles interwoven, as he explained, to show the interdependence of each on the other two. Many of his comments shattered the rigid distinctions among the parallel arts. In an epistle to his academic friend Benedetto Varchi, Michelangelo declared "to no purpose" Benedetto's referendum on the relative art of painting and sculpture, since they are parts of the same phenomenon, are of equal nobility "philosophically speaking," serve the same purpose, and derive from the same experience and inspiration. Despite a personal preference for the practice of carving, he knew that, "philosophically speaking," the underlying unity of art was to be found in design (in Renaissance Italian the "fine" arts were then called the "arts of design"), and his insistence on the oneness of the arts issued from this knowledge. Design, draftsmanship, drawing—this was the basic talent he demanded of his colleagues and assistants. In the Roman dialogues transcribed by Francisco de Hollanda, Michelangelo views all arts as subservient to design. While others, such as Varchi and Leonardo da Vinci, were stressing the dissimilarities of the arts, Michelangelo held consistently to the unitarian view.

Interdependence of Michelangelo's arts

Thus, this unitarian tendency as a Renaissance thinker led the artist, so ungregarious and unaccommodating as an individual (Raphael called him "the hangman") to become conciliatory in the areas of aesthetics and philosophy, uniting artistic theory and practice, Christianity and paganism, and the three arts of design.

MAJOR WORKS

SCULPTURE: "Madonna of the Stairs" (*c.* 1492; Casa Buonarroti, Florence); "Battle of the Centaurs" (*c.* 1492; Casa Buonarroti); "Crucifixion" (before 1494; S. Spirito, Florence); "Bacchus" (1497; Bargello, Florence); "Pietà" (1499; St. Peter's, Rome); "David" (1501–04; Accademia, Florence); "Madonna and Child" (1501–04; Notre-Dame, Brugge); "The Madonna and Child with the Infant St. John" ("Taddei Tondo"; tondo, *c.* 1503; Royal Academy, London); "Madonna and Child" ("The Pitti Madonna"; tondo, *c.* 1503; Bargello); "Moses" (for Tomb of Pope Julius II, 1505–45; S. Pietro in Vincoli, Rome); "Christ Bearing the Cross" (1519–20; Sta. Maria Sopra Minerva, Rome); Medici Tombs ("Dawn," "Dusk," "Day," "Night"; 1520–34; S. Lorenzo, Florence); "David" ("Apollo"; *c.* 1530; Bargello); "Victory" (*c.* 1532–34; Palazzo Vecchio, Florence); "Pietà" ("Deposition"; 1550–56; Duomo, Florence); "Palestrina Pietà" (*c.* 1550; Accademia, Florence); "Rondanini Pietà" (1552–64; Castello Sforzesco, Milan).

DRAWINGS: Major collections: British Museum; Royal Library, Windsor; Ashmolean Museum, Oxford, England.

ARCHITECTURE: Chapel of Leo X (dedicated to SS. Cosmas and Damian; facade, 1514; Castel Sant'Angelo, Rome); Laurentian Library (1523–59; S. Lorenzo, Florence); Palazzo Farnese (facade completed 1547; Rome); St. Peter's dome designed 1557–58; Rome); Porta Pia (1561, Rome).

PAINTINGS: "Holy Family" ("Doni Madonna"; *c.* 1503–05; Uffizi); frescoes, Sistine Chapel ceiling (1508–12; Vatican); "The Last Judgment" (1533–41; Sistine Chapel, Vatican); "Conversion of St. Paul" (1542–45; Cappella Paolina, Vatican); "Crucifixion of St. Peter" (1542–50; Cappella Paolina, Vatican).

BIBLIOGRAPHY

Major published bibliographies: ERNST STEINMANN and RUDOLF WITTKOWER, *Michelangelo Bibliographie* (1927), covers the period 1510–1926; *Michelangelo im Spiegel seiner Zeit* (1930), covers the period 1927–30; P. CHERUBELLI, "Supplemento alla bibliografia michelangiolesca," in *Michelangelo Buonarroti del IV centenario del Giudizio Universale* (1942), covers the period 1931–42; *Zeitschrift für Kunstgeschichte*, brings bibliography from 1942 to the present.

Reproductions and sources: LUDWIG GOLDSCHEIDER (ed.), *The Sculptures of Michelangelo*, 2nd ed. rev. (1950); *The Paintings of Michelangelo*, 2nd ed. with revised plates (1948); *Michelangelo Drawings*, 2nd ed. (1966); *The Complete Work of Michelangelo*, foreword by MARIO SALMI (1966); *The Complete Poems and Selected Letters of Michelangelo*, trans. by CREIGHTON GILBERT, rev. ed. (1965); *The Letters of Michelangelo*, trans. by E.H. RAMSDEN, 2 vol. (1963).

Biography and criticism: JAMES S. ACKERMAN, *The Architecture of Michelangelo*, 2 vol. (1961), views Michelangelo's buildings "conceived as if the masses of a structure were organic forms capable of being molded and carved, of expressing movement, of forming symphonies of light, shadow and texture, like a statue" (this challenging thesis less widely accepted in Italy than elsewhere); GEORG BRANDES, *Michelangelo Buonarroti*, 2 vol. (1921; Eng. trans., *Michelangelo: His Life, His Times, His Era*, 1963), a classic biography that maintains its reliability as well as its readability as a dramatic portrait; ROBERT J. CLEMENTS, *Michelangelo's Theory of Art* (1961), a collection of verifiable references to Michelangelo's ideas on the subject of the fine arts, both philosophical and practical, set in the greater context of Renaissance aesthetics; *The Poetry of Michelangelo* (1965), textual analyses of Michelangelo's poetry; HERMANN GRIMM, *Leben Michelangelos*, 2 vol. (1860–63; Eng. trans., *The Life of Michael Angelo*, 2 vol., trans. by F.E. BUNNETT, 1865), a pioneer biography by a German scholar that has now been superseded, despite many valid insights; CHARLES HOLROYD, *Michael Angelo Buonarroti, with Translations of the Life by Ascanio Condivi and Three Dialogues from the Portuguese of Francisco d'Ollanda*, 2nd ed. (1911), accurate translations of two major sources of information about Michelangelo, his contemporaries Ascanio Candivi and Francisco da Hollanda; CHARLES H. MORGAN, *The Life of Michelangelo* (1960), a sensitive, accurate record of the artist's life and accomplishments, splendidly written and interpreted; GIOVANNI PAPINI, *Vita di Michelangiolo nella vita del suo tempo* (1949; Eng. trans., *Michelangelo: His Life and His Era*, 1952), the Florentine author and scholar sheds new light on many obscure points in Michelangelo's life, especially his relations with such contemporaries as Francesco Granacci, Contessina de'Medici, II Pontormo, Sofonisba Anguissola, and many others; ROMAIN ROLLAND, *La Vie de Michel-Ange*, 2 pt. (1905–06; Eng. trans., *Michelangelo*, 1962), a highly-charged presentation of the artist's life and problems, drawing from German scholarship the most effective and dramatic facts about Michelangelo's life; CHARLES DE TOLNAY, *Michelangelo*, 5 vol: *The Youth of Michelangelo* (1943), *The Sistine Ceiling* (1945), *The Medici Chapel* (1948, reprinted 1970), *The Tomb of Julius II* (1954, reprinted 1970), and *The Final Period* (1960), a definitive presentation of the basic known facts about the artist's life and works, with little personal conjecture (an exception being de Tolnay's theory that the histories of the Sistine Vault be viewed not in the order presented by the Old Testament, but in reverse, as a Platonic *ascensio*); *Werk und Weltbild des Michelangelo* (1949; Eng. trans., *The Art and Thought of Michelangelo*, 1964), a distillation of Michelangelo's thought on politics, philosophy, religion, and art.

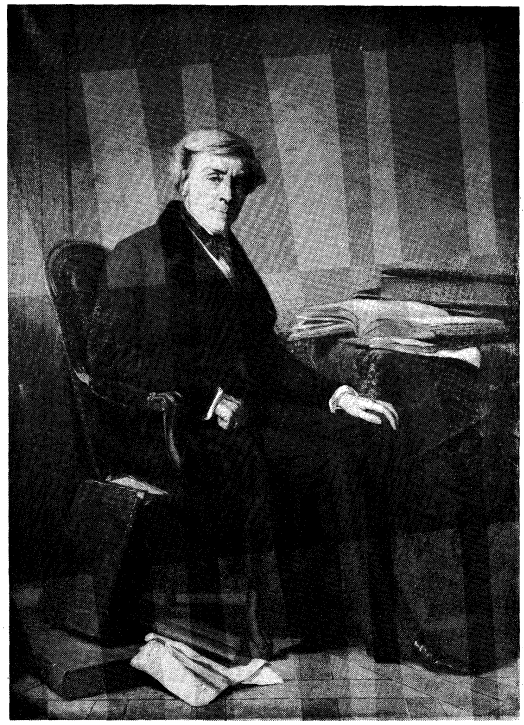
(R.J.C.)

Michelet, Jules

One of the first and greatest of the nationalist and romantic historians of France, Jules Michelet was born in Paris on August 21, 1798. His father, a modest printer, managed to give him an education. A brilliant and laborious student, Michelet at 29 was teaching history and philosophy at the École Normale Supérieure. He had already published various textbooks and a translation (1827) of Giambattista Vico's *Scienza nuova* ("New Science"). The French Revolution of 1830 confirmed Vico's influence on Michelet in stressing man's own part in the making of history, conceived as a continuous struggle of human freedom against fatality. This, the main theme of the *Introduction à l'histoire universelle* (1831), was to underlie all of Michelet's later writings.

After the *Histoire romaine* (2 vol., 1831), Michelet devoted himself to medieval and modern history; his appointment as head of the historical section of the Record Office in the same year provided him with unique resources for carrying out his monumental life's work, the *Histoire de France*. The first six volumes (1833-43) stop at the end of the Middle Ages; they include the "Tableau de la France," in which the emergence of France as a nation is seen as a victory over racial and geographical determinism; they also include his treatment of Joan of Arc as the very soul of France and the living symbol of his own patriotic and democratic ideals. Michelet deliberately threw his intimate self into his narrative, convinced that this was the way to achieve the historian's ultimate aim: the resurrection (or re-creation) of the past. Such a resurrection must be integral: all the elements of the past—artistic, religious, economic, as well as political—must be brought back, intertwined, as they once were, in a living synthesis. Arbitrary and overambitious as the undertaking seems, Michelet's compas-

Treatment of Joan of Arc



Michelet, oil painting by Thomas Couture (1815–79). In the Musée Carnavalet, Paris.
Giraudon

sionate genius and romantic imagination enabled him to conjure up an effective evocation, unsurpassed for poetic and dramatic power.

Toward the end of this period of life, which was marked by private crises reflected in his work (the death of his first wife, in 1839, and the death of his friend Mme Dumesnil, in 1842, cast their shadows over whole periods of his *History of France*), Michelet turned away from Christianity and began to profess a messianic belief in democratic progress. His increasing hostility to the church, expressed in his lectures at the Collège de France (where he had held the chair of history since 1838) and in *Duprêtre, de la femme et de la famille* (1845), brought him into conflict with the Jesuits and caused his lectures to be suspended in January 1848. A month later, the revolution that he had heralded in *Le Peuple* (1846) seemed to bring about the realization of his dreams. But they were soon shattered: in 1852 Michelet, having refused allegiance to the Second Empire, was deprived of his posts. In 1847 he had interrupted the sequence of the *Histoire de France* to write the *Histoire de la Revolution française* (7 vol., 1847–53). He visualized the French Revolution as a climax, as the triumph of *la Justice* over *la Grâce* (by which he meant both Christian dogma and the arbitrary power of the monarchy). These volumes, written at a feverish pace, are a vivid and impassioned chronicle.

Hatred of priests and kings

Michelet then resumed the *Histoire de France* from the Renaissance to the eve of the revolution (11 vol., 1855–67). Unfortunately, his hatred of priests and kings, his hasty or abusive treatment of documents, his mania for symbolic interpretation, continually distort these volumes into hallucinations or nightmares. Also thus distorted is *La Sorcière* (1862), an apology for witches considered as godforsaken souls, victims of the antinatural interdictions of the church. A new and happier inspiration produced a series of books on nature: *L'Oiseau* (1856); *L'Insecte* (1858); *La Mer* (1861); *La Montagne* (1868). They reflect the influence of his second marriage to Athénaïs Mialaret, 30 years his junior, in 1849; written in a lyrical vein, they contain some of the most beautiful pages of a supreme prose writer. *L'Amour* (1858) and *La Femme* (1860), written under the same influence, are erotic and didactic divagations. *La Bible de l'humanité* (1864) is the story of man's quest for God, a story immensely enriched by the 19th-century philological and archaeological discoveries in the Oriental field.

The Franco-German War of 1870 shattered Michelet's idealism and his illusions about Germany. He died at Hyères on February 9, 1874. His widow tampered with his diaries, and their publication as a whole was begun only in 1959 (*Journal*, vol. 1, 1959, vol. 2, 1962; *Écrits de jeunesse*, 1959). They record his travels through Europe, and, above all, they give a key to his personality and illuminate the relationship between his intimate experiences and his work.

BIBLIOGRAPHY. G.J.J. MONOD, *La Vie et la pensée de Jules Michelet, 1798–1852*, 2 vol. (1923), a classic biography that unfortunately does not go beyond 1852; G.P. GOOCH, *History and Historians in the Nineteenth Century*, 2nd ed., pp. 168–177 (1952), a short but excellent assessment, with an outdated bibliography; J. GUEHENNO, *L'Évangile éternel: Étude sur Michelet* (1927), a sympathetic view by a liberal critic; R. BARTHES (ed.), *Michelet par lui-même* (1954), an anthology with a brilliant essay on Michelet's obsessions; P. VIALLANEIX, *La Voie royale* (1959), a scholarly and valuable essay on Michelet's notion of "the people," with a full chronology and bibliography.

(J.J.Se.)

Michelson, A.A.

Albert Abraham Michelson, the first American to receive a Nobel Prize in science, is best known for his determination of the speed of light and the Michelson-Morley studies of the Earth's motion in relation to surrounding space, work that made an important contribution to Albert Einstein's theory of relativity.

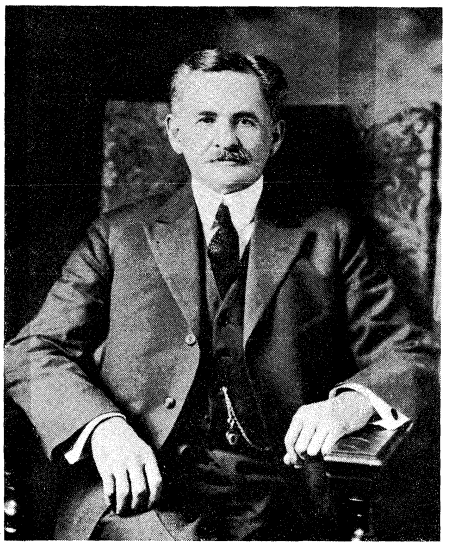
Michelson was born in Strelno, Prussia (now Strzelno, Poland), December 19, 1852. His parents brought him to the United States when he was two years old. From New York, the family made its way to Virginia City, Nevada, and San Francisco, where the elder Michelson prospered as a merchant.

At 17, Michelson entered the United States Naval Academy at Annapolis, Maryland, where he did well in science but was rather below average in seamanship. He was graduated in 1873, then served as science instructor at the academy from 1875 until 1879. In 1878 Michelson began work on what was to be the passion of his life, the accurate measurement of the speed of light. He was able to get useful values with homemade apparatus. Feeling the need to study optics before he could be qualified to make real progress, he travelled to Europe in 1880 and spent two years in Berlin, Heidelberg, and Paris, resigning from the navy in 1881. Upon his return to the United States, he determined the velocity of light to be 299,853 kilometres per second, a value that remained the best for a generation. When it was bettered, Michelson bettered it.

While in Europe, Michelson began constructing an interferometer, a device designed to split a beam of light in two, send the parts along perpendicular paths, then bring them back together. If the light waves had, in the interim, fallen out of step, interference fringes of alternating light and dark bands would be obtained. From the width and number of those fringes, unprecedently delicate measurements could be made, comparing the velocity of light rays travelling at right angles to each other.

It was Michelson's intention to use the interferometer to measure the Earth's velocity against the "ether" that was then thought to make up the basic substratum of the universe. If the Earth were travelling through the light-conducting ether, then the speed of the light travelling in the same direction would be expected to be equal to the velocity of light plus the velocity of the Earth, whereas the speed of light travelling at right angles to the Earth's path would be expected to travel only at the velocity of light. His earliest experiments in Berlin showed no interference fringes, however, which seemed to signify that there was no difference in the speed of the light rays, and, therefore, no Earth motion relative to the ether.

In 1883 he accepted a position as professor of physics at the Case School of Applied Science in Cleveland, Ohio, and there concentrated his efforts on improving the delicacy of his interferometer experiment. By 1887, with the



Michelson.
By courtesy of the University of Chicago

help of his colleague, American chemist Edward Williams Morley, he was ready to announce the results of what has since come to be called the Michelson-Morley experiment. Those results were still negative; there were no interference fringes and apparently no motion of the Earth relative to the ether.

Michelson-Morley experiment

It was perhaps the most significant negative experiment in the history of science. In terms of classical Newtonian physics, the results were paradoxical. Evidently, the speed of light plus any other added velocity was still equal only to the speed of light. To explain the result of the Michelson-Morley experiment, physics had to be recast on a new and more refined foundation, something that resulted, eventually, in Albert Einstein's formulation of the theory of relativity in 1905.

In 1892 Michelson, after serving as professor of physics at Clark University at Worcester, Massachusetts, from 1889, was appointed professor and the first head of the department of physics at the newly organized University of Chicago, a position he held until his retirement in 1929. From 1923 to 1927, he served as president of the National Academy of Sciences. In 1907 he became the first American ever to receive a Nobel Prize in the sciences, for his spectroscopic and meteorological investigations, the first of many honours he was to receive.

Michelson advocated using some particular wavelength of light as a standard of distance (a suggestion generally accepted in 1960) and, in 1893, measured the standard metre in terms of the red light emitted by heated cadmium. His interferometer made it possible for him to determine the width of heavenly objects by matching the light rays from the two sides and noting the interference fringes that resulted. In 1920, using a 20-foot interferometer attached to a 100-inch telescope, he succeeded in measuring the diameter of the star Betelgeuse (Alpha Orionis) as 240,000,000 miles (300 times the diameter of the Sun). This was the first substantially accurate determination of the size of a star.

In 1923 Michelson returned to the problem of the accurate measurement of the velocity of light. In the California mountains, he surveyed a 22-mile pathway between two mountain peaks, determining the distance to an accuracy of less than one inch. He made use of a special eight-sided revolving mirror and obtained a value of 299,798 kilometres per second for the velocity of light. To refine matters further, he made use of a long, evacuated tube through which a light beam was reflected back and forth until it had travelled 10 miles through a vacuum. Michelson died in Pasadena, California, on May 9, 1931, before the results of his final tests could be evaluated, but in 1933 the final figure was announced as 299,774 kilometres per second, a value less than 2 kilometres per second higher than the value accepted in the 1970s.

BIBLIOGRAPHY. The only full-length study of Michelson is BERNARD JAFFE, *Michelson and the Speed of Light* (1960).

(I.A.)

Michigan

Since its admission in 1837 as the 26th state of the Union and the fourth to be carved from the old Northwest Territory, Michigan has become a mainspring in the nation's economic life; the name of its largest city, Detroit, has become a byword throughout the world for the American automotive industry. The state also has retained its prominence in agriculture, and, because of its many inland lakes, its borders on four of the five Great Lakes, and its many wilderness tracts, has evolved into one of America's leading tourist regions. A young and rapidly increasing population, kept in the state by ample employment opportunities and wage scales well above the national average, has helped Michigan to become a symbol of the dynamic growth of the United States as a whole.

Michigan is the only one of the 49 continental states to be split into two large land segments: the sparsely populated but mineral-rich Upper Peninsula, which slices eastward from northern Wisconsin between Lakes Superior and Michigan, and the mitten-shaped Lower Peninsula, which reaches northward from Indiana and Ohio. The two landmasses have been joined since 1957 by "Big Mac," the five-mile-long Mackinac Bridge across the Straits of Mackinac, which separate Lake Michigan to the west from Lake Huron to the east. Between Lake Huron and Lake Erie, in the southeast, the Lower Peninsula is separated from the Canadian province of Ontario by Lake St. Clair and the St. Clair and Detroit rivers. The state's name is derived from a Chippewa Indian word meaning, approximately, "large lake." Though its 58,216 square miles (150,779 square kilometres) rank it only 23rd nationally in size, the inclusion of Great Lakes waters over which it has jurisdiction would raise the figure to 96,720 square miles (250,504 square kilometres), placing it 11th.

Nearly 75 percent of the 8,875,083 Michiganders counted in the census of 1970 are urbanized, heavily concentrated in the industrialized centres of the Lower Peninsula. This factor, together with a broad array of ethnic and national stocks among the people and a constant influx of lesser skilled workers attracted by the union-dominated labour scene, has created in many cities the typical marks of economic progress and poverty existing side by side and a sometimes tenuous social stability. The state government, in Lansing, coordinates a vast network of programs attempting to reduce these contrasts, and it has provided a system of public higher education that is among the most diversified and renowned in the nation. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; DETROIT; and GREAT LAKES.)

THE HISTORY OF MICHIGAN

In the 17th century, the Indian population of present-day Michigan was estimated at between 12,000 and 15,000. The majority, including the Ottawa, Chippewa (Ojibwa), Miami, and Potawatomi, belonged to the Algonkian linguistic group or family. A lesser number, primarily in southeastern Michigan were Hurons and Wyandots. The Ottawa and Chippewa, conditioned extensively by their Great Lakes environment, were of great assistance to the French in the development and expansion of the fur trade. The Ottawa, with their primitive commercial interests, had developed a type of canoe that was highly serviceable in the Great Lakes area. The Miami and Potawatomi were identified more with the culture of the woods. The Hurons were the most advanced in their agricultural practices. All of the Indians lived in small communities and lacked the concept of territorial boundaries.

French heritage

Settlement. Étienne Brulé, the first European to visit the area (1622), was the forerunner of numerous explorers, missionaries, and fur traders who paved the way for French control over Michigan. The oldest community in Michigan is Sault Ste. Marie, founded in 1668 at a site

where in 1641 missionaries held services for 2,000 Chippewas. In 1701 Antoine de la Mothe Cadillac established Detroit as a fur-trading centre and administrative post; it soon became the leading French community in the entire Great Lakes area.

In 1760 the French garrisons were surrendered to an English force, and by the Treaty of Paris, in 1763, England acquired jurisdiction over Canada and the French empire east of the Mississippi River except New Orleans. Under English rule, Michigan remained a part of Canada, and a handful of English-speaking traders helped to give Detroit and other communities a more English character. During the Revolution, Detroit was a major supply centre for British troops who raided the Kentucky country until 1779, when the British general Henry Hamilton was captured.

American territory. Though Michigan had been awarded to the United States in 1783, the British refused to leave Detroit and other major posts until 1796. In 1787 it was made a part of the newly created Northwest Territory. Indian opposition to American rule in the area was ended by the victory of Anthony Wayne at the Battle of Fallen Timbers, near present-day Toledo, Ohio. After 1796 the Americanizing of the regions was accomplished within a few years. Detroit became the capital of the Michigan Territory, which was separated in 1805 from Indiana. Although the first governor, William Hull, surrendered Detroit to the British early in the War of 1812, American rule was restored by the victory of Oliver Perry at the Battle of Lake Erie, late in 1813.

Process of Americanization

The real growth of Michigan Territory began soon after the war. The personality of the new governor, Lewis Cass, and the encouragement that he gave to settlement helped to promote its growth. New modes of transportation were even more significant. In 1818 steamship navigation linked Detroit and Buffalo, New York, inaugurating a new era in lake transportation. Cass's crude highway chain from Detroit to Chicago, Saginaw, and Port Huron helped to establish the patterns of settlement in the interior. Completion of the Erie Canal in 1825 made Detroit a leading distribution point for settlers seeking new homes in the Great Lakes area.

Statehood and growth. Michigan was anxious for statehood so that it might undertake a more ambitious program of internal improvements. The first constitution was drafted in 1835, but admission was delayed until 1837 by the Toledo War, a boundary dispute with Ohio. For relinquishing its claims to the mouth of the Maumee River, at Toledo, Michigan was awarded the western half of the Upper Peninsula as well as the eastern portion, which was historically part of the territory.

The state grew rapidly through the 1840s and 1850s. Thousands of prospective agricultural settlers, including many foreign-born, established new homes in the state. Detroit and other leading cities profited, and in the 1840s the rich iron and copper resources of the Upper Peninsula became known. Tension over the slavery issue resulted in the formation of the present-day Republican Party at Jackson in July 1854. The new party captured the Michigan delegation to Congress and immediately made impressive nationwide gains.

Throughout the Civil War, Michigan made major contributions, losing some 14,000 of its 90,000 men who served. A Negro regiment from Michigan included enlistees from many states and Ontario. Political life saw the Republican Party become dominant after the war, and in the 1890s Hazen Pingree implemented progressive legislation as mayor of Detroit and subsequently as governor.

The 20th century. Before 1900 all of Michigan's 83 counties had been settled. Agriculture, lumbering, mining, and manufacturing created a stirring economic tempo. Throughout the 20th century the economy has been modified greatly and often dominated by the automotive industry, which has contributed further to the rapid growth of Detroit, Flint, and other cities. During World War I, industrial production at all levels was intensified. The emergence of new problems connected with this urban–industrial growth was recognized by features of the new constitution approved in 1908.

Increasing economic tempo

The 1920s were characterized by more economic expansion and a new urban sophistication, as both state and nation continued to lose their rural character and values, and urban crime rates began to rise. The Depression was unusually severe in Michigan, the industrial products of which were not among the necessities of life. Unemployment and deflation were far above national averages: early in 1933, one-third of Detroit's labour force was unemployed.

In 1932 Michigan departed from the Republican fold, and thereafter it became one of the doubtful, or swing, states. Organized labour became a powerful political and economic factor during the 1930s. In 1937 the United Automobile Workers (UAW) became the bargaining agent for production workers at General Motors Corporation and by the outbreak of World War II it was the dominant union in all automotive plants. During the war, Detroit became known as the "Arsenal of Democracy," and industrial production continued at a peak afterward to restock the nation with new cars and other war-depleted consumer goods.

In Detroit the inner-city population reached its peak during the early 1950s. An influx of blacks to war plants created social tensions that brought on a major race riot in 1943, and in 1967 a major racial disturbance resulted largely in burning and looting. Other cities had less severe disorders in following years. Labour tension continued to be part of the Michigan scene, with large-scale UAW strikes invariably being felt throughout the national economy.

THE NATURAL AND HUMAN LANDSCAPE

The physical environment. The Great Lakes cool the hot winds of summer and warm the cold winds of winter, giving Michigan a milder climate than some other north central states. Though the Upper Peninsula is cooler, difference in temperature range between far northern and far southern cities is not excessive. Northern Sault Ste. Marie has average maximum temperatures of about 50° F (10° C) and average minimums of 30° F (−1° C). Detroit, in the southeast, has average maximums of about 58° F (14° C) and minimums of 42° F (6° C). The greatest amount of moisture is received in southern Michigan, with the state's average about 33 inches (838 millimetres). The central portion has less precipitation than does the Upper Peninsula, and the coastal strip along Lake Michigan receives an unusually large snowfall from westerly storm fronts moving across the lake. The growing season ranges from approximately three months in the Upper Peninsula to six months in the southern portions of the Lower Peninsula.

Climatic ranges

The mildly rolling country appealed to the early agricultural settlers. Even the elevations are not high. The highest point in the Lower Peninsula, near Cadillac, rises only about 1,200 feet (370 metres). Elevations approximate 2,000 feet (600 metres) only in the western Upper Peninsula.

Michigan's rivers, the majority of them in the southern portions of the Lower Peninsula, drain vast interior areas. With few exceptions, they are rather shallow and not wide. Their navigability and the ease of bridging them were factors that encouraged settlement. Several of the rivers, especially in the Upper Peninsula, have sufficient falls to serve as a basis for water power.

Rivers, lakes, and islands

Michigan's 11,000 inland lakes range from a few acres in size to the 19,000 acres (7,700 hectares) of Houghton Lake in the north central part of the Lower Peninsula. The shores of many lakes are ringed with summer cottages, as are the shores of the Great Lakes in many places.

About 500 islands dot the lakes and rivers. Belle Isle, a public recreation centre, and Grosse Ile, largely residential, are well-known features of the Detroit River. Mackinac Island became a resort on which motor vehicles are prohibited. Isle Royale, a virgin wilderness of nearly 850 square miles, is a national park in northern Lake Superior.

Several physical features of Michigan are unusually appealing to lovers of the scenic. The dunes on the Lake Michigan shore are used annually by thousands of vaca-

tioners, and the 3,000-odd state and local parks, containing more than 4,100,000 acres of wooded land, present varied landscapes that have helped Michigan to become a major tourist attraction of the Middle West.

Animals native to the area are numerous. Whitefish in abundance swim in the Great Lakes, and many streams contain various edible trout. The Conservation Department operates hatcheries and encourages tourism around the inland lakes, where carp and bass abound. The beaver was sought eagerly by early traders, and other fur-bearing mammals were also found in large numbers. Deer, as well as quail and ducks, remain numerous in many counties. Some 80 out of 83 Michigan counties were once heavily wooded, with genuine prairies or clearings found only in southwestern Michigan. Hardwood timbers included the hickory, ash, oak, and hemlock, though the white and Norway pine were the most common timber in the north.

Regions. Understandably, many Michiganders think of their state as divided into the Upper and Lower Peninsula, but the two physiographic divisions, the Superior Upland and the Central Lowland, follow this plan only in part. The Upland comprises only the rugged western region of the Upper Peninsula, where the abundance of copper and iron ores has made the area economically dependent on the extraction industries. Calumet, Hancock, Houghton, and Marquette are among the better known cities of the area. Agriculture has been of little significance, but tourist and recreational potentialities offer some hope of diversifying the economy.

Geographical divisions

Geographical as well as historical forces unite the eastern counties of the Upper Peninsula and the Lower Peninsula north of a Bay City–Muskegon corridor to create the popularly known "northern Michigan." Here the white-pine forests were ruthlessly exploited. Some agriculture is successful, for even the sandy soil, with adequate moisture, can produce an abundant yield of potatoes and grain. A fruit-growing region near Traverse City has succeeded because Lake Michigan's influence prevents early killing frosts. The towns serve as regional centres for tourists and farmers and often as headquarters for governmental services. Many of the larger communities have attracted small-scale manufacturing establishments. Thriving cities include Sault Ste. Marie, Petoskey, Ludington, Manistee, and Cadillac.

South of the Bay City–Muskegon corridor a fertile clay soil and a longer growing season permit a wider variety of crops, especially grains. The southern agricultural counties, resembling those of northern Ohio and Indiana, produce much of the state's agricultural wealth. The region also contains the great industrial concentrations, and large cities are numerous. As in Detroit, the fifth largest of America's cities, the automotive industry played an important role in the growth of Flint and Lansing. Grand Rapids attained prominence through its furniture industry, and Battle Creek is known nationally for its cereals. The majority of the cities, however, have more widely diversified industries than their reputations suggest.

THE PEOPLE OF MICHIGAN

The diverse backgrounds of Michigan's people, in nationality, religion, and ethnic stock, has been important in shaping the state's character. By the early 1970s, however, the percentage of foreign-born and first-generation citizens was small, even in Detroit.

National and ethnic groups. Germans, the most numerous of the early non-English-speaking immigrants, settled in both rural and urban areas. Detroit had a large German community by the mid-1830s, as did several rural counties by 1850. The large Irish population remained basically urban, although Irish farmers were found in southern Michigan and by 1860 in the Upper Peninsula. Other groups left an impact, though their numbers were not large. Dutch influences are still observable in western counties around Holland, where Dutch settlers pioneered successfully in 1847. Finns have been important in the economic and cultural life of the Upper Peninsula. Early Polish immigrants settled in rural areas until the 1890s, when a large number of Poles concentrated in Detroit. The city's present-day population has a large percentage of persons of Polish ancestry.

The most significant population phenomenon of the 20th century has been the growth of the black population from fewer than 16,000 in 1900 to just under 1,000,000 in 1970. Two-thirds live in Detroit, which had become nearly 45 percent black by 1970. Throughout the 1960s, blacks moved to suburbs that had long-established black neighbourhoods and substantial black populations. They also moved into older neighbourhoods where the prices of homes were not exceptionally high. The trend was similar to that in the central city, where blacks replaced whites. The majority of the newer Detroit suburbs, however, have very insignificant black populations, since much of the price of housing is above what many blacks could afford. Many Detroit blacks expressed a preference for the pattern of life that has evolved in the inner city. Among the Detroit suburbs with majority or large minority black populations were Highland Park and Inkster, 55 percent, and Pontiac, more than 25 percent. In other areas, Flint and Muskegon had sizable black communities, and the smaller city of Benton Harbor became nearly 60 percent black.

The growing black population

Religious affiliations. Michigan's religious history differs somewhat from that of many Middle Western states. Since the first European settlers in Detroit were Roman Catholic, many immigrants of that faith were attracted to the city even before the large Irish, Italian, and Polish immigrations of the 19th century. Detroit was created a diocese in 1833 and an archdiocese in 1937. Other dioceses were established at Marquette, Grand Rapids, Lansing, Saginaw, Gaylord, and Kalamazoo. Of the Protestant denominations, Lutheranism has had many German and Scandinavian adherents, while Methodism has been important in both rural and urban Michigan. The Baptist Church has a large membership throughout the state. The first Dutch settlers were members of the Christian Reformed Church, which was opposed to the state church of the Dutch. The Michigan group was highly independent and conservative in doctrine and social mores. Altogether, Michigan has almost 200 Protestant denominations, some with only small memberships. The first Jewish immigrants were of German background. In 1851 Detroit Jews founded a synagogue. Synagogues throughout the state today reflect all shades of Jewish theological persuasion.

Demography. The natural increase in population—the preponderance of births over deaths—has been important in Michigan's growth. Both birthrates and death rates declined during the 1960s, though they remained near national averages. The death rate among blacks is approximately 30 percent higher than among whites. All projections indicate continued population growth in future decades. The state's population is younger than the national average, suggesting a continuing preponderance of births, and the far greater immigration than emigration should also help to maintain the trend.

The major increases in population are indicated for the metropolitan areas. Some demographers suggest that by 2010 the population of the six-county Detroit area will reach 7,500,000. This urban growth has been continuous, from less than 8 percent of the state's population in 1850 to nearly 75 percent in 1970. As in most of the nation, suburbs have grown at the expense of the inner cities. The population of Detroit proper, 1,845,000 in 1950, had declined by 1970 to less than 1,575,000. Among Detroit's suburbs having unusual growth were Warren, which leaped from 727 in 1950 to nearly 180,000 in 1970, and Livonia, which grew from less than 18,000 to more than 110,000 in the same period.

Urban population growth

THE STATE'S ECONOMY

Economic overview. Manufacturing has been the continuing basis of Michigan's economic strength, ranking it sixth among the states in 1969 in value added by manufacturing. In that year nearly 1,200,000 persons were employed in manufacturing, about 970,000 of them in the production of durable goods. The high wages paid by the automotive industry help to explain the state's mag-

netic appeal in attracting workers from other states, for they have tended to increase wages for the 2,500,000 Michiganders employed in other areas of the economy.

Natural resources

With new methods of iron-ore processing, Michigan's huge mineral reserves should be exploitable through the 20th century. Some deposits of gypsum for cement have not been touched, and the continuing search for petroleum and natural gas promises rich new fields. The fertile soils will continue to make Michigan a major agricultural state. New power sites have added to the industrial potential, though in the early 1970s accommodation was being sought between conservationists and advocates of an extensive network of thermonuclear generating plants that might add to the growing evidences of air and water pollution in the Great Lakes as well as in the inland streams and lakes.

Michigan's tourist industry, which ranks as the state's second largest industry, is dependent largely upon these water resources threatened by pollution. Some 18,000,000 people take recreational trips in Michigan annually, spending about $1,300,000,000. Tourism employs about 155,000 persons, and another 400,000 receive indirect employment from it.

The future of Michigan's agriculture is dependent in part on prices. Its farms average 153 acres in size, in comparison with a national average of 378 in 1969. Many farmers work parttime in industry or for the government, supplementing their income and providing an important factor in the economy of many rural communities.

The automotive industry. Many reasons have been advanced for the rise of the automotive industry in Detroit. The city long had been noted for the manufacture of carriages, wagons, bicycles, and marine engines. The large number of skilled and semiskilled labourers was another important factor. Other cities, however, offered inducements equal to those of Detroit, and in fact, during the pioneering phase of the industry, Detroit had a number of rivals.

Unquestionably, personalities played a major role in making Detroit the world's automotive centre. The industry began with Ransom E. Olds of Lansing, whose father manufactured gasoline engines. Subcontractors found themselves pioneers in this new and profitable industry. Olds's success by 1901 focussed the attention of automotive people on Detroit. Henry Ford brought even greater fame. Organizing the Ford Motor Company in 1903, he was by 1908 confining production to the standard, low-priced Model T. He emphasized ease of repair, garage service, and the utility of his product. In that same year W.C. Durant of Flint recognized that the automobile would be purchased by persons who desired transportation rather than by faddist motorists. He hoped through the formation of a company with large-scale capital to speed up technological advances and thus capture a large portion of the untapped middle class market. The General Motors Corporation stands as a testimony to his thinking. The successful Michigan manufacturers succeeded because their models were designed for a national market, whereas many companies had models that could not be operated in mountainous country.

The industry paid high wages, and many automotive workers became members of the middle class. Because of the high wages, there was little interest in unionism until the Great Depression of the 1930s.

Government and industry. The state government encourages a climate favourable to economic expansion. Many sections of government maintain research divisions that offer assistance to various segments of the economy. The Department of Agriculture administers a division of animal health, a division of marketing, and a division of soil conservation. The state tax structure has been evaluated as one that attempts to avoid discrimination against the private sector of the economy. For the labour sector, Michigan has had a workmen's compensation law since 1912. Extensive programs of social legislation, including unemployment compensation, were undertaken even before World War II and have since been broadened.

Transportation. The first railroad in Michigan, the Erie and Kalamazoo, was completed between Toledo and Adrian in 1836, and by 1870 the state had more than 1,600 miles of rail. The peak of 9,000 miles in 1910 was reduced to 6,300 by the time the semipublic Amtrak system took over passenger service in 1971.

In turn, the gains in motor-vehicle ownership have been staggering, reaching more than 4,600,000 registrations in 1971. State and local governments have combined to give Michigan a modern system of state highways, county roads, and city streets. The interstate express system, built largely with federal assistance, will comprise 1,175 miles when the initial phases are completed.

Air services in Detroit

Air passenger service in Michigan began in 1926 with schedules between Detroit and Grand Rapids. In 1968 Detroit Metropolitan Wayne County Airport accounted for 80 percent of Michigan's 9,000,000 airline passengers, and there are 11 other major airports throughout the state and several small commuter air lines. Much of the air freight of metropolitan Detroit is handled at the Willow Run Airport, constructed as a bomber site during World War II.

The waterways, the sole basis of travel for the first pioneers, still carry tremendous tonnage, with many of the state's exports shipped from Detroit harbour. Ores and other bulk materials destined for the interstate trade are ordinarily sent by water.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. In 1967 Michigan had more than 2,900 local governmental units including counties, municipalities, townships, school districts, and such special districts as park authorities. The basis of this complex system is the new constitution of Michigan, ratified by the voters in 1963. The new document included a number of innovations designed to streamline the government and to make it more responsive to the problems of a modern industrial and urban society. Amendments may be submitted to the electorate by the legislature or by initiative petitions, but all amendments must be approved by a referendum of the voters.

Executive power is vested in the governor, who serves for four years. He is nominated by a primary election, but the lieutenant governor is chosen by party convention. The governor appoints a large number of administrative commissions that are responsible to him and several advisory commissions. The majority of the important governmental services are combined under 20 departments also responsible to the governor. A few of these bodies, such as the State Highway Commission, must be bipartisan.

The legislative branch comprises a Senate of 38 members elected for four-year terms and a House of Representatives of 110 members elected for two-year terms. The legislative districts are redefined by a special bipartisan commission after each federal census. The highest court is a seven-member state supreme court. The state also has a court of appeals, circuit court, probate court, and courts of limited jurisdiction that are specified by the legislature.

Local government

The organization of units of local government is very flexible. Though the majority of counties are governed by a board of supervisors, the home-rule privilege allows larger counties to entrust management to county commissioners. Extensive privileges of home rule are authorized for the cities as well. School districts are classified by population and enjoy differing privileges of government.

Politics. The precinct is the primary unit in the structure of political party organization, and the precinct delegates carry considerable importance in the annual party conventions. These conventions nominate the candidates for lieutenant governor, attorney general, and the members of the boards that govern the state system of higher education and the Board of Education. They also nominate justices of the supreme court, who, however, do not bear a partisan label on the ballot. The convention also selects delegates to national presidential conventions.

A new interest in politics has resulted from the civil rights movement and from the formation of coalitions among college students. The civil rights groups have been responsible for a new awareness of politics on the part of

the black voters. Blacks have been nominated for major state offices by both parties. Unions have been very active in Michigan politics, and the UAW has endorsed candidates at the municipal, state, and national levels.

Finance and social services. The state sales tax is the major source of revenue, though receipts from personal and corporate income taxes play major roles in assuring the financial structure. Gasoline and vehicle taxes are reserved exclusively for highways.

Michigan's annual budget of well over $1,500,000,000 reflects in its allocations the concern with the well-being of its citizens. A mental institution at Kalamazoo that received patients as early as 1859 was a forerunner of a number of other institutions caring for the unfortunate. Since 1936 many of the state welfare services have followed a federal formula in transmitting aid to units of local government. Payments to recipients of old-age assistance, for example, and aid to dependent children, the blind, and the disabled have required special personnel at the state level, though disbursal of most funds is done locally.

Education. Michigan's major governmental service lies in the field of education. Many of the legislature's major appropriation bills are for support of the numerous public institutions of higher learning and the many community colleges. Large sums, however, are returned to the local districts to enable public elementary and secondary schools to maintain a high standard. Many of these schools, both in the inner-city areas and in the suburbs, are under frequent study for the identification of problems, and experiments in improved methods of instruction are often conducted by educators and psychologists from the universities.

Higher education

The greatest of the monuments to Michigan's cultural heritage, combining the traditional and the popular, are its some 90 colleges and universities, both public and private. In 1817 Judge Augustus Woodward, one of the major figures in the state's early history, conceived of the idea of a "Catholepistemiad," an academy of universal knowledge. This was achieved in some measure in 1841, when the University of Michigan opened its doors in Ann Arbor. During the 20th century it has come to be widely regarded as one of the nation's, and perhaps the world's, finest universities, with programs on both the undergraduate and graduate levels. In 1852 a teacher-training institution began instruction at Ypsilanti, and in 1857 Michigan Agricultural College, now Michigan State University, was formally established in East Lansing. The latter shares much of the wide regard of its rival in Ann Arbor, having moved far beyond its early identification with agriculture in many areas of research and scholarship.

The several teacher-training institutions became state universities during the 20th century, as did the Michigan Technological University at Houghton. In 1956 the state acquired Wayne University, a Detroit municipal university, founded in 1868. As Wayne State University, it has fostered much educational experimentation, and in the 1960s its campus and physical plant have become landmarks in American educational architecture through the designs of the Japanese-American architect Minoru Yamasaki.

Michigan is also the home of widely recognized specialized schools. In 1927 the School of Music was founded in Interlochen, the forerunner of the National Music Camp that now offers instruction in music, dramatics, and related arts. The Cranbrook and Kingswood schools, in Bloomfield Hills, designed by the famous Finnish-born architect Eliel Saarinen, pioneered advanced art courses for students of high-school age.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu

Detroit dominated the cultural life of early Michigan. Many of its citizens had the leisure and wealth that allowed them to devote time to cultural pursuits, and the multifacetted backgrounds of its population gave it a more cosmopolitan atmosphere. It received the state's first travelling theatrical companies, and an opera house was erected before the Civil War. In 1819 the Young Men's Society was organized by Lewis Cass, who led many early civic endeavours, to promote debates, lectures, and general intellectual life.

The pioneer farmer, however, had little time for cultural pursuits. Often he respected only the more elementary aspects of cultural life that might have utilitarian value —the community dance that eased the strains of plowing or harvesting, the county fairs that provided a ritual summation of the year's achievements. Thus Michigan's culture, perhaps like that of the nation as a whole, tended to become broadly based in popular life, having many of the aspects of "mass culture." In the major parks the city of Detroit promoted band concerts and, later, symphony concerts to bring better music to thousands. The Detroit Institute of Arts, maintained municipally, has always emphasized exhibits that appeal to a broad public.

Cultural commemorations. With the emphasis upon popular culture, Michigan increasingly is commemorating its past. The Michigan Historical Commission has designated sites of historic importance, such as the location of battles and the first home of the Ford Motor Company. Many cities emphasize Homecoming Day, usually the anniversary date of the incorporation of the community. Local folk festivals have been given a greater emphasis. Holland's Tulip Festival, held each May, has become an event of more than local importance. The annual Bavarian Festival, in Frankenmuth, appeals to others besides those of German background. Traverse City sponsors a popular Cherry Festival, and the Lake Huron cities commemorate the lumbering era with their highly publicized Paul Bunyan Days. Many ethnic groups in Detroit and other cities sponsor folk festivals that recall their cultural ties to Europe. The International Afro-American Museum in Detroit is a continuing reminder to the black population of its distinguished heritage extending far back into pre-American centuries.

Folk and regional festivals

Museums and libraries. The Detroit Institute of Arts, which can be traced to a private museum established in 1889, holds one of the nation's major collections. In 1969 it was the recipient of the Robert Hudson Tannahill collection, which included five Picassos, two van Goghs, four Cézannes, three Renoirs, and two Degas, as well as works by Rouault, Modigliani, Rousseau, Manet, Chardin, and Corot. The institute also has noted Baroque, African, and German Impressionist collections. The Hackley Art Gallery in Muskegon was founded in 1911 by a wealthy philanthropist, and the Kalamazoo Institute of Arts and the Grand Rapids Art Museum are among other art museums that have won recognition. Lansing is the home of the Michigan Historical Museum, famous for its military and Indian collections, while many county museums commemorate local history.

The state library was founded in 1828. A state library board was created in 1936, but major public libraries are found only in the larger cities. Among the outstanding special libraries are the William L. Clements Library at the University of Michigan, a depository for records pertaining to the American Revolution, and the Burton Historical Collection of the Detroit Public Library, specializing in local history and genealogy.

Communications. Michigan's press dates from the one issue of *The Essay*, published by a priest in 1809, and from the weekly *Detroit Gazette* founded in 1817. Today's *Detroit Free Press* became the state's first daily in 1833 as the *Free Press and Intelligence*. Journalism was a wildly competitive activity by 1900, when both Kalamazoo and Grand Rapids had three daily newspapers. By the early 1970s, however, the *Detroit Free Press* was the only surviving morning daily in the state, along with about 50 evening dailies. Weeklies and semiweeklies number about 150, while neighbourhood and trade journals have increased in number. The foreign press has been important: in 1870 Detroit had two German dailies, but the majority of ethnic newspapers were weekly.

Broadcasting in Michigan began in 1920 with broadcasts of local Detroit news, and today there are about 190 AM and FM radio stations and about 20 commercial television stations. These are supplemented by several noncommercial radio and television outlets, usually sponsored by educational organizations.

PROSPECTS

Both immigration and the normal fertility of the young population seem to assure Michigan of a continued increase in population, especially in the metropolitan areas. The major bases for further economic expansion probably will lie in the service industries and in government employment. Increases in income and employment in tourist and recreational areas can be expected as well. Employment increases in manufacturing, however, will be nominal, and it is possible that employment in agriculture and in mining industries will register a slight decline.

The agricultural areas in Michigan are losing their distinctive character, with some 50,000 acres being swallowed up annually by suburban or exurban residential developments. The development of recreational areas in the north is also responsible for the decline of farming acreage, though the long-term results of this may prove positive.

Major social gains can be anticipated also from a state that has been in the vanguard in seeking the well-being of its citizens: public welfare programs and all levels of education will become increasingly important to Michiganders. Since such services are related to income, however, taxation will continue to be a major issue. Slight increases are anticipated from higher rates on liquor and racing, though the largest increases will come from the sales and personal- and corporate-income taxes.

Social progress may well be marred by further racial disturbances. The disorders of the larger cities have reached down to the smaller cities. Short-term disruptions reflecting these problems can be expected also in the schools. The electorate unquestionably will be called upon, through the referendum, to vote on numerous issues affecting the social and political life of the state.

BIBLIOGRAPHY. Good general histories include: F. CLEVER BALD, *Michigan in Four Centuries*, rev. ed. (1961); WILLIS F. DUNBAR, *Michigan: A History of the Wolverine State* (1971); and MILO M. QUAIFE and SIDNEY GLAZER, *Michigan: From Primitive Wilderness to Industrial Commonwealth* (1948). More specialized studies of state history are: FLOYD RUSSELL DAIN, *Education in the Wilderness* (1968); WILLIS F. DUNBAR, *All Aboard: A History of Railroading in Michigan* (1969); GEORGE N. FULLER, *Economic and Social Beginnings of Michigan: A Study of the Settlement of the Lower Peninsula During the Territorial Period, 1805–1837* (1916); CALVIN GOODRICH, *The First Michigan Frontier* (1940); and ANGUS MURDOCH, *Boom Copper* (1943), a history of the mining counties. Interesting studies of the Michiganders include: RICHARD M. DORSON (ed.), *Negro Folktales in Michigan* (1956); W. VERNON KINIETZ, *The Indians of the Western Great Lakes, 1615–1760* (1940); and C. WARREN VANDERHILL, *Settling the Great Lakes Frontier: Immigration to Michigan, 1837–1924* (1970).

(S.Gl.)

Microbiology

The 17th-century discovery that living forms exist that are invisible to the naked eye was a dramatic one in man's history, for, from the 13th century onward, it had been postulated that "invisible" organisms were responsible for decay and disease. The word microbe was coined in the latter quarter of the 19th century to describe these organisms, all of which were thought to be related. As microbiology eventually developed into a separate science, it was found that microbes comprise a very large group of extremely diverse organisms; thus, microbiology became subdivided into various disciplines—*e.g.*, bacteriology, protozoology, and virology. The diversity of microbes, or micro-organisms as they are now commonly called, has meant that it is almost impossible for one person to be knowledgeable in all of the disciplines grouped under microbiology.

Microbiology involves the identification of micro-organisms and the study of their structure and function. It encompasses the study of bacteria, rickettsiae, small fungi (*e.g.*, yeasts and molds), algae, and protozoans, as well as problematical forms of life such as viruses. Because of the difficulty of assigning plant or animal status to micro-organisms—some are plantlike, others animallike—they are sometimes considered a separate group called protists. Microbes can also be divided into procaryotes, which have a primitive and dispersed kind of nuclear material—*i.e.*, the blue-green algae, bacteria, and rickettsiae—and eucaryotes, which display a distinct nucleus bounded by a membrane. Such are small algae other than the blue-greens, yeasts and molds, and protozoans. (All higher organisms are eucaryotes.)

The biological status of micro-organisms

Man's daily life is interwoven inextricably with micro-organisms. They abound in the soil, in the seas, and in the air. Everywhere abundant, although usually unnoticed, micro-organisms provide ample evidence of their presence, sometimes unfavourably, as when they cause decay of objects valued by man or generate disease, and sometimes favourably, as when they ferment alcohol to wine and beer, raise bread, flavour cheeses, and create other dairy products from milk. Micro-organisms are of incalculable value in nature, causing the disintegration of animal and plant remains and converting them to gases and minerals that can be recycled in other organisms.

HISTORICAL BACKGROUND

Beginnings. Microbiology can be said to have begun with the development of the microscope. Although others may have seen microbes before him, Antonie van Leeuwenhoek, a Dutch draper whose hobby was lens grinding and microscope making, was the first to provide proper descriptions of his observations, which included protozoans from the guts of animals and bacteria from teeth scrapings. His descriptions and drawings were excellent because his lenses were of an exceptional quality. Leeuwenhoek conveyed his findings in a convincing and logical manner in a series of letters to the British Royal Society during the mid-1670s. Although his observations stimulated much interest, no one made a serious attempt either to repeat or to extend them. Leeuwenhoek's "animalcules," as he called them, thus remained mere oddities of nature to the scientists of his day, and enthusiasm for the study of microbes gained ground slowly. It was only later, during the 18th-century revival of a long-standing controversy about whether or not life can develop out of nonliving material, that the significance of micro-organisms in the scheme of nature and in the health and welfare of man became evident.

The early Greeks believed that living things could originate from nonliving matter; their goddess Gea was credited with creating life from stones. Although Aristotle discarded this notion, he still held that animals could arise spontaneously from other unlike organisms or from soil. His influence regarding this concept of spontaneous generation was still felt as late as the 17th century. Toward the end of the 17th century, a chain of observations, experiments, and arguments began that dealt a deathblow to the idea that life could be generated from nonlife. It was an involved series of events, with the forces of personality and strong wills often obscuring the facts.

Although Francesco Redi, an Italian naturalist, disproved that higher forms of life could originate spontaneously, proponents of the concept claimed that microbes were different and did indeed arise in this way. Such illustrious names as John Needham, Lazzaro Spallanzani, Franz Schultze, and Theodor Schwann, among numerous others, figured in the debates.

It remained for Louis Pasteur to settle the matter. He proved in a series of masterful experiments that only pre-existing microbes could give rise to other microbes —at least under current earthly conditions (see also LIFE; that life arose spontaneously from nonlife at some earlier time, under appropriate physical and chemical conditions, is an undisputed postulate of chemical evolution; see EVOLUTION).

Microbes and disease

Regarding microbes and disease, Girolamo Fracastoro, an Italian scholar, advanced the notion as early as the mid-1500s that contagion is an infection that passes from one thing to another. The "thing" that is passed along eluded discovery until the late 1800s, when the work of many scientists, Pasteur foremost among them, determined the role of bacteria in fermentation and disease. Robert Koch, a German physician, defined the procedure for proving that a specific organism causes a specific disease.

The age of discovery. The foundation of microbiology was securely laid during the period from about 1880 to 1900. The students of Pasteur, Koch, and others discovered in rapid succession a host of bacteria capable of causing specific diseases (pathogens) and elaborated an extensive armamentarium of techniques and laboratory procedures for revealing the ubiquity, diversity, and power of microbes.

All of these developments occurred in Europe. Not until the early 1900s did microbiology become established in America. Many of the microbiologists who worked in America at this time either had studied under Koch or at the Pasteur Institute, in Paris. All microbiologists of the early 20th century, however, were influenced by such men as Koch. Once established in America, microbiology flourished, especially with regard to such related disciplines as biochemistry and genetics.

Contemporary advances. Since the 1940s, microbiology has experienced an extremely productive period, during which many disease-causing microbes have been identified and methods to control them have been developed. Micro-organisms have also been effectively utilized in industry; their activities have been channelled so that valuable products of commerce and agricultural benefits result.

The study of micro-organisms has also advanced man's knowledge of living things. Microbes provide easy-to-work with material for studying the complex processes of life; *e.g.*, metabolism. Correlated with the intensive probing into the functions of microbes have been numerous, and often unexpected, dividends that can be applied to solving existing problems. Knowledge of the basic metabolism of a pathogenic bacterium, for example, often leads to a means for controlling the pathogen. Nutritional requirements of bacteria thus may be of value in combatting an infection.

THE SCOPE OF MICROBIOLOGY

The study of bacteria, which were among the first objects of microbiological study, is called bacteriology. Various subdisciplines deal exclusively with particular micro-organisms.

From another standpoint microbiology can be subdivided into theoretical, or pure, microbiology, and practical, or applied, microbiology. The latter can be further subdivided according to specialties, such as medical, industrial, agricultural, food, and dairy microbiology (see below *Applied microbiology*).

Interdisciplinary nature. The science of microbiology has been influenced by, and in turn has influenced, other sciences. Micro-organisms are no longer the exclusive concern of microbiologists. Biochemists, geneticists, cytologists, and molecular biologists have discovered the value of microbes as experimental tools in the study of such fundamental biological processes as metabolism, photosynthesis, enzyme action, gene action, and population dynamics. Micro-organisms are well suited to such uses; they represent a vast range of metabolic types, and genetic changes are correlated with their rapid proliferation. In addition, they can subsist on relatively simple inorganic nutrients and, because they multiply rapidly, are available in extremely large numbers in a relatively short period of time. They are also relatively easy to maintain and handle under laboratory conditions.

The content of microbiology. *Bacteriology.* Modern and accurate knowledge of the forms of bacteria dates from the researches of the German botanist F.J. Cohn, the chief results of which were published at various periods between 1853 and 1892. Cohn's classification of bacteria, published in 1872 and extended in 1875, dominated the study of these organisms thereafter. While various observers added to the knowledge of the structure of bacteria, others laid the foundation of what is known about the relations of bacteria to fermentation and disease. When Pasteur showed in 1857 that lactic acid fermentation depends upon the presence of an organism, it was already known that fermentation and putrefaction are intimately connected with the presence of organisms in the air. In 1862 Pasteur placed beyond reasonable doubt the fact that production of ammonia by the fermentation of urea is caused by the action of a minute bacterium, named in 1874 *Micrococcus ureae.*

Isolation of disease-causing bacteria

After the introduction of bacteriological techniques (see below *The study of microbiology: Cultivation of micro-organisms*) came the isolation of many bacteria. It was discovered in 1882, for example, that a bacillus (a rod-shaped bacterium) is the cause of glanders, a disease of horses. In 1883 Koch isolated the organism of Asiatic cholera, and the same year that of diphtheria was found. In 1885 the tetanus bacillus was observed in pus produced in mice and rabbits inoculated with soil; only in 1889, however, did the Japanese bacteriologist S. Kitasato discover the way in which the organisms could be cultivated (they grow only in the absence of oxygen). W.D. Miller, a U.S. dentist, studied the micro-organisms of the human mouth in the 1880s, noting their possible relationship to the decay of teeth.

Pasteur and his associates found that animals vaccinated with a specially cultivated anthrax bacillus showed immunity to disease when reinoculated with the deadly wild form. These findings were destined to lead to a study of the principles of immunity, which underlie the prevention and treatment of disease by vaccines and immune serum. Questions relating to causes and to the nature of changes occurring both in the bacteria and in the host, as well as the development of immunity in the latter, continue to be subjects of great interest and importance.

While investigations on infectious diseases and immunity were under way, it became apparent that other activities of bacteria also are of importance to man. In 1878, only two years after Koch announced the discovery of the anthrax bacillus, the U.S. botanist T.J. Burrill discovered the bacterial cause of fire blight in pears, thereby establishing that certain plant diseases are caused by bacteria. The importance of some bacteria in soil and their contribution to soil fertility was recognized in the 1880s and 1890s. The significance of bacterial activities in many aspects of the dairy industry was recognized at about the same time.

The further application of bacterial activities to industrial processes, other than early studies on alcoholic and lactic acid fermentations, began later. Thus, within the span of a few decades the study of bacteria had progressed enormously, and by the early 20th century it was recognized that the activities of bacteria bear an intimate relation to many aspects of human activity.

The original interest of bacteriologists centred upon what bacteria do. This interest has broadened to include the study of the bacteria themselves: what they are, their relationships to each other, and their relationships to other organisms.

Protozoology and others. The name Protozoa is derived from the Greek words meaning "first animal," and, indeed, protozoans are sometimes considered the most simple of all animals. First observed by Leeuwenhoek in the 17th century, protozoans are almost as ubiquitous as bacteria. The study of protozoans at one time centred on the parasites that cause malaria and sleeping sickness, thereby stimulating research in tropical medicine. Improved laboratory techniques for cultivating this diverse group of organisms have made them valuable tools in many types of investigations—from physiological studies, such as cytoplasmic motion in *Amoeba*, to ecological studies involving their role in the food chain of the oceans.

The term fungi, Latin for "mushroom," applies not only to mushrooms but to all of the large and diverse group of plantlike organisms to which the mushrooms belong. The fungi of most interest in microbiology are the yeasts and molds. Many molds are studied because of their economic importance. *Mucor*, for example, not only causes spoilage of foods such as vegetables and fruits but also is used to manufacture some cheeses.

Lichens and slime molds

Some molds form unique relationships with other organisms; lichens, for instance, are composed of both algae and fungi. The unusual nature of lichens was not discovered until after the invention of the microscope. The name lichen appeared long before, however, in the

writings of Theophrastus, a disciple of Aristotle, about the 4th century BC. At that time, and for many centuries thereafter, lichens were confused with mosses and the hepatics. Early students thought that the green cells, now known to be an alga, were specialized fungal structures. Not until 1867 were the green cells in lichens indentified as algae; at the same time it was announced that the fungal cells were parasitic upon them. The introduction of pure culture techniques permitted the lichen constituents to be isolated and grown apart from each other. In 1873 the German botanist H.A. De Bary suggested the name symbiosis to describe the mutual benefit between the components of lichens.

Once objects of interest for their curious form, slime molds—as their individual nutritional, environmental, and genetic characteristics have become better understood—have become objects useful in teaching and research. The creeping sheet (plasmodium), or vegetative stage of a slime mold, has the characteristics of a primitive animal and resembles a primitive amoeba. The reproductive stage (sporangium) has the characteristics of a mold. The double life of these organisms is reflected in the name Mycetozoa, meaning fungus animals, which was coined by De Bary in 1858.

Algae, a heterogeneous group of primitive organisms, are of great interest to all biologists because they are capable of photosynthesis and are of evolutionary interest as well. Intensive research is currently directed toward the small forms making up much of the free-floating microscopic life in water (phytoplankton) because they provide food for aquatic animals and thus are of great importance. Many freshwater algae produce undesirable tastes and odours, and studies have been aimed at eliminating them from domestic water supplies. Although algae occasionally form mats on water surfaces, sometimes causing suffocation of aquatic life, heavy growth of certain species also has been found to reduce water hardness and to remove the salts found in brackish water, thereby making the water more suitable for human consumption. Algae are also being studied for their potential value as a food source for man.

Virology and others. Viruses comprise a heterogeneous assemblage of self-reproducing agents, smaller than the microscopically visible bacteria, that multiply only within living susceptible cells. Certain diseases caused by viruses have been known since the late 1700s, when the British physician Edward Jenner developed a vaccine from material isolated from cowpox lesions. He did not see the causative agent. Pasteur, in the mid-1850s, developed an attenuated strain (*i.e.*, one made less virulent) of the virus that causes rabies, but he was not then aware of the viral nature of the disease. An associate of Pasteur, Charles Chamberland, discovered that bacteria would not pass through a porcelain filter but that the causative agent of rabies did. The term filter-passing, or filterable, virus was used to describe these agents.

The Russian bacteriologist D. Ivanovski in 1892 found that a filtrate of sap from tobacco plants infected with mosaic disease could be used to transfer infection to healthy plants; the Dutch microbiologist M.W. Beirjerinck later confirmed the work. The further finding in 1902 that the agent of foot-and-mouth disease of cattle is also filterable made clear the fact that the agents cause disease in animals as well as plants. By 1930 the term filterable had been dropped, and virus is routinely used by microbiologists as a name for these agents.

Nobel Prize winner Max Theiler found in 1951 a method for attenuating virulent yellow fever virus; the technique has since been modified to produce vaccines against other viral diseases. Thus, viruses have been studied intensively in recent years not only because they provide important information about life processes but also because vaccines against them are constantly sought.

The study of viruses

The study of rickettsiae, which resemble very small bacteria but grow only in susceptible cells, is intimately associated with those forms causing human disease. Elucidation of the microbes causing such diseases as epidemic typhus, Rocky Mountain spotted fever, and scrub typhus began in 1906 with the work of Howard Taylor Ricketts, for whom the organisms are named; he described the organism of spotted fever. In 1910 Ricketts and an associate described the organism of typhus fever, which in 1916, was named *Rickettsia prowazekii* in honour of Ricketts and Stanislas von Prowazek, both of whom died of the disease. Research continues as better methods for prevention, control, and treatment of rickettsial diseases are sought.

Subjects of current interest. One of the more current studies involving micro-organisms is their possible occurrence in outer space and on planets other than Earth. A branch of exobiology, space microbiology, includes the investigation of microbes as providers of food and oxygen in the closed environment of spaceships.

A less positive development of microbiology has been biological warfare—the selection and cultivation of microbes as weapons of war, to cause disease or injury to domestic plants, animals, and man.

Although most of the subdisciplines of microbiology are directed at the occurrence of microbes in specific environments, the study of gnotobiotics is concerned with the exclusion of microbes except those involved in any given experiment. Such germfree organisms are important research tools in investigating parasitic diseases, immunological processes, nutrition, stress, shock, and aging (see GERMFREE LIFE).

THE STUDY OF MICROBIOLOGY

The study of microbiology is channelled in two major directions. Pure microbiology—that is, the study of a particular group of micro-organisms in order to learn about their morphology, physiology, taxonomy, occurrence, variation, heredity, and evolution—seeks to understand the nature of microbes. Applied microbiology, on the other hand, is motivated by the desire to exploit the effects of micro-organisms that are of benefit to man and to control the activities of those that are harmful.

Bacteriology as a synonym for microbiology

The methods used to study bacteria have been widely adapted to the study of other micro-organisms; indeed, the techniques employed in the microbiological sciences probably have more in common than do the origin and evolution of the organisms themselves. For this reason, the term microbiology is often considered to be synonymous with bacteriology. Similarly, the term micro-organism no longer refers only to microscopic organisms. With the adoption of bacteriological methods for the study of many types of organisms, the meaning of the term micro-organism has gradually been extended to include any organism that can grow and be studied using the methods originally developed for bacteria. The study of microbiology, therefore, encompasses many organisms that are not microscopic (*e.g.*, molds) but usually does not include some that are (*e.g.*, rotifers).

Microscopy. That invisibly small organisms exist was believed from early times. Lucretius, who expressed an atomic view of matter, wrote about AD 75 that even the plague must be caused by a kind of atom. He considered the atoms, or seeds, as lifeless, however. The writings of Lucretius influenced 16th-century scientists, but the means for actually observing micro-organisms—a magnifying lens—was not developed until the 17th century by Galileo.

The convex lens of Leeuwenhoek allowed a greatly enlarged image to be seen, enabling him to see protozoans, filamentous fungi, and yeasts. In 1676, using his simple lens, which was capable of magnifying 280 times, he first saw bacteria. A few years before Leeuwenhoek's observation of bacteria, Robert Hooke, the father of the cell theory, had observed filamentous fungi (1667) through a compound microscope—one with two lenses, which gives a larger image than that given by a simple lens.

By 1786 various micro-organisms, including bacteria, had been viewed through the compound microscope. The compound microscope of today is widely used; it differs from those of the 18th century mainly in the addition of a device, called a condenser, for lighting the object being viewed with a wide cone of light. A condenser is necessary to provide sufficient light when large magnifications are used.

Types of light microscopy used in microbiology

Of the types of microscopy now used in microbiology, most utilize light microscopes; magnification is obtained by a system of optical lenses. The types of light microscopes include the commonly used bright field, in which the background is brightly lighted, the objects studied are dark, and the power of magnification is about 1,000; the dark field, in which the background is dark and the objects studied are bright, a phenomenon particularly useful for examining unstained organisms suspended in fluid; the ultraviolet, the greatest advantage of which is that magnifications two to three times greater than those obtained with the light microscope can be achieved because ultraviolet light has a shorter wavelength than visible light; and phase contrast, in which controlled illumination is obtained by the use of special equipment that enables the refraction, or bending, of light passing from one material to another to be seen, thereby revealing differences in cells not discernible with other microscopic methods. Fluorescence microscopy utilizes the ultraviolet microscope; a chemical substance with the property to absorb ultraviolet waves and emit visible ones is used with a mixture of micro-organisms, some of which take up the substance, and thus can be distinguished from those that do not. Still another type of microscopy, electron microscopy, allows very great magnification; waves of electrons (negatively charged particles), rather than light, and magnetic fields, rather than lenses, are used to achieve an image. Although electron microscopy has the advantage of great magnification, it also has several limitations, most important of which is that the material must be dry. This not only eliminates the possibility of studying living specimens but also raises the possibility that the drying process may alter the characteristics of the specimen.

Cultivation of micro-organisms. *Sterile culture.* Micro-organisms growing on a nutrient medium are referred to as a culture. Early experiments (1776) by Spallanzani, begun as the result of the previously mentioned controversy over whether or not life could develop out of lifeless matter, laid the foundation for the technique of sterile culture. This involves first freeing a suitable medium of micro-organisms by heating and, second, keeping the medium sterile—*i.e.*, keeping micro-organisms out of the medium. In his experiments, Spallanzani boiled various kinds of seeds in a flask and stoppered it. After a few days, many organisms could be found in the flask; Spallanzani distinguished the larger ones, which were destroyed by boiling for one-half minute, and microbes, which survived boiling and developed even after the flask had been sealed. Eventually, he discovered that, after boiling sealed flasks for as long as 45 minutes, no micro-organisms developed.

The objection then voiced to Spallanzani's work—that the quality of the air in the flasks had, by heating, been rendered unable to support life—was overcome later (1836) when air was passed into the flask after having been slowly drawn through solutions containing sulfuric acid. The meat extract in the flask remained uncontaminated because the micro-organisms in the air were killed by its passage through the solutions. In 1853 it was discovered that flasks did not have to be sealed after boiling but could be closed with a plug of cotton, which effectively filtered the incoming air; this procedure is still used. Because many micro-organisms can produce bodies (spores) that are extremely heat-resistant, mediums for the growth of micro-organisms are usually sterilized by heating under pressure to 120° C (250° F) for 15 or 20 minutes.

Simplification of the method for sealing flasks

Dilution method and pure culture techniques. The Scottish surgeon Joseph Lister contributed to culture methods in microbiology with the introduction in 1878 of the dilution method. During studies probably concerning the souring of milk, he used sterile water to dilute a small amount of milk, then diluted some of this mixture with more sterile water. He continued the dilution process until a sample of the milk and sterile water mixture no longer caused the milk to sour. The last dilution that resulted in souring thus contained a minimum number of organisms, and Lister considered it a pure culture—that is, containing only one species of organism. The dilution method has become an essential part of pure-culture methods.

The dilution method was modified in 1896 by a Danish microbiologist, Emil Hansen, who studied yeasts. He added one drop of a yeast culture to the first of a series of small drops of sterile water, then removed a drop of the culture and added it to the second drop of water. Eventually, he obtained a drop containing just one yeast cell.

Koch also made a significant contribution to culture techniques with his development of a simple method for obtaining pure cultures of bacteria. He added a solidifying agent (gelatin) to a nutrient medium. After the medium had been heat sterilized and partially cooled, he added some micro-organisms. By cooling the solution further and spreading the organisms before the gelatin solidified, he was able to isolate the micro-organisms from each other, with the result that each organism gave rise to a separate colony, or crop of cells.

The introduction of agar-agar

In 1883 a woman in Koch's laboratory, Frau Hesse, further improved the technique by substituting agar-agar for the gelatin. Unlike gelatin, agar-agar can be liquefied by only a few micro-organisms and does not provide a food source, thus allowing better control of the nutrient content of the medium. Silica gel has also been used as a solidifying agent.

Enrichment cultures. One final advance in culture technique was developed in the 1890s by the Dutch microbiologist M.W. Beirjerinck and by the Russian microbiologist Serge Winogradsky. They selected a medium that favours the growth of one organism over another. If an organism, one from the soil, for example, is capable of utilizing a certain substance (*e.g.*, a specific sugar), growth of the organism can be induced by using the substance in the medium. Only the organisms capable of attacking the substance will grow in great numbers. Successive transfers of a relatively small number of organisms (inoculum) to a fresh medium will eventually result in a culture strongly enriched with the organism that utilizes the desired substance.

The maintenance of pure cultures. After a species of micro-organism has been obtained as a pure culture, it is necessary to maintain it alive and as a pure culture. Microbiology laboratories in schools, universities, and industry usually maintain collections of pure cultures of the particular species they use. Various organizations throughout the world maintain pure cultures of micro-organisms; such collections are important in that they make available pure cultures of species when needed.

The Kral Collection in Prague, established in 1900, was the first known culture collection. The names and locations of a few of the other collections are: the American Type Culture Collection (Rockville, Maryland), the Japanese Type Culture Collection (Tokyo), and the (British) National Collection of Type Cultures (London). Other collections also exist worldwide to serve particular needs. A section on culture collections was established in 1966 by the International Association of Microbiological Societies in order to promote the exchange of information among culture collections.

Whenever a microbiologist proposes a new species, he provides one or more of the national culture collections with a pure culture of the species.

Staining techniques. Until the latter third of the 19th century, micro-organisms were observed only in the natural state; the similarity of their refractive index to water and their small size, however, made them difficult to see when viewed through a microscope. Thus, the discovery of a method that would allow them to be seen more easily was an important development.

In 1875 the German pathologist Carl Weigert, using techniques developed more than a decade earlier for staining animal tissues with dyes, found that dead ("fixed") bacteria would become heavily stained with the dye picrocarmine. Other dyes were then introduced—*e.g.*, methylene blue, fuchsin, and crystal violet. Hans Christian Gram, a Danish bacteriologist, discovered in 1884 a simple procedure for placing bacteria into either a gram-negative class or a gram-positive class, depending

on whether or not the organisms retained crystal violet when treated in a specific way. Although the gram stain is used mostly with bacteria, other micro-organisms also show a reaction. Yeasts and actinomycetes, for example, are gram-positive; rickettsiae are gram-negative. Stained preparations now are used particularly to observe structural features of micro-organisms.

Other methods. Twentieth-century developments in microbiological methods have depended largely upon the techniques of other disciplines—namely, biochemistry, physiology, organic chemistry, and physics.

The gas exchange of respiring micro-organisms, called the respiratory quotient (the ratio of carbon dioxide produced to oxygen consumed), is measured by following the pressure change of respiring organisms in a closed vessel. Other processes—*e.g.*, fermentation—can also be studied using this method.

The transfer of hydrogen ions (positively charged atoms) in micro-organisms has been studied by adding to the organisms or extracts prepared from them a substance (*e.g.*, the dye methylene blue) the colour of which changes when hydrogenated. Many other dyes behave in this way.

Techniques for breaking apart micro-organisms

The techniques used to break up micro-organisms include such biochemical techniques as alternate freezing and thawing, prolonged grinding, vigorous shaking with glass beads, and exposure to supersonic vibration; certain micro-organisms (*e.g.*, yeast) are especially difficult to disrupt because of their resistant cell walls.

The apparatus of modern physics has greatly aided the advance of virology. Around 1930 the steady development of new technical methods for studying the physical, chemical, and biological properties of micro-organisms completely changed the outlook of biologists toward viruses. An important advance in this new approach was the development of graded filter membranes of known pore size. This created for the first time the opportunity to assess the actual size of virus particles.

Another important advance of the 1930s was the growth of the virus of fowl pox in the tissues of a developing chick embryo. Methods for the growth in the laboratory of rickettsiae are similar in that living host cells are necessary. In 1949, tissue culture methods, long used in half-hearted fashion for the cultivation of viruses, were shown to be suitable for the growth of poliomyelitis viruses. This not only opened the way to immunization against polio but supplied a method by which many new types of virus could be isolated (see also TISSUE CULTURE).

APPLIED MICROBIOLOGY

Practical applications of microbiology began to develop soon after the first sterile cultures were prepared. In the 18th century, a Parisian cook applied Spallanzani's technique of sealing and heating flasks to jars containing fruits and vegetables. His discovery that they were preserved indefinitely not only made him wealthy but prepared the way for the canning industry.

Applications in industry and medicine. *Industrial microbiology.* Many substances are produced by micro-organisms on a large scale. Bacteria, for example, are used in the commercial production of lactic acid, derivatives of which are used as solvents and to treat calcium deficiency and anemia. Vinegar is formed by bacteria, as are amino acids, the latter being used as supplements for foodstuffs.

The ethyl alcohol produced by yeasts from carbohydrates is important to brewers of malt beverages, wine makers, distillers, and others; and yeasts were used in baking even before Roman times. Attempts have been made to cultivate many types of micro-organisms as food; yeasts are attractive because they are easy to grow and have a high nutritive value. Research is active in this area because yeasts can be grown as by-products of industry—*e.g.*, in sulfite liquor resulting from paper manufacture and in dairy wastes—thus reducing the problems associated with waste disposal and pollution.

Molds are used in the production of industrial chemicals; for example, citric acid, which is used in medicines, flavourings, extracts, foods, dyeing, and the manufacture of ink. Molds are also used to produce antibiotics, such as penicillin, which was the first one to be produced commercially, and enzymes. (For detailed discussions of these processes see BAKING AND BAKERY PRODUCTS; BREWING; WINE MAKING; PHARMACEUTICALS, PRODUCTION OF.)

Food microbiology. The preservation of food is an important aspect of applied microbiology. Foods consumed by man are largely perishable—that is, they provide nutrient sources for the growth of many micro-organisms; *e.g.*, bacteria, yeasts, molds. Some of these organisms cause chemical changes in food that make it unsuitable for human consumption; others can cause poisoning if the contaminated food is eaten.

Undesirable changes caused by micro-organisms in food

In ancient times, salting, drying, and smoking were practiced to preserve food. Modern techniques of food preservation involve refinements of the primitive methods and the utilization of newer ones: heating, low temperature, dehydration, and treatment with chemicals and radiation. Such techniques either prevent or remove contamination, prevent microbial growth or metabolism, or kill micro-organisms. (For details of methods of food preservation see FOOD PRESERVATION.)

Dairy microbiology. Milk, which has been called the "most nearly perfect" food, not only provides great nutritional value for man but also is an excellent medium for the growth of many micro-organisms. In fact, sterile skim milk is commonly used as a growth medium for many bacteria.

Milk from the cow, although essentially free of micro-organisms, is usually contaminated as it is processed. The pasteurization process—the heating of milk at a specific temperature for a specific amount of time—greatly reduces the number of bacteria in milk. If milk is not properly stored, micro-organisms can affect it by producing acid and gases such as carbon dioxide and hydrogen; by causing the texture to become viscous; and by causing curdling, abnormal flavours, odours, and colours.

Micro-organisms are of great value in the dairy industry, however. Fermented milks (*e.g.*, buttermilk), butter, and cheeses are produced as a result of microbial activity (see DAIRYING AND DAIRY PRODUCTS).

Medical microbiology. Many micro-organisms, especially species of bacteria, protozoans, and viruses, cause infections in man, other animals, and plants. The control of disease-causing organisms is thus of great importance, regarding both economic losses and the well-being of plant and animal life.

Commonly used methods of combatting plant diseases include the breeding of disease-resistant plants, sterilization of the soil, and crop rotation. Certain infectious diseases of animals, called zoonoses, are of special importance because they can be spread from infected animals to man. Some methods used to control animal diseases are similar to those used to control human diseases—*e.g.*, immunization, adequate diet and sanitation, and prevention of transmission by quarantine of infected animals (see DISEASES OF ANIMALS; DISEASES OF PLANTS; INFECTIOUS DISEASES).

Microbiology of water, soil, and air. Since 1854, when waters polluted with man's domestic and industrial wastes (sewage) were proved to have caused an epidemic of cholera, bacteriological studies of water have been concerned with detecting, identifying, controlling, and eliminating disease-causing species. Although water-purification systems have been very effective in protecting man against polluted waters, pollution problems have increased as the human population has grown (see SEWAGE SYSTEMS). Another aspect of water microbiology involves the microbes that occur in natural bodies of water—*e.g.*, ponds, lakes, rivers, and oceans; their enormous importance in the economy of nature is increasingly recognized.

Microbes mediate the conversion of organic matter in the soil to nutrients that can be utilized by plants. The condition of soil in a given area is therefore of great significance in agriculture.

Micro-organisms do not grow in air, but the particulate matter in air—*e.g.*, dust and water droplets—may be laden with them, the numbers and types being determined by

the contaminating source. The control of airborne pathogens is of great significance not only in the spread of disease but also in circumstances in which sterile conditions are required, such as in certain industrial processes.

MICROBIOLOGICAL SOCIETIES AND PUBLICATIONS

Certain microbiological societies publish several journals related to specific aspects of microbiology. The American Society for Microbiology, for example, publishes *Journal of Bacteriology; Bacteriological Reviews; Journal of Virology; Applied Microbiology;* and *Antimicrobial Agents and Chemotherapy*.

Some societies publish journals of a general nature—*e.g.*, the Society for General Microbiology in Great Britain publishes the *Journal of General Microbiology*. Societies that publish journals relating to a specific group of micro-organisms include the Mycological Society of America (*Mycologia*); the British Mycological Society (*British Mycological Society Transactions*); and the Society of Protozoologists (*Journal of Protozoology*).

Other publications relating to microbiology include the *Japanese Journal of Microbiology*, published by the Japanese Bacteriological Society and the Society of Japanese Virologists and the *Canadian Journal of Microbiology*. *Mikrobiologiya* is published in the U.S.S.R.; the *Institut Pasteur Bulletin* is published in France; and *Acta Pathologica et Microbiologica Scandinavica* is published in Denmark. *The Journal of General and Applied Microbiology* is also published in Japan.

BIBLIOGRAPHY. G.C. AINSWORTH, and P.H.A. SNEATH (eds.), *Microbial Classification* (1962), a series of essays about the problems of classifying micro-organisms, includes names and addresses of institutions maintaining culture collections; T.D. BROCK (ed. and trans.), *Milestones in Microbiology* (1961), a collection of original papers important in the history of microbiology; ALBERT DELAUNAY *et al.* (eds.), *The World of Microbes*, vol. 4 of the *Encyclopedia of the Life Sciences* (1965), a well-illustrated account of microbial activities in relation to man; C.L. DUDDINGTON, *Micro-organisms As Allies* (1961), an interesting account of the industrial uses of bacteria and fungi; R.K. JENNINGS and R.F. ACKER, *The Protistan Kingdom* (1970), an informative and entertaining account of protists and viruses; H.A. LECHEVALIER and MORRIS SOLOTOROVSKY, *Three Centuries of Microbiology* (1965), a reconstruction of the main aspects of the development of microbiology; JOHN POSTGATE, *Microbes and Man* (1969), an introduction to the importance of micro-organisms to life on Earth; THEODOR ROSEBURY, *Life on Man* (1969), a witty and informative book.

(Ed.)

Microclimates

Though "climate" normally refers to the general, large-scale conditions of the atmosphere, it is often important to consider quite small areas and to distinguish between sharply contrasting climates at points close together. The term microclimate is used to describe this finer structure. In particular, within a very few feet above and below the earth's surface and within the canopies of vegetation, the temperature, humidity, and wind undergo much stronger gradients and stronger patterns of cyclical behaviour and variability than anywhere else. Within the height of a man, temperature differences of 10° C or more are commonplace; horizontally finding such climatic differences may involve a journey of continental dimensions.

Microclimate is strongly influenced by soil and vegetation, by slope and orientation, and by exposure. On a somewhat larger scale (the effects, for instance, of a small range of hills on the temperature and wind fields around them), the term mesoclimate is employed. Climate in general is described as macroclimate. Various other terms are also used, but their definitions are not universal.

An insect that chooses to live on the underside of a boulder or a cabbage leaf thus identifies its microhabitat. All of the different microhabitats encompassed by a given crop and soil type, including the variety of conditions within the soil itself, constitute the micro-environment. The complex of microclimatic factors as they involve individual plants or animals, including modifications due to the presence of those plants or animals, constitutes the bioclimate or proprioclimate.

Marine and terrestrial microclimates

Although any small species may tolerate only a limited range of bioclimate, the close proximity of strongly contrasting microclimates provides a total environment within which different animal and plant species can coexist and interact. The complexities of microclimate, therefore, constitute a requirement for the variety of life forms. The greater variety of terrestrial life, as compared with marine forms, follows not so much from the greater equability of the marine climate, although this is important, as it does from the greater variability of land surfaces and, hence, far wider range of microclimatic contrasts. These overlap the main climatic variations in space and time, though the overall tendencies remain; microclimates, for example, are generally coldest in winter and high latitudes. Furthermore, there is strong correlation between the type of macroclimate and the range of microclimates within its compass. Clear skies and dry conditions, as in desert climates, engender the strongest microclimatic contrasts in temperature. In permanently moist and overcast conditions, these contrasts may almost disappear.

Plants occupy their preferred microhabitats through processes of natural selection. The majority of animals seek theirs out by movement, although many forms do achieve marked modification of microclimates by nests, shelters, and other artifices. In all cases, the preferred conditions may vary with season and with the maturity of the plant or animal. Chief of the modifiers is man, who lacks great microclimatic tolerances and thus adjusts his shade and shelter and his house and clothing to suit his comfort, convenience, and efficiency.

MICROCLIMATIC CONDITIONS

Temperature and moisture. The interface between the air and the ground is heated by short-wave radiation received from the sun and scattered from the sky. The interface emits radiation, as do the gases of the atmosphere, in the long or terrestrial wavelengths. The time pattern of radiation processes is dominated by annual and diurnal cycles. In both cycles, the extremes of temperature are found at or very close to the surface, at or shortly after the times of extreme (net) radiation.

Heating and cooling propagate upward and downward from the interface. The wave amplitude is progressively attenuated, and the times of occurrence of maxima and minima become progressively later, with increasing height and depth. The rate of decrease of air temperature with height is known as the lapse rate; when the air temperature increases with height, it is termed an inversion.

Temperature ranges

The profiles in Figure 1 illustrate typical temperature ranges in air and soil on a clear summer day in middle latitudes. The daily range at the ground–air interface may compare with and even exceed the annual macroclimatic range so that surface-dwelling insects can experience a year's extremes in a single day. The attenuation of the two cycles with depth is, however, very different, the annual penetrating about 20 times as deep as the diurnal so that significant annual variations of temperature in most soils will be found up to depths of about ten metres. The maximum variation is delayed about one month at one metre and by three or four months at five-metre depth. The minimum will be delayed similarly, but the delay will be longer at the smaller depths in climatic zones where snow is prevalent.

The penetration of the annual wave into the atmosphere is under macroclimatic rather than microclimatic control. For reduction in daily amplitude to 10 percent of surface value, 300 metres is a characteristic height, but this depends on wind and weather. The maximum surface temperature occurs soon after noon, some two hours later at one-metre height, three to five hours at 100 metres. By night the cooling is slow, but it continues until sunrise; thereafter the surface temperature may rise rapidly, and the warming will spread upward relatively quickly. Thus, there is a strong asymmetry in the diurnal march of temperature.

It is clear that strong temperature gradients will be found near the interface in both media. Generally the atmospheric temperature gradient decreases approximately in inverse proportion to the distance from the surface. The

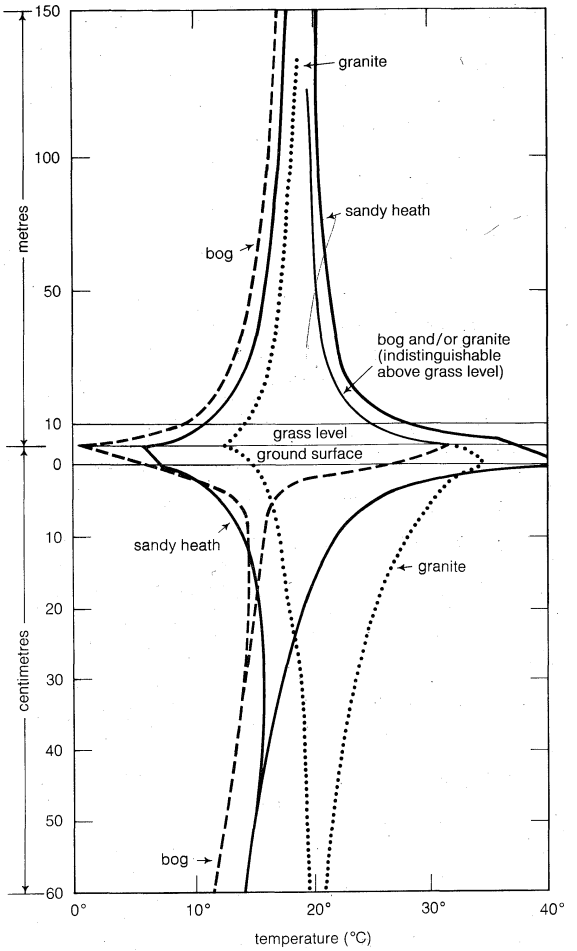


Figure 1: Maximum and minimum temperature curves, or heating and cooling paths, above and within the ground for different kinds of ground material.
From R. Geiger, *The Climate Near the Ground* (1965); Harvard University Press

change per unit distance in soil is much stronger, but the pattern of behaviour is harder to categorize.

The humidity (weight of water vapour per unit volume of air) behaves in much the same manner as temperature. Both the absolute value and the gradient decrease with height by day and increase at night. General statements about diurnal and annual moisture distributions in soil tend to be unreliable because recent rainfall and evaporation history exert a governing influence. Strong moisture gradients are found in partially wet soil, and moisture then migrates only slowly, mainly through distillation from warmer toward colder. Bare soils, once they have started to dry out, are thus defended against further daytime evaporation, but vegetation circumvents this defense by drawing water from depth and transpiring directly to the atmosphere.

Wind and turbulence. Friction reduces the wind to zero at a solid boundary and almost to zero where surface drift can occur, as on water surfaces. Near the ground, the average speed increases with height at a rate varying inversely with the height. Low-level winds tend to be stronger by day than by night.

Even a steady wind will show irregularities in strength and direction down to periods of seconds or less with corresponding fluctuations in temperature and other variables. This turbulence plays a vital role in microclimate. It controls the diffusion of heat and moisture in the air, the dispersal of seeds and spores, and of atmospheric pollutants.

Turbulence, especially in lighter winds, varies markedly with time of day and state of sky. By day, under clear skies, the air is heated from below, and the lighter air underlying heavier air is unstable and more turbulent. Nocturnal temperature inversions represent stable stratifications; heavier air underlies lighter air, and turbulence may virtually disappear. Very strong and irregular layering effects can then develop with marked layered contrasts of temperature, smoke content, etc. continuing throughout the night with layers sliding over each other with very little intermixing.

Dew and frost. When the surface temperature falls below the saturation level (temperature at which no additional moisture can be held), dew is deposited. This may represent true dewfall, a deposit from the vapour diffusing down through the air, or it may result from water diffusing upward through soil or plants and condensed at the coldest point. When this occurs at temperature below freezing, the result is a white frost or hoar frost. Black frost, by contrast, occurs when the water in the surface layers freezes *in situ*.

Dew and frost formation interrupt the cooling on clear nights because of the latent heat released as the water condenses or freezes. As the air warms up again, heat is required to evaporate or condense. The overall effect of this process is always to moderate the temperature and retard its changes.

Heat balance and evaporation. The temperature and moisture behaviour is dominantly controlled by the heat or energy budget at the interface. The interface possesses no heat capacity, and, as it absorbs daytime radiational energy, it must dispose of it by its own radiation, by conduction and convection, and by using energy for evaporation. Similarly, the heat radiated away at night must be conveyed to the surface. The daily radiation amounts are governed by macroclimate; that is, by latitude, season, and cloudiness. The temperature response is then conditioned by the heat capacity and effective conductivity of air and underlying medium.

Approximate quantitative rules may be given in terms of a combination of physical properties known as the admittance or conductive capacity of the medium. Table 1 illustrates the conductive properties of various soils and corresponding effective values of air and ocean. These latter may exhibit considerable variation because the effective conductivity is really dependent upon the degree of turbulence.

Surface absorption of heat

Heat absorbed at the surface will be shared between the atmosphere and the underlying medium in proportion to their conductive capacities. The magnitude of the temperature response will vary inversely as the sum of the two conductive capacities. Accordingly, the lapse rate is unstable by day and more heat will go into the air than into the ground. But on still nights, with stable stratification, most of the heat radiated is supplied by the ground. The difference between various underlying surfaces also is indicated by Table 1. The depth of penetration of surface temperature variations, whether periodic or irregular, varies as the square root of the conductivity. The more conductive the underlying material, the greater and more rapid the penetration and the smaller the temperature response.

The other major factor in the heat balance is evaporation. With soil saturated or vegetated with live growth, or over open water, energy is used in evaporating water at the interface. This often consumes the major fraction of the radiation energy available. By so reducing the energy available for heating, evaporation decreases the daily range in temperature. Over land, the daily march of evaporation follows closely the net radiation.

Over water, radiation is absorbed appreciably below the interface so that the heat balance is more complicated. Another important difference is that water, by virtue of its mobility, is effectively a much better conductor of heat than any soil. Thus, the daily cycles of temperature, and the temperature gradient, are much weaker in maritime than in continental environments. The same holds true for the annual cycle and the irregular variations from day to day.

FACTORS AFFECTING MICROCLIMATES

Soils and surfaces. The effects of different soil types and conditions on temperature and heat balance may be understood from the values in Table 1 and the principles of the preceding section. Sandy soils and other coarse,

Table 1: Conductive Properties of Air, Soils, and Other Media
(CGS units)

medium	new snow	dry sand	sandy clay	wet sand	ice	water*		air near ground*	
						stable, moderate currents	strong currents	stable	unstable
Conductivity (thermometric)	0.006	0.0013	0.004	0.01	0.1	0.1	100	10^3	10^5
Conductive capacity	0.002	0.011	0.04	0.04	0.05	0.3	10	0.01	0.1

*For water and air, these are effective values that allow for degrees of turbulence.

loose, and dry soils are subject to high maximum and low minimum surface temperatures with relatively little penetration. Compacted and wet soils, and clays, are better conductors and exhibit more equable surface temperatures and deeper penetration of daily and seasonal variations. For bare soils, the surface reflection characteristics are also important; darker soils reflect less and respond more to the daily heating. Experiments have shown that a 10° C reduction in surface temperature can be achieved by sprinkling white powder over a dark soil. Specimen reflection coefficients (albedo) for various natural surfaces are given in Table 2.

Table 2: Approximate Values of Albedo
(percentage)

Fresh snow	85
Old snow	50
Salt flats, sand dunes	45
Clean ice	40
Desert	25
Green trees	20
Cultivated soil	
Light	20
Dark	10
Green fields	15
Built-up areas	15
Pine forest	10
Water	
High sun	5
Sun at 5° elevation	25

A surface layer of material of low-conductive capacity tends to insulate the soil below from effects of the daily cycle and from longer spells of abnormal temperature. Thus, freshly fallen snow will protect against a subsequent frost. Snow following frost is only harmful when the frost has already been prolonged, otherwise there will be conduction of heat upward from greater depths. Other soil characteristics that are important to the microclimate include the ability to accept and retain moisture and the tendency to cake and crack. These depend on the composition of the soil itself and its use.

Local variability of soils results in adjacent pockets of quite strongly contrasting microclimates. Large-scale irregularities and the presence of stones and boulders all go to diversify the microhabitats.

Vegetative cover and forest canopies. Vegetation is another insulator, and dense low cover reduces the temperature variability in the soil below. The exposed parts become the radiational interface and the site of the strongest temperature change. Most plants exposed to strong radiation fields require evaporation to prevent damage by overheating. Only typical desert species are adapted both to conserve moisture and to withstand the consequent high temperatures.

Plants and plant communities grow in different configurations and result in an indefinite variety of heat-balance and temperature-distribution situations. A special case is the forest, where canopy and ground are separated effectively by an airspace. Vertical temperature gradients within the airspace are reversed from those in the atmosphere above, with inversions by day and stable, calm conditions providing insulation from the daytime heat. At night the canopy cooling destabilizes the air and causes drafts that bring colder air down in pockets. The airspace is warmer than it would be without a canopy, but the distribution is patchy. Evaporation occurs mainly from the leafy canopy but also from the ground wherever sunlight penetrates. At night, distillation occurs with water evaporated from warmer and condensing on colder spots. These conditions bring about a high humidity during the day and night, a feeling of clamminess in the air, and the accumulation of copious water droplets on the cooling forest litter.

Topography. Air expands and cools when pressures decrease. Elevated places experience low pressures, low temperatures, and high relative humidities. Air ascending a mountain range releases rain and snow; the condensation reduces its rate of cooling, and the warming in its subsequent descent occurs much faster. As a consequence, hot, dry, enervating winds are common in the lee of mountain ranges. These are called fohn winds and also carry local names like the chinook, east of the Rocky Mountains, and the zonda of Argentina.

Topography therefore affects the microclimate by affecting macroclimate, but there are also many direct topographic influences. Figure 2 illustrates the amounts of solar radiation that normally are incident on south-facing slopes of different angles, calculated for the New York area using observed cloud amounts. The great benefit gained by steeper south-facing slopes in winter is evident; even the small summer loss can be turned to advantage by growing crops that do not benefit from excessive heat. The very steep and vertical slopes that are included in the diagram have relevance to architecture.

By courtesy of F.A. Brooks

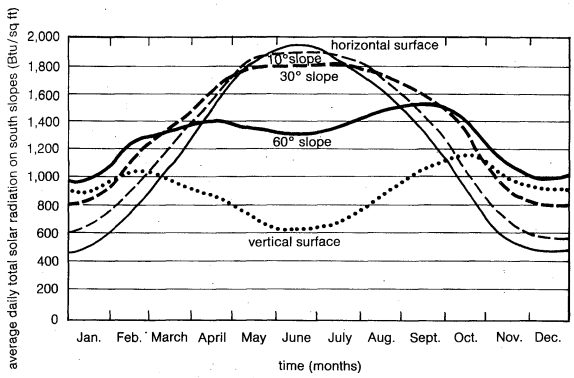


Figure 2: Daily solar radiation on south-facing slopes of varying declivity and the monthly distribution of radiation.

Undulating country also achieves microclimatic variety through air motions generated by differences in density. Nocturnal radiation from a slope creates a lowest layer of cold air that then flows downhill as a wind, which is termed the katabatic wind. The strength and vertical extent of the katabatic wind depend on both the angle and length of the slope. On the vast unbroken slopes of Greenland and Antarctica, it becomes a permanent feature of the winter night, capable of speeds of more than 100 miles an hour and depths of several hundred metres. More commonly, a katabatic is shallow and gentle and drains the cold air, accumulating it in cutoff depressions, or frost pockets, with night temperatures up to 20° below that of the surrounding country. The converse, anabatic wind, flowing uphill by day on slopes that are preferentially heated, lacks the same tenacity because the warm air breaks away from the surface and is more readily mixed. In steep-sided valleys, oriented north–south,

The katabatic wind

rotary motions can result, down the east and up the west side in early morning and oppositely at dusk. Similar circulations due to unequal heating or cooling underlie the familiar phenomena of land and sea breezes.

From R. Geiger, *The Climate Near the Ground* (1965); Harvard University Press

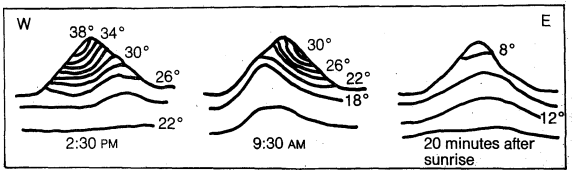


Figure 3: Surface temperatures (in degrees Celsius) on a hill at three times a day.

Mountain and valley winds are a complex of these thermal and gravitational influences on various scales coupled to channelling effects. A steep-sided valley will channel a wind that otherwise would cut across at an angle. On exposed ridges across a prevailing wind, the maximum speed is at the crest or else just below on the windward shoulder, which affords a preferred site for generation of wind power.

Shelter and roughness. Abrupt downslopes and the lee of obstacles gain significant shelter from wind. A good rule of thumb is that a barrier across wind will reduce the speed appreciably for a distance of some 20 times its height. A solid wall or fence provides good shelter for animals but can produce strong lee eddies that buffet crops. More open shelters, like hedges, reduce rather than stop the wind but avoid the damaging lee eddy. Drifting sand or snow accumulates in the sheltered zone behind a windbreak, and this effect commonly is exploited in setting protective fences some distance to the side of exposed sections of traffic routes.

A succession of shelters spaced across country will reduce the overall wind speed near ground level. Any scattered or continuous irregular surface cover constitutes such a succession of shelters. In this way, the low-level wind is reduced by surface roughness. To illustrate magnitudes of the roughness effect, Table 3 presents wind speeds at a height of one metre above some different type of surfaces, expressed as percentages of the speed in the free atmosphere. The importance of these differences is emphasized by the fact that wind force on structures varies as the square of its speed, and the power of windmills as the cube.

Table 3: Relative Wind Speeds at One Metre above Various Surfaces

Smooth mud flats or sea without ripples	52%
Bare soil, very short lawn, cropped pasture, sea in moderate winds	43%
Scrubland, rough cultivated land with row crops	26%
Uniform young orchard in full leaf	19%

Advection. Advection is the conveyance to an identified place of a property different from that currently obtaining there. The strongest advection phenomena occur at sharply contrasting boundaries, such as coastlines. Thus a sea breeze will represent the advection of cooler and moister air to the land. Advection is generally accompanied by a progressive microclimatic adjustment, modifying the properties of the air toward those of its new surroundings, as in the warming of the sea breeze as it moves inland. A small desert lake or oasis will evaporate more rapidly because of the dryness of the wind, drawing on advected energy in excess of that available locally from radiation. Evaporation will be strongest at the upwind edge and progressively decrease downwind as the air grows moister. For this reason, one lake with twice the downwind extent of another will evaporate less than twice as much. This is called the blanketing effect, and it applies equally to heat transfer when cold or warm air invades a surface at different temperature.

While the adjustment process following advection is essentially microclimatic, its effects extend strongly into weather phenomena on larger scales; for example, most rainfall over land occurs in air of maritime origin. Cooling due to advection results in fogs and dews in favourable locations, whereas heating causes convection and cloud and rain showers, and in other ways enhances temperature fluctuations and local differences in moisture. Thermal and moisture gradients of almost any magnitude can derive from advection if the contrasts in the underlying medium are sufficiently sharp.

RELATION OF MICROCLIMATES AND MACROCLIMATES

Macroclimatic influences. In terms of average moisture and temperature, the microclimate of a region is strongly bound to the climate, latitude, elevation, and the season. But the diversification of microclimates, the degree to which conditions vary between adjacent locations or within the daily cycle, is controlled by more special factors. Wet climates have less temperature diversity, both in space and time, than dry ones, partly because of the increase in soil conductivity and partly because the energy that is involved in evaporation and condensation would otherwise appear in the form of temperature changes. Wind and cloudiness are other elements that moderate the microclimates, although each has also one diversifying influence. The greater turbulence of stronger winds increases the air's effective conductivity; this reduces the intensity of temperature changes, but wind diversifies through advection. Cloudiness lowers the net incoming radiation by day and the outgoing radiation by night; however, it diversifies through its patchiness, in the magnitude of temperature change from sunshine to shadow.

The highest temperatures and the strongest daily variations are found not in the tropics but in the high pressure belt of latitudes 20° to 35° where dry, clear, calm conditions prevail. The extreme microclimatic cold is found on high polar plateaus, in places where the cooling of continuous winter night is magnified by clear skies, a poorly conducting undersurface, and locally by the frost-pocket effect (small pockets of cold air in depressions).

Effects of macroclimate on soils

Freezing and thawing in soils are other moderating influences, although less so than evaporation–condensation because of the sevenfold smaller latent heat involved. Because they occur at a single temperature, however, their moderating and retarding effects are sharply concentrated at 0° C. Thus, temperatures close to 0° C occur with relatively high frequency, and microclimatic winter may persist well into climatic summer wherever snow cover remains on sheltered and shaded slopes. An extreme effect known as permafrost (*q.v.*) extends over Antarctica and much of northern America and Eurasia. The ground is permanently frozen, in places for hundreds of metres, and is overlain in less severe parts by an active layer that thaws in summer down to a few metres, permitting growth of mosses and some trees. Deep drainage of the active layer is impossible, so the low-lying parts are waterlogged. The depth of the layer depends on slope, vegetation, and snow, which deepens it through winter insulation. Mosses act in the opposite way by insulating from the summer heat. The prevailing view is that permafrost in many places is a retardation effect from earlier colder periods and that its extent has declined during this century. Large areas with active layers could be more productively developed, although there are problems in construction, maintenance of roads, buildings, underground pipes, etc.

Microclimatic influences on weather and climates. Many influences of microclimate on climate are brought about through variations in surface characteristics. These characteristics determine how the solar radiation is apportioned between the different forms of energy (terrestrial radiation, convection, conduction, and evaporation) into which it is transformed.

Evaporation, for example, accounts for much of the continental–maritime contrast in climate, but it also exerts within the former a sharp cutoff effect on the daily maximum temperature. When water is available, maxima of 32° and 33° C are commonplace but above 34° C almost unknown except through dry-air advection. With dry soils, maxima are widely distributed from 30° to 45° C and even beyond.

The temperature contrasts between air masses are a prime cause of the development of cyclonic depressions. Because an air mass modifies its temperature through contact with the underlying medium, the conductive properties of that medium are vital. Cyclonic development occurs preferentially over certain oceanic regions, coastlines, and over major inland or semi-enclosed bodies of water, but not in seasons when these are frozen. The contrast in effective conductive capacity between water and land or ice testifies to the microclimatic nature of the control.

Effects of abnormal weather

Dramatic effects occur when an unusual weather condition modifies the surface in ways that operate to sustain or amplify the condition. Abnormal runs of weather and certain climatic trends are widely believed to have their roots in such feedback microclimatic processes rather than in external influences. Once more, evaporation enters as a major factor. Wet ground promotes evaporation and increases atmospheric moisture. Conversely, the drying of bare soil creates a crust with an adverse temperature gradient that inhibits upward diffusion of water, and the dry atmosphere persists. Moreover, ensuing dusty conditions create more condensation nuclei, and clouds that then form are colloidally more stable and less likely to give rain. Smoke particles from bush fires or burning of sugarcane, etc., may also reduce rainfall by the overproduction of nuclei.

Another feedback process derives from early winter snowfall in North America and Europe. A storm travelling a southerly track leaves a carpet of snow that increases the albedo; thereafter, less solar energy is retained. This tends to influence ensuing storms also to follow a southerly course. Thus, snow begets snow is one of many weather saws well rooted in scientific fact.

Similar trigger actions may underlie the occurrence of glacial epochs in climate. A sequence of severe winters decreases the fraction of open water in higher latitudes. For this reason, more solar energy is reflected away, less is conducted and stored, and still more ice forms, and an unstable sequence is initiated. This instability may be seen in miniature in small ponds. Once freezing starts, it extends quite rapidly despite the high latent heat, which, as already seen, at any other level than the interface exerts a retarding rather than a destabilizing influence.

THE IMPORTANCE OF MICROCLIMATES

In the hydrologic cycle. Microclimate controls evaporation from surfaces, and it influences precipitation. Both are principal factors in the hydrologic cycle.

Water vapour is transpired into the atmosphere by plants, is evaporated from open water or from bare soil, and is diffused upward from the water table. Over land, provided water is freely available, the total rate is regulated by the local energy supply or net radiation; all surfaces of the same reflectivity, which are uniform for sufficient distance to remove advective effects, will evaporate equally. This provides the concept of potential evaporation, the rate at which the least reflecting moist surface will evaporate. The potential evaporation thus depends solely on weather conditions, among which general temperature level (high temperature, strong evaporation) and sunshine are most important. With heterogeneous surfaces, wind speed also becomes important and evaporation will vary with position. It is strongest over wet areas in drier surroundings (the oasis effect) and where the surface protrudes (the clothesline effect). Where such advective influences prevail, potential evaporation loses significance. Given sufficient contrast, there is no limit to the evaporation rate at the upwind edge of a preferred area. Proceeding downwind, the local rate progressively declines through the blanketing effect of advection until the potential evaporation rate obtains.

Over extensive bodies of water, the local rate of energy supply is no longer a limiting factor because the water itself can store and transport energy in abundance. Limitation resides in the atmosphere's diffusing power, and this depends on turbulence and hence on wind strength near the surface. The evaporation rate E is commonly expressed by the formula

$$E = Cu\,(e_s - e_a)$$

in which u is wind speed, e_a the atmospheric vapour pressure, and e_s the saturation vapour pressure at the surface temperature. C is a quasi-constant whose value is known except for very strong winds, in which accurate experiments have so far proved impossible. Hence evaporation is proportional to the difference between the saturation vapour pressure and the atmospheric vapour pressure. Evaporation and its estimation are important in water-conservation schemes. Losses can exceed 100 inches per year. Research aims at reducing losses by reducing the effective value of e_s in the above formula, for example, by spreading monomolecular films of suitable substances like cetyl alcohol.

Rain and snow

Rain and snow fall from clouds formed by condensation on appropriate nuclei, whose nature and concentration determine the subsequent development of the distribution of drop sizes; this controls stability and, hence, the readiness to precipitate. Nuclei originate as sea salt from evaporated spray, as dust whirled up from exposed soil, or from combustion or other pollution. Thus, surface processes influence the precipitation, though details of cause and effect are often hard to identify.

Once formed, the heaviest rain is little influenced by the ground beneath, but driven rain and snow are subject to local currents that influence the moisture of microhabitats through both precipitation and evaporation. Some minor forms of precipitation are specially sensitive to surface characteristics and can prove decisive in marginal regimes. Dew falls most unevenly, being attracted to tips of leaves and other protuberances that are coldest because of poor conduction. This principle was exploited by ancient Mediterranean civilizations, who erected loose stone cairns to attract and retain copious deposits. Similarly, dew ponds result from attraction by a poor conductor, aided by topography to sustain a katabatic flow of cooled air past or into the sink. Without advection, dew deposit per unit of horizontal area is minute, but it may yet be significant in sustaining some desert species whose tendency to grow in clumps achieves the attraction of dew from surrounding bare areas. In deserts, emergency drinking water is obtainable by digging a hole, filling it with fresh shrubbery, and then covering with clear plastic on to which solar heat will distill soil and plant moisture. Collection of drippings will yield one to three pints of water per day from a hole three feet square.

Another marginal effect occurs when cloud or drifting fog is frequent at the canopy level of vegetation, as in the redwood forests of California. With clouds of small droplets that would otherwise not precipitate, the vegetation can act as a comb against whose teeth the droplets impinge, so collecting into larger drops that drip or trickle to the ground. The ancient Cedars of Lebanon were reputedly so nourished, but once cut down, the precipitation was lost and they could not regenerate.

Rock weathering and soil formation. Fracture of the harder rocks is initiated by rapid heating and cooling and by freezing of water trapped in the porous parts. This is particularly so in clear, calmer macroclimates and at high elevations where microclimatic extremes are greatest. Microclimate accounts only for this initial fragmentation, and the final weathering (*q.v.*) of hard rock into the clay and mineral-particle constituents of soils (*q.v.*) is largely a chemical process. Even here, however, warmth and moisture control the pace and degree; soils in cold, clear climates tend to remain thin and rocky. In stormier climates, and everywhere with softer rocks, the prolonged mechanical action of wind and rain is another soil-forming process. There are indirect influences of microclimate also through plant growth, which accelerates weathering by ionic exchanges (addition or subtraction of charged particles) through the roots, increasing the proneness to subsequent chemical decomposition. The ionic exchange is strongest for lichens and other more primitive plant forms, and so the nutrition of the plant is an inseparable part of weathering.

In agriculture. Agricultural practices are aimed at achieving the best possible arrangement of chemical, biochemical, and physical factors in the micro-environment

of plants and animals. Man strives to manipulate the temperature, moisture, and wind exposure to suit each crop according to its stage of maturity, combatting the constraints imposed by macroclimate, soil, and topography. He is specially concerned to avoid microclimatic disasters such as frost or gale damage, scorching, wilting, and soil erosion.

Proper study of siting and exposure and contingent decisions of cropping practice come first and may outweigh in importance the subsequent tillage of the land. Sun-facing slopes accommodate the heat-loving species, whereas poleward-facing slopes harbour the hardier plants. Exposed sites must be avoided or shelter provided for fragile crops, and attention must be given to soil and air drainage according to the projected crop. Hedges across slopes dam up cold air, producing frost-prone strips wherein all sensitive buds or young fruit may be killed up to hedgetop level; gaps for air drainage are effective but will increase the severity of frost pockets further downhill.

The choice of crop follows study of siting. Any but the most adverse microclimates can grow some crop, so the agriculturalist's task is to optimize; in continuous exercise of judgment and experience, understanding and application of microclimatology become virtually instinctive.

General principles of temperature and moisture control through cultivation have already been explained, but generalizations about optimum conditions and practices must be treated with care because of variations between species. The germination of seeds in cold ground in spring depends on the conduction of sufficient heat to the seed. Shallow-sown seeds will prefer a dry, loose soil, deeper ones a firmer and more conducting soil. Later, the supply of water and nutrients in the root zone becomes significant. As soon as the seedling emerges, surface moisture is required to protect the tender tips from extreme heat and cold. Thereafter, its total growth will depend on a complicated factor of accumulated temperature or of accumulated temperature above some threshold (5° C) below which growth is negligibly slow. Moisture then acts mainly as a passive part of the transpiration-growth process and as a protection against high temperatures and wilting. For many crops, success depends not on total growth but on some special property of flower, fruit, or leaf that responds to a somewhat different combination of heat units, moisture, and timing.

Seed germination

Agricultural practice must vary accordingly. Aside from controlling the physical properties of the soil by tilling and mulching, there is scope for changing its geometry by row cultivation, hilling, etc., with considerable impact on low-growing and root crops. Two examples described by Rudolf Geiger, the German microclimatologist, are illustrated in Figure 4. Young trees in a nursery threatened

From R. Geiger, *The Climate Near the Ground* (1965); Harvard University Press

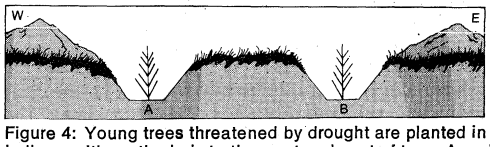


Figure 4: Young trees threatened by drought are planted in hollows with earth clods to the west and east of trees A and B as shown. Trees in hollow B thrived relative to those in A because they were shaded from the morning sun and had a few more hours of evening dew and humidity.

with drought were planted in hollows, with displaced clods left around the hollows. Plants B, with clods to the eastward, were shaded in the morning and had a few more hours of night dew and humidity. These thrived relative to plants A, which dried out early without any compensating benefit from the afternoon shade. The lower part of the Figure shows the temperature distribution at three different times of day, in unplanted drills running north–south at a height of about 15 centimetres and 30 centimetres apart.

Scorching and wilting are best avoided by watering, but frost prevention may take many forms. Smudge fires in orchards heat the air directly and achieve additional benefits by elevating the effective level of the outgoing radiation. Water sprays producing a glazed frost around buds and fruit protect them from falling below 0° C. Soil water, weeding, rolling, and frost fans achieve protection by increasing the effective conductivity of soil or air. Combinations of these methods can achieve benefits of more than 5° C in radiation frosts on calm, clear nights but are ineffective during prolonged freezes when day temperatures remain below 0° C. But even here some benefit is possible in marginal cases in which the surface is manipulated to retain more sensible heat, as by adding a covering of sand to the cranberry marshes in Wisconsin.

Soil erosion is the most permanent of all microclimatic disasters. Gully and other forms of water erosion are avoided or reduced by contour ploughing and by vegetation or artificial holding devices. Dust-bowl conditions leading to wind erosion are more insidious because of their progressive spreading if unchecked by means of water, shelter, and holding vegetation.

In air pollution. The dispersal of local pollutants is governed by turbulence and its variation with height. This may be observed from characteristic patterns of behaviour (coning, looping, fanning, lofting, fumigation) of smoke plumes from chimneys. The variation in turbulence through the diurnal cycle, and its dependence on weather conditions, thereby exerts a dominant influence on the time and space variations of air pollution in urban environments, thus creating urban climates (*q.v.*).

The interaction between microclimate and topography is another vital consideration, as evidenced by the katabatic accumulation of smog in low areas. Steep-sided valleys are specially prone to organized fumigations in which a concentrated stack effluent, accumulating clear of the ground by night, is brought to the surface by thermal circulation set up after sunrise. This may occur far downstream, as in fumigations in the U.S. from the smelter in Trail, British Columbia. Where chimneys are badly sited, their emission may have to be governed according to micrometeorological criteria.

Smog effects

In optical and radio transmission. Differences in optical refractive index accompany differences in temperature and density of the atmosphere. Shimmering on hot days and the lack of clear outline of distant objects results from disorganized refraction variations due to turbulence. Systematic layering of density causes more organized effects known as mirages. If the line of sight is near horizontal, images are displaced from their true position. The displacement is downward by day in the inferior mirage, when lighter air lies below heavier air and the rays are convex downward. The image of the sky, looking like water in the surface, has proved the despair of desert travellers. The superior mirage is convex upward and results from temperature inversions. These are most prevalent at sea with warm air overrunning colder water. Under special conditions, the ray may be trapped in a relatively shallow layer. Where this occurs, or when the curvature of the ray is close enough to that of the earth, very distant objects are seen well beyond the horizon and may appear deceptively close. Lateral mirages can appear over hilly country, where the density distribution conforms to the topography. Significant corrections for all systematic refractive effects are required in surveying, visual navigation, and astronomy.

These phenomena are paralleled in radio transmission, with the difference that the refractive index depends more on humidity than temperature. Because humidity decreases with altitude, near-horizontal rays bend back toward the earth (the superior mirage); the radio line of sight commonly extends well past the true horizon. Thus, anomalous propagation, in which waves travel large horizontal distances with little loss of energy, occurs more frequently in radio than in the optical wavelengths. If temperature decreases with height along with humidity, the resulting turbulence will tend to break up the regular stratification. The most extreme condition occurs, therefore, when dry air, as of a fohn wind, flows over cool water. Radar reflections can be obtained from humidity discontinuities in perfectly clear air, and other more complicated propagation effects are observed under special conditions.

Sound waves are similarly influenced by the density distribution in the air. Distant sounds are heard much more clearly at night, or with warm air over cold sea, when the bending is downward and the muffling effects of turbulence are absent.

SPECIAL AND ARTIFICIAL MICROCLIMATES

Artificial microclimates for crops. Whereas man in agriculture manipulates an essentially natural microclimate, his cropping and storage practices extend to the creation of artificial and radically new environments. Historically, the art was first practiced in the construction of walled gardens that provided shelter and directed the receipt and disposal of solar radiation.

Greenhouses

Glass has made possible the fullest concentration and exploitation of nature's heat sources. Short-wave radiation penetrates glass to warm soil and plants, and the convection plums, which in the open would rise high, can be trapped to heat a much smaller depth of air. Long-wave radiation from the glass makes a further contribution to the warming. Combinations of sheltering and the greenhouse effect are exploited in less ambitious structures (cloches, cucumber frames, plastic covers, etc.) for protecting sensitive crops or growth of valuable products at unseasonable times or in marginal climates. If sunshine is available, almost any degree of temperature increase is possible with devices no more complex than concave mirrors or blackened receptors; and these are in use in appropriate climates for cooking, water heating, and other purposes. With modern aids of air conditioning and artificial sunlight, supergreenhouses known as phytotrons are used to study plant growth and responses over the complete range of temperature, humidity, airspeed, and effective day length, though fully realistic modelling of the gradients of the variables that surround and affect a plant in nature have yet to be achieved.

Storage of crops presents different microclimatic problems. In many countries, winter supplies of root vegetables and fruit are stored outdoors in clamps, insulated by straw and earth to keep the temperature at about 3° C and ventilated to avoid excessive heat from fermentation. The great depths and pressures involved in silo storage of grain magnify this danger; moisture will migrate as in soils from the warm core to the colder parts and there cause molding. Suitable cooling by aeration will check this migration and, by keeping temperatures low enough, will inhibit breeding insects (which must otherwise be controlled by chemical methods).

The habitat of man. Although man feels comfortable at temperatures between 18° and 32° C, he has extended his habitat into climates ranging from −75° to +60° C. Clothing is a prime insulator against cold and excessive strength of sun. The microclimate of enclosed dwelling spaces (cryptoclimate) is today fully controllable by air conditioning, but there is much diversification where control is incomplete. An extensive bibliography of conditions in houses, cellars, tents, and other places has been compiled. Heat inside automobiles may be extreme because of exposure and poor insulation: temperatures 20° C higher than outside have been measured in a black sedan under certain conditions; excesses are halved by light colouring and further reduced by motion and ventilation.

The external climate and microclimate of cities differs from their natural surroundings because of different surface material, the changed effective roughness or shelter complex, and artificial alteration to heat and water budgets. Dust, smoke, and fumes affect the radiation balance, and the heat supplied by artificial fuels in winter may greatly exceed the receipt from the sun. The degree of modification depends so much on these factors and on topography that generalizations are difficult. An overall warming of 1° C is common, much more on cold, clear nights and in low-lying parts. Artificial convection in winter may be sufficient to increase rainfall over parts or all of the city.

Microclimate of caves. Caves and cave systems (*q.v.*) provide nature's cryptoclimates. Cold air seeps preferentially into the entrance, the average temperature of which is below that outside. Elsewhere, conditions depend strongly on the configuration. In static (single entrance) caves, downslopes admit only cold air, and the coldest seeps downward and stagnates. Rising caves are warmer, and the temperature increases inward and upward. Bats, insects, and other animals, therefore, are more numerous. Temperatures may be almost constant. Plant life occurs only near the entrance, where there is light for photosynthesis. Transit or dynamic (multiple entrance) caves are ventilated by air currents that arise because of temperature differences at the entrances. The current velocities can reach several metres per second at constrictions. The direction of circulation may vary with season or even diurnally if the entrances are differently exposed. Artificial caves, such as mines or railway tunnels, are normally ventilated or air-conditioned to avoid these adverse effects.

The marine and littoral microclimate. In oceans and seas (*q.v.*), microclimate is almost uniquely determined by macroclimate; and, with macroclimate more equable, microclimate has far less scope for diversity. Surface temperatures range from 30° C to freezing and vary only slowly with time, except at convergences or on the borders of strong meandering currents. Daily variation at the surface is about ¼° C. The bottom temperature of all deep oceans is generally between 0° and 2° C. The main vertical contrast is concentrated in the thermocline, extending through a few hundred metres from close to the surface. The increase in pressure with depth strongly influences life forms, but the most important physical variable controlling the habitat is light intensity. One percent of incident energy would penetrate to depths of several hundred metres in pure seawater, but more representative magnitudes for real oceans are 100 metres for the clearest waters, ten metres in coastal waters, and one metre in very turbid water near river mouths. One hundred metres is about the limit for plant growth.

Contrastingly, the intertidal zone of the shoreline provides unique and strongly time-dependent microclimates. Shallow rock pools in strong sun may rise to 40° C before the tide returns, and bare-rock surfaces become even hotter. Rock fauna, seaweeds, and animals that dwell among them must tolerate extremes of temperature, regular inundations of saltwater, and occasional inflows of freshwater from storms. The duration of exposure to microclimatic contrasts decreases from the high-water to the low-water mark. At neap tides, some shore is left uncovered for days on end. The strong microclimatic contrast results in zonations of plants and animals throughout the intertidal zone. Under stones and small rocks there exists another unique microhabitat; it is uniformly moist, saline, and without plants. Mangrove swamps, whose interlacing trunks and roots accumulate a permanently shaded mud layer, provide yet another unique microclimate and animal population.

HABITATS AND RESPONSES OF ORGANISMS

Heat budgets and animal modification. Just as the heat budget controls the microclimates at and near the ground–air interface, so the heat budget of the individual plant or animal governs its own proprioclimate and its own climatic responses. Energy is absorbed as radiation from the external field or is created by metabolism within. This must be balanced by long-wave radiation, convection, conduction of the ground, or evaporation (plants use only about 1 percent of their available energy in photosynthesis). Plants, and animals that sweat or pant, use principally evaporation; if this is inadequate, the surface temperature will rise until other processes can take over. Generally, only specially adapted desert plants can conserve moisture and withstand the consequent high temperatures without collapse. Animals are more versatile; the relative efficiencies of the various heat-disposal mechanisms will depend on size, shape, and exposure and on external factors such as wind speed.

Limitations on animal size

Animal size is related to the various heat quantities by the following rough rules. Metabolic heat production varies as the volume (for a given degree of activity); absorbed and emitted radiation varies as the projected area; and

free and forced convection varies less rapidly than the area because of the blanketing effect described earlier. That most members of the Negroid races of man have long slender limbs as compared with those of most of the short, squat Eskimo reflects their relative efficiencies in heat disposal and heat retention (large and small surface areas), respectively. Warm-blooded animals in cold climates require insulation, as do the larger animals exposed to strong sun unless they have evaporative cooling. Wool insulates the sheep; the temperature at the ends of its fleece may rise to 90° C before long-wave radiation can dispose of the heat intercepted: much smaller animals require no insulation from sun because smaller areas yield a higher convection-to-interception ratio and heat balance is reached at a lower surface temperature. For insects in flight, metabolic heat represents the major load, and this is shed by forced convection.

Such self-regulating systems imply adaptability to microclimatic changes. Organisms with little adaptability must achieve their narrower tolerance range; plants accomplish this by selective survival, and animals, generally, by locomotion. The habitat may even accompany the animal. In heat, snails and winkles retire to the recesses of their shells to avoid desiccation. Birds' nests, ant heaps, and animal burrows are made for comfort and protection of young and old and show attention to microclimatic positioning and complexity of structure that compare with man's own artifices. Notable are mound-building birds such as the Australian mallee fowl. The male's beak is temperature sensitive, and he daily tends the mound, adjusting soil insulation to the conditions of the day and season and to the heat of fermentation of its vegetable content.

Life forms in special environments. The broad divisions of macroclimate and environment generally determine the first-order distinction between larger species. Smaller species depend on microclimate, but because its ranges overlap the macroclimatic zones, smaller species have often the wider geographic distribution.

The great variety of microclimates engenders enormous diversification of species, and in practice only two criteria appear ultimately limiting to live growth: temperatures too extreme to sustain any liquid in vital tissues and, for plants, insufficient light. Animals are not restricted by the latter, for food may be brought to them by water, wind, or gravitational settling.

Unique micro-environments produce unique populations: earthworms and mud crabs, caterpillars at selected locations in plants, echinoderms on the sea floor, prawns and squids with luminescent organs in the depths. From Antarctica to tropical rain forests, the interface is populated by organisms of varyingly adapted species. Beach worms and clams, in densities of thousands per cubic foot, characterize the intertidal environment and its fecundity. Cave species show genetic adaptation to darkness and to the absence of rhythm and change in their subterranean environment. The microclimatic correlation is generally understood only qualitatively because the fine instrumentation for quantitative studies is of relatively recent development.

Variability effects on populations. Macroclimate controls the migrations of birds and fish, and its vicissitudes influence their abundance. In the same way, microclimate reacts on the daily behaviour of all life forms and their variations in space and time. Again, the marine environment presents an inherently more simple picture than does the land. Light confines anchored plant growth to shallow waters. Animal plankton rise at dusk and descend in the morning, seeking constant light intensity. Fish follow their food supply. The thermocline inhibits these daily migrations, but where major vertical circulations exist, the population density is profoundly influenced. Vast aggregations of life are found, for example, off Antarctica and in other nutrient-rich upwelling waters.

On land, the abundance of life is strongly correlated with temperature and moisture. The limits of the area that any particular species inhabits is a boundary zone in which the species is found to exist in pockets. Moisture especially exerts a buffering influence in regions otherwise subject to strong temperature gradients and rapid changes. Given adequate nourishment, enormous concentrations of smaller insects and microbial forms inhabit soil layers that remain moist but not waterlogged. The same applies to the permanently moist patches under stones and rocks in the intertidal zone, sometimes crowded with life to the point of suffocation.

Massing of populations in air is, by its nature, more spasmodic, and the causes are several. With birds, as with larger land animals, grouping is socially motivated, and only the composite movements of the flock are climatically regulated. Smaller flying insects mass from instincts that are microclimatically oriented, be it to seek shelter from wind buffetings or a favourable temperature in and around canopies in the morning. Massing of locusts in upcurrents appears a largely inert response, although it is motivated toward finding a good wind for transport.

Daily behaviour patterns of all animals relate to cycles of temperature and light, but in an almost infinitely varied manner. Speed of movement of smaller cold-blooded species is directly responsive to temperature. Snails, caterpillars, and crickets migrate up and down plants to suit their heat and moisture balance according to time of day. Some animals and plants vary their orientation to maintain maximum interception of radiation (see also PERIODICITY, BIOLOGICAL).

MEASUREMENT OF MICROCLIMATES

Where quantitative understanding of the microclimate is sought, the most important single measurement frequently is net radiation, which governs the heat balance. It can be measured directly by a thermopile with sets of junctions in contact with two opposed-plane blackened surfaces. These are so adjusted that the sensitivity to solar and terrestrial wavelengths is the same. Surface temperatures are also best measured by a suitable radiometer. Both instruments can be miniaturized to about one centimetre for study of spatial variations on this scale.

Soil temperatures can be recorded by any standard method or read from thermometers with long stems. Soil moisture *in situ* is determined indirectly from moisture tension or by electrical or thermal methods that determine moisture-dependent properties such as capacitance or conductivity. Temperature and humidity determinations of the air and within plant canopies frequently call for precision, and the strong gradients involve small sensing elements and sensitivity to rapid fluctuations. Probes are constructed as single thermojunctions, fine resistance wires, etc.; and when wrapped with a fine wick, they yield wet-bulb temperature and humidity. Small instruments avoid the need for special housing or aspiration. They can be used to measure temperature or humidity of individual eddies and, linked with a correspondingly fine measurement of air motion, can yield the rate of heat convection or evaporation by the eddy flux, an absolute method. The airflow instrument is most appropriately a hot wire anemometer or very light vane or propeller. Heat flux into the ground may be determined from temperature difference across a thin plate, whose conductivity is close to soil conductivity so that the heat flow is undistorted.

For measuring average conditions over periods of several minutes, and how these vary with height, etc., masts that hold arrays of more robust instruments are used. For mean wind speed, sensitive, rotating cup anemometers are commonly employed.

For evaporation, direct weighing of soil blocks is an alternative to the eddy flux as an absolute method, provided no disparity with surrounding terrain and surface cover is introduced. Various evaporation pans and devices similar to the Piche evaporimeter are also widely used. By interfering with the natural surface, however, they introduce advection errors, and their values must be corrected or restricted to comparative uses. This also applies to most dew gauges, in which a collector attracts the dew and the amount of deposit is determined. The use of a net radiometer and application of energy-balance principles will help to detect and avoid this sort of error.

BIBLIOGRAPHY. R. GEIGER, *Das Klima der bodennahen Luftschicht*, 4th ed. (1961; Eng. trans., *The Climate Near the*

Ground, 1965), the standard and most comprehensive work, including an extensive bibliography; O.G. SUTTON, *Micrometeorology* (1953), a scientific textbook dealing with the physical processes; C.H.B. PRIESTLEY, *Turbulent Transfer in the Lower Atmosphere* (1959), a monograph for professionals concerned with the turbulent processes; F.A. BROOKS, *Introduction to Physical Microclimatology* (1959), a treatise for professional men seeking a working knowlege; R.O. SLATYER and I.C. MCILROY, *Practical Microclimatology* (1961), a manual for practical investigators. Additional references and information on this subject may be found in C.E.P. BROOKS and G.J. EVANS, "Annotated Bibliography on the Climate of Enclosed Spaces (Cryptoclimates)," *Met. Abstr. Biblphy.*, 7:211–264 (Feb. 1956); V.I. VITKEVICH, *Agricultural Meteorology* (1963; orig. pub. in Russian, 1960); J.Y. WANG, *Agricultural Meteorology* (1963); H.U. SVERDRUP, M.W. JOHNSON, and R.H. FLEMING, *The Oceans* (1942); M.N. HILL (ed.), *The Sea*, 3 vol. (1962–63); and H.E. LANDSBERG, *Physical Climatology*, 2nd ed. (1958).

(C.H.B.P.)

Micronesian Cultures

The name Micronesia is derived from two Greek words meaning small islands. It is generally considered to include the Caroline Islands from the Palau Islands in the west to Kusaie in the east, the Mariana Islands to the north of the western Carolines, the Marshall Islands to the east of the Carolines, the Gilbert Islands to the south of the Marshalls, and the isolated island of Nauru west of the Gilberts and south of the eastern Carolines. The islands of Micronesia are scattered over an area of the tropical western Pacific, mostly north of the Equator, for an east–west distance of about 3,000 miles (5,000 kilometres) and a north–south distance ranging up to more than 1,000 miles, but the total land area of all the islands combined is only about 1,100 square miles (2,800 square kilometres), of which 207 square miles are found in the island of Guam and about 130 square miles each in the two next largest islands, Babelthuap in the Palaus (western Carolines) and Ponape (eastern Carolines). Many of the permanently inhabited atolls contain less than one square mile of land. The inhabitants of the Ellice Islands to the south of the Gilberts and the isolated atolls of Nukuoro and Kapingamarangi west of Nauru and south of the eastern Carolines speak Polynesian languages and by culture and physical type are properly classed as Polynesians rather than Micronesians (see map in OCEANIC PEOPLES AND CULTURES).

Nearly all the islands of Micronesia can be conveniently classified as high or low islands according to their physical relief. The high islands are typically volcanic; they support fairly large populations, and their soil sustains a greater variety of wild and cultivated plants than does that of the low islands. The high islands have traditionally served as ports of call for the inhabitants of nearby low islands.

THE TRADITIONAL CULTURAL SETTING

Seven major high-island cultures can be distinguished in Micronesia: those of the Chamorros, inhabiting the Marianas, principally the four southern islands; the Palauans; the Yapese; the Trukese, inhabiting about 12 high islands of varying size in the large Truk Lagoon; the Ponapeans; the Kusaieans; and the inhabitants of the isolated island of Nauru, which is geologically a raised atoll (without exposed volcanic rock).

Diversity of the Micronesian peoples

The inhabitants of most of the low islands in Micronesia are culturally distinct from the high islanders, though in contact with them. In the east are found two culturally distinctive groups of atolls, the Marshalls and the Gilberts, ranging from the northwest to the southeast over about 1,400 miles. The culture of Ocean Island, or Banaba, a raised atoll, is quite similar to that of the Gilberts and will be treated herein as a variant of Gilbertese culture. Three atolls within sailing distance of Ponape—Mokil, Pingelap, and Ngatik—show closer cultural relationships to the people of Ponape than to any other large population but are clearly distinct from them. The Hall Islands, atolls to the north of Truk, and the Mortlock (Nomoi) Islands, atolls to the south, are culturally closest to Truk. The remaining low islands to the west of Truk also show linguistic and cultural relationships to Truk, with the differences becoming increasingly marked the farther west one goes. The low islands between Namonuito and Yap were once part of a ceremonial exchange system called by anthropologists the Yap Empire. Linguistically and culturally, however, these low-island people were closer to the Trukese than to the Yapese.

The cultural diversity of Micronesia is indicated by the number of mutually unintelligible languages in the area. The following all have mutually unintelligible languages: Nauru, the Gilberts, the Marshalls, Kusaie, Pingelap, Mokil, Ponape, Ngatik, the Mortlocks, Truk, the Puluwat area, the Woleai area, Yap, the Palaus, and the Marianas (Chamorro language). The languages of the islands from the Marshalls through the Woleai area in the above list appear to be somewhat more closely related to each other than to the remaining languages and have been referred to as Nuclear Micronesian. The languages of Yap and the Gilberts may also be more closely related to this group than to any other Pacific languages. The languages of the Palaus and the Marianas are relatively distinct from each other and from other Micronesian languages, although they are clearly Austronesian (formerly called Malayo-Polynesian) in their general affiliation. The Chamorro language, spoken in the Marianas, has undergone much influence from Spanish and probably also from the Philippine Tagalog language after more than four centuries of Western contact.

Traditional Micronesian cultures were characterized by a belief in the stability of society and culture, perhaps even more so than most other nonliterate cultures. Man suffered occasional natural disasters, such as typhoons or droughts, but his goal after encountering one of these was to reconstitute the previous state of affairs. Wars occurred in most areas from time to time but often resulted in few casualties and no changes in land ownership.

Living in small communities on small territories, Micronesians learned to adjust to their neighbours and remain on good terms with most of them most of the time and to develop techniques of reconciliation when fights did break out. In general, Micronesians valued good manners. Even today, Micronesian adults and children are likely to impress the outsider as unusually quiet, polite people.

Local loyalties

Micronesians traditionally depended on the cultivation of plant crops and on fishing in shallow reef waters. Since both arable land and reef were in short supply for the relatively dense population, the Micronesian had a strong practical basis for his attachment to his locality and lands. Land rights were usually held through some sort of lineage or extended family group, often backed up by traditions of ancestral origins on the land.

The strong local loyalties of the Micronesians may also be partly explained by the difficulty of travelling to any place very far from home, especially for the high islanders who lacked ocean-going canoes. Of the high-island peoples, only the Yapese practiced very much navigation in the open seas at the time of their discovery by Europeans, sailing to Palau and to some of the atolls in the central Carolines. The remaining high-island peoples mostly sailed around their home islands without losing sight of land, although they were visited from time to time by low islanders. The low islanders had reason to visit the high islands, with their more fertile soil and greater elevation; they could obtain food products not found on their home islands or seek refuge after a typhoon or drought. Low islanders also had reason to visit other low islands in search of mates outside the limits of the incest taboo and for help after typhoons. Some of the low islanders, especially in the storm-swept central Carolines area between Truk and Yap and in the Marshalls, were skilled navigators of oceangoing canoes.

The populations of the Micronesian islands appear to have been in good balance with the natural resources at the time of their discovery by the Europeans. Since the climate is equable and varies little during the year, moderate amounts of labour were sufficient for comfortable survival, and much leisure was available for activities such as dancing, feasting, romantic affairs, and visiting

friends and relatives. Some kinds of subsistence activities, especially fishing, were regarded as something like play. The period of youth was often prolonged, since the older people could afford to indulge their children. This positive attitude toward the enjoyment of leisure was especially characteristic of the high islanders, with their more fertile soil and more secure life.

MICRONESIAN SOCIETY

Political units

Throughout most of Micronesia the maximum independent autonomous political unit was the high island or atoll, often subdivided into more than one unit. Satawan atoll in the Mortlocks had four separate communities aboriginally, each with its own chief; sometimes they fought each other. The Palaus had two confederations of villages or districts aboriginally, each independent of the other, and the villages themselves had considerable autonomy. Ponape had five petty states at the time it was discovered, although there were traditions of a unified rule for the whole island at an earlier period. Truk was extremely fragmented politically, with several independent communities on each of the six larger high islands. The Marshalls and the Gilberts had larger political units, integrating groups of separate atolls under a high chief, although the fullest development of these conquest states was achieved after the introduction of firearms by Europeans.

The low islands between Truk and Yap have been described as belonging to the Yap Empire. The so-called empire consisted mainly of a chain of trading and ceremonial relationships with one of the petty states of Yap. The Yapese exerted no military force against the low islanders but did claim the ability to punish them by magically sending typhoons, disease, and famine if they should fail to fulfill their obligations. According to widely accepted myths, the low islanders originated from the Yapese, although they are culturally and linguistically closer to the Trukese.

In most of the area from the Palaus in the west to the Marshalls in the east, the community was considered to be owned in some sense by a clan, the head of which was also the chief of the community. Other clans had land rights by their relation to former chiefs or to the men of the ranking clan: their women had married men of the chiefly clan, or their men had rendered the chiefly clan service in the past. The chiefly clans in some cases claimed their position by virtue of ancient military conquest and in others by virtue of being the first human beings to occupy the land. The aboriginal pattern of political and community organization in the Marianas has been obscured by the early Spanish conquest, which exterminated most of the population and concentrated the remainder under close Spanish military and religious control. Chieftainship in the Gilberts was variable and in some cases patrilineal.

Family and kinship patterns. The majority of Micronesians before European contact lived in some form of extended-family group. In most areas the organization of this group probably had considerable flexibility, some newlywed couples living with the husband's family, others with the wife's relatives. Major determinants in the choice of residence were the relative availability of agricultural land and the need for additional labourers on one or the other side of the family. In much of the area the wife's family was considered the ideal place of residence; but frequently exceptions were allowed in practice, and children often had rights to use land on their father's mother's side. In Yap, on the other hand, patrilocal residence and patrilineal inheritance of land were considered ideal.

As in other parts of Oceania, the adoption of children of relatives was frequent. This was, in part, a way of caring for illegitimate and orphaned children; in part, a way of relieving young adults of the chores of child care and providing old people with young children to do minor work for them; in part, a way of insuring more equitable distribution of land rights; and, in part, a way of providing heirs for specialized knowledge when the proper natural heir was lacking or unsatisfactory.

Where matrilineages existed in Micronesia, they were traditionally exogamous—*i.e.*, members did not marry within the same lineage. While matrilineage membership was considered basically unalterable in some communities, there was probably some flexibility everywhere. If a lineage grew too large, it tended to split into two parts, one of which would adopt a new name and no longer observe exogamy with the other. If outsiders came into a community, they would often be taken into an existing lineage as honorary or fictive members and eventually might merge with the patron group and observe rules of exogamy with them.

Marriage

Marriage in Micronesia varied in formality. In the west, in Palau and Yap, marriages were marked by formal payments from the groom's family to the bride's. In the area from the central Carolinian atolls to the Marshalls, marital relationships were usually rather loose and informal for most people, although people of high rank may have had public ceremonies with some exchange of wealth. In this area considerable premarital and extramarital sex was traditionally expected; marriage for ordinary people consisted simply of openly living together and being spoken of by the community as spouses. Apparently, there was more formality to marriage and more control of premarital sex in the Gilberts.

Polygyny was generally permitted to some extent in Micronesian societies, although it was not very common. It was more likely to be practiced by high-ranking men, and in eastern Micronesia it may have been restricted to chiefs.

Birth order is widely important in Micronesian societies. The seniority of the eldest child is recognized in representing the family or lineage in public and in inheriting political office. The eldest child generally directs the use of lineage or family lands. Younger siblings generally exhibit formal respect to older siblings.

Brother–sister avoidance relationships are well developed in parts of Micronesia, perhaps most strongly in the central Carolines from Truk through the atolls to the west. In this area women were traditionally expected to avoid speaking to their brothers, to crouch in their presence, etc. In Ponape, farther to the east, a similar relationship existed between a man and his next younger sister but did not extend with the same force to other sisters.

Social stratification. A certain amount of hereditary social stratification was found universally in Micronesia, but its degree varied considerably from some of the smaller Carolinian atolls, which had nominal hereditary chiefs with little special power or wealth, to the high island of Yap, which had several ranked endogamous castes. Other cultures that showed relatively marked social stratification were Palau, Ponape, Kusaie, the Marshalls, and the Gilberts. It seems likely that the Marianas may have also had distinct social classes before the Spanish conquest. In all the above areas there appear to have been some chiefs who were supported principally by tribute from their subjects; who were the object of considerable deference; who could punish offenses (especially against themselves and their own relatives) by fines, destruction of property, or death; and whose principal wives were chosen mostly from other high-ranking families. In these areas the high chiefs also had subordinate chiefs and officials.

Egalitarian tendencies

At the same time, there were limitations on the development of the hierarchical principle in most of these societies. The notion of crime to be punished by the state was weakly developed. Palau and Ponape had an abundance of nominal political titles, the principal significance of which might be called recreational and dramatic. In both these societies there was considerable opportunity for individuals to rise above their initial hereditary status through personal achievement, thus introducing much flexibility into the system of social stratification. The principle of birth order worked at cross purposes with the principle of hereditary rank of clans and lineages; a child of low birth order in a high-ranking lineage would not necessarily receive much advantage from his relationship to the chief.

On the high islands of Truk and on many of the low islands, especially those in the central Carolines and the smaller communities in the Marshalls and Gilberts, social relationships were quite egalitarian. The local chief was expected to work for his living like anyone else. He received occasional ceremonial deference but traditionally had little power to order people to do anything or punish them for offenses. His principal function was to preside at meetings and to try to define and state the community consensus as it developed. He had a wide choice as to whom he might marry.

Socialization and education. Micronesians are indulgent with infants. Because of the extended family, someone is usually available to amuse the baby and take care of its needs. Babies are generally exempt from physical punishment. In the tropical climate, toilet training is mainly a matter of teaching the baby to go outside the house for elimination when it is able to walk.

Children were inducted into adult life gradually through observation and participation; there was little in the way of formal schooling or initiation ceremonies in most of the Micronesian societies. Several forms of bodily ornamentation by mutilation were practiced, mostly performed around puberty or in early adulthood. The most widespread of these was tattooing, practiced by both sexes. This and other forms of bodily mutilation were generally done on individual initiative, to demonstrate bravery and increase a person's attractiveness to the opposite sex.

Training in cultural specialties, such as medicine, magic, esoteric mythology, canoe and house building, and navigation, was often delayed until young adulthood or middle age. It was usually given by an older relative, although sometimes outsiders could receive it by payment of food and goods. The most formal training in esoteric knowledge appears to have been given by the specialists known as *itang* (in Truk). Each *itang* had himself been trained under an older expert adept in traditional history, oratory, war strategy and tactics, and magic. At the end of his training he was formally awarded the title or degree of *itang* and thenceforth could serve as an orator, ambassador, counsellor, and executive officer for a chief. An *itang* sometimes trained a group of several young men at a time, receiving prescribed payments from them.

Settlement patterns. Micronesians preferred to live in dispersed, extended-family homesteads, in contrast to some Melanesians and Polynesians. In atolls the inhabitants generally preferred the lagoon side of the larger islands for ease in launching canoes and for protection from typhoons. On the high islands, people again wanted access to the lagoon, although easily defensible sites were sometimes also preferred, such as the top of a steep, cleared slope.

The "Venice" of Ponape

One ancient "city" in Micronesia is the archaeological site of Nan Matol or Metalanim on Ponape. This is a sort of tiny Venice, containing about 100 artificial islands, separated by shallow canals, covering one-third of a square mile in all, including water. The islands were used for residence and for ritual, and the total population may have been several hundred to 1,000. There is a similar site at Lelu (Lele) in Kusaie.

The typical Micronesian community had one or more meetinghouses. These served both as social gathering places and as places to plan community affairs. The number and elaborateness of the meetinghouses is greatest in Yap and Palau. Palauan meetinghouses were built by the men's clubs. In Palau, Yap, and the western atolls, meetinghouses were used mostly by men, while farther east, women and children also entered them freely much of the time.

Houses in most areas were built on slightly raised platforms made of coral rock and gravel on the low islands and volcanic rock and dirt on the high islands. They generally had thatched roofs with low eaves and poor ventilation. The smoke from a small hearth may have been a way of controlling mosquitoes, although plaited mosquito sleeping bags were also used at times. This form of construction may have also provided good protection against heavy rainstorms. Some of the houses in the Marianas appear to have been constructed on stone pillars. The *latte* stones of this area—double rows of stone pillars with capstones—are now thought to have been piles for raised houses, perhaps for chiefs and wealthy men, since the *latte* sites are relatively few for the reported population. Other houses may have been built on wooden piles that have since disappeared.

Canoe houses were another important form of building throughout Micronesia. Those big enough to store the larger canoes were on the scale of meetinghouses and often used as such in some areas.

Small buildings for the isolation of menstruating women existed in the western Carolines and continued in use in Yap until recent times.

ECONOMY

Production and technology. The islands were probably first settled by small groups of immigrants with few technical specialists among them. The prehistoric Micronesians had basically a subsistence technology with a few part-time specialists producing luxury items, such as personal ornaments and the shell and stone valuables used, mostly in the western Carolines, for ceremonial payments.

Subsistence everywhere was based primarily on horticulture and fishing, with fishing somewhat more important in the low islands and horticulture more important in the high islands. Domestic animals were limited to dogs and chickens, found only in some areas.

Food crops

Rice was grown in the Marianas but later was largely replaced by maize, introduced from Mexico by the Spanish conquerors. Coconuts were used everywhere, both as food and for other purposes, such as thatch, lumber, and cordage. Some form of taro, either the true taro (*Colocasia*) or the giant swamp taro (*Cyrtosperma*), was probably cultivated everywhere except on some of the drier low islands in the east, where the groundwater tended to be too brackish. The breadfruit tree, which produces a large starchy fruit in abundance in the early summer, was also widely distributed. Bananas were an important food crop on the high islands. Tropical yams (*Dioscorea*), which produce large starchy tubers, are found today on a number of the high islands but are of greatest importance as a cultivated crop on Ponape, where they have high prestige and provide an important source of food in the winter. The sweet potato and cassava, introduced by Europeans, now serve as alternate subsistence foods on some of the high islands. On some of the drier atolls in the Marshalls and the Gilberts, the pandanus tree is a major subsistence crop. The edible fruit of some cultivated varieties, in appearance resembling a pineapple, contains starch and sugar that can be made into a flour and stored. Other varieties have large edible nuts. Some varieties of pandanus are cultivated for their leaves, used principally in making plaited sleeping and floor mats and sometimes for thatch.

Many kinds of fishing were practiced traditionally. Often there was a division of labour, with men using techniques requiring canoes in deep water, such as seines and line fishing, and women wading on the waters of the fringing reefs. Deep-sea trolling with sailing canoes was practiced mostly by low islanders. The low islanders also at times made expeditions to certain small, uninhabited reefs and low islands to fish and collect turtle and seabird eggs.

Cutting tools were made from stone and shell. Iron was introduced by Europeans, although some iron tools may have been obtained from Oriental sources in western Micronesia.

Simple pottery has been found only in the western high islands of Micronesia: the Marianas, Palau, and Yap. Suitable clay apparently exists in Truk, Ponape, and Kusaie, but if pottery was ever introduced to these islands, it was not widely used.

Ocean travel

All Micronesians, even the high islanders, relied often on water travel, although the high islanders used canoes principally in the sheltered coastal waters of their home islands. Micronesian canoes had a single hull with one outrigger. Canoes used in protected waters were often simple dugout canoes, but the oceangoing canoes, found

especially in the central Carolinian atolls and the Marshalls, had sides built up of irregular planks that were caulked and sewn together with cord made from coconut-husk fibre.

Some of the atoll dwellers regularly went on trips requiring several nights in the open sea. Extra provisions were usually taken along as gifts and for emergency needs if the canoe were blown off course. In historic times vessels blown off course have ended up in a variety of islands inside and outside of Micronesia. Conversely, vessels from other areas of the Pacific and from the Orient have been blown to Micronesia. It seems clear that, prehistorically, there was communication in chain fashion from the Ellice Islands and western Polynesia through the Gilberts, the Marshalls, and the Carolines to Palau and the southeast atolls and beyond them into the northern Moluccas in Indonesia. Probably there was also deliberate communication across the larger gap between the central Carolinian atolls and the Marianas; in the latter there is a long-established population of cultural Carolinian immigrants retaining their original language.

Property and exchange systems. The most important property among the Micronesians was land. Land-tenure customs varied considerably from island to island, even among those with related cultures.

Land was owned primarily by extended-family groups or lineages. Individuals acquired use rights to particular plots through their kin connections, acknowledging the rights of the group by periodic offerings of first fruits to the kin-group head. Often a community chief and sometimes a superior chief received offerings from each household or larger kin group holding land under his jurisdiction. In some areas, especially in eastern Micronesia, strong chiefs effectively exercised the power of eminent domain and confiscated land from those who fell out of their favour, awarding it to loyal followers.

Since land was a scarce resource in much of Micronesia, various mechanisms developed to govern its distribution. In areas where matrilineal inheritance of land rights was practiced, there were usually provisions that, if necessary, children of the men of the lineage could also inherit use rights to the land worked by their fathers or receive a share outright if the lineage had more than it needed. In general, land was not sold, but it might be given in payment for medicine and nursing services or as compensation for an injury.

In the larger high islands, interior areas not under regular cultivation were considered community property and used for collecting wild food and for temporary gardens. In the atolls practically all land was owned by one or another family group, even small islets with four or five coconut trees.

Feasts as exchange rituals

Large feasts were common throughout Micronesia. The people assembled much food, offered it to the chief, and he in turn redistributed much of it to the people. Sometimes kin groups or communities held alternating semicompetitive feasts in which each tried to outdo the other. Feasts were held to commemorate important transitions in the life cycle, especially marriage and death. A kind of delayed exchange might take place at these feasts, certain relatives being obliged to make presentations of food and goods in return for past or anticipated services.

Substantial payments were sometimes made to practitioners of traditional medicine, especially when the practitioner was not a close relative. Such payments consisted of food and other goods. The precise amount was generally left to the family of the patient, with the understanding that a stingy family might not get the most energetic and effective treatment.

In Palau and Yap, shell and stone valuables were used in transactions at life crises and for certain other specific compensations, such as payment for injuries. These are often called money, but their use was much more limited and specific than that of most money. The best known of them were the large stone discs that the men of Yap manufactured on Palau and carried home in sailing canoes.

A certain amount of trade developed between the low islanders and high islanders, the low islanders providing handicraft products that the high islanders could have made if they had needed to, and the high islanders providing things obtained more easily there. This trade was especially well developed between the low islanders of the central Carolines and the high islanders of Yap and Truk. The most important high-island export was turmeric, used for medicinal and cosmetic purposes, which formed a bright orange body paint when mixed with coconut oil. The low islanders provided shell beads, plaited pandanus mats, and coarse cloth woven from banana or wild hibiscus fibre and used for women's wraparounds and for men's loincloths. Trade was often with particular partners who regarded themselves as distant relatives. An important function of the trade was to provide the low islanders with aid and a temporary dwelling place when their islands were devastated by the periodic typhoons.

RELIGION AND ART

After 1900 Christianity became well established in most major centres in Micronesia. Native religions were no longer practiced in their full original form, although in Yap and some atolls of the central Carolines, traditional religion continued to be practiced until the middle of the 20th century. Missionaries and travellers have left descriptions of certain aspects of the traditional religions, but there is no complete and systematic account.

The basic patterns of religion were probably similar throughout most of Micronesia. The Micronesians were polytheists, believing in several high gods, a large number of spirits attached to specific localities or performing specific functions, and a number of ancestors and deceased neighbours who could sometimes make contact with their living descendants and friends. Practices connected with these three major categories of supernatural beings tended to be distinct and to be handled by different specialists, although a figure might gradually shift from one category into the other.

Micronesians generally believed in at least three vertically arranged levels of the universe: the earth proper, the underworld, and heaven or the sky world. Some of them may have believed in multiple heavens, as did people elsewhere in Polynesia and Indonesia, but the details are now unclear.

Micronesian origin myths generally tell of the origin of particular islands or descent groups or the initial discovery or later conquest of an island by the ruling descent group. The notion of the supernatural creation of the whole human species or of the whole world is either not found or little emphasized in Micronesian mythology.

Ceremonies for the high gods appear to have been principally seasonal offerings of first fruits, performed often in private by a specialist priest with a few helpers. Special appeals to the high gods were probably also made at times of community crisis, such as wars or typhoons. The priests were very likely relatives of the ruling chiefs. The lesser spirits were called on by magicians for specific purposes, most notably for the diagnosis and cure of disease but also for such purposes as success in fishing and agriculture, control of weather, success in love, and prowess in athletic contests, battle, canoe building, house building, and other difficult pursuits.

Communication with spirits

Ancestral spirits were often contacted in dreams and in the trances of spirit mediums, as were the high gods and other nonhuman spirits. They would give people information about the causes of disease, deaths, and other misfortunes and would sometimes prescribe new medicines or new varieties of magic. At times, the spirit mediums would also order their human protégés to perform songs and dances for the entertainment of the spirits, to win their goodwill and insure the prosperity of the community.

Micronesia has little permanent visual art. The best known consists of the painted bas-relief scenes on the timbers and gables of the men's houses in Palau. Stylized wooden masks and human figures have been found in the Mortlock Islands, south of Truk. There was some decoration. Decorative lashings of light and dark sennit twine are found in important buildings. Painting and decoration of canoes is common, following traditional patterns

for each island. Patterns were woven into clothing. Wooden bowls were carved and painted with aesthetic intent.

Much of the most-developed traditional art of Micronesia was ephemeral: body painting and ornamentation with garlands of flowers, singing, dancing, and the recital of myths, tales, and poetry. Oratory was also an important traditional art and still survives. A good orator was expected to make mythological and historical allusions and to use special figures of speech.

THE CHANGING CULTURE

Factors in cultural change. Micronesian culture has been heavily influenced by contacts with outsiders: the Spanish conquest of the Marianas at the end of the 17th century; the entry of European whalers, traders, and missionaries into the rest of the area in the 19th century; the Spanish conquest of the Carolines; German occupation of the Marshalls; the British occupation of the Gilberts at the end of the 19th century and beginning of the 20th; the Japanese occupation of former German and Spanish holdings at the start of World War I, extending temporarily to Guam and the Gilberts in World War II; and the U.S. occupation of the Japanese holdings at the end of World War II under a United Nations trusteeship.

Each of these waves has resulted in substantial racial and cultural admixture in at least parts of the area, especially in the larger and more populous islands but extending out to the small atolls that have traditionally looked to the larger islands as models.

Effects of outside contacts

How one evaluates the effects of these contacts on modern Micronesian culture depends on whether one is interested in the preservation of traditional local cultures or the strengthening of pan-Micronesian culture. Christian missionaries, for instance, have often been less than sympathetic to the local religions and mores. On the other hand, they have encouraged the development of a sense of Micronesian identity by teaching foreign languages with which Micronesians from different islands could speak to each other, by sending out Micronesian missionaries and catechists to other islands and thus increasing inter-island contact, and by encouraging a wider feeling of human brotherhood.

Foreign trade has caused the disappearance of some traditional crafts, such as weaving and pottery, by providing superior imported products and by diverting labour to export products such as copra (dried coconut meat). At the same time, trading vessels have often carried Micronesians as passengers or crewmen to islands in Micronesia and beyond. Internal mobility within Micronesia has led to inter-island migration and to a sharper awareness of the similarities in much of the content of Micronesian cultures. Foreign trade has brought new similarities in ways of life based on the adoption of foreign products and practices. Thus, all the main population centres now use outboard motors for trips in coastal waters, motor scooters and automobiles on land, and many other products of modern technology.

In the United Nations Trust Territory, administered by the U.S., and in Guam, the government payroll has in recent times become a much greater source of cash income for Micronesians than foreign trade and tourism. Nevertheless, these funds are dispensed mainly at the large population centres, while most of the low-island communities continue with much of the traditional subsistence economy, purchasing clothes, tools, cigarettes, and some luxury foods from the sale of copra. Royalties from the mining of phosphate are locally important for the small independent island of Nauru and the island of Angaur in Palau; they have also been important for the Gilbertese inhabitants of Ocean Island, who have moved away to Fiji to accommodate the mining.

The imposition of foreign political authority on Micronesia has tended to lessen the power of the traditional chiefs and political systems, although the foreign rulers have often paid lip service to the ideal of preserving indigenous culture and social organization. In the Trust Territory the Japanese and Americans have tried to establish more or less uniform elected local governments, each usually corresponding to what was originally a small independent society. These elected governments were intended to take care of modern political life, including the execution of policies and laws prescribed by the territorial government, while the traditional political system was left to take care of "social" matters. These arrangements have often been understood differently by the Micronesians involved; the elected local governments have often been regarded by the electorate primarily as a sort of buffer to minimize the interference of the foreign territorial government in local affairs.

In the 1960s the Trust Territory government sharply expanded its budget and personnel. More and more Micronesians were becoming involved in government above the local level. The new Congress of Micronesia was given powers to formulate territorial policy and laws.

U.S. and Japanese immigrants

So far, most of Micronesia has been spared the effects of massive permanent immigration of outsiders. The greatest concentration of non-Micronesians is on the island of Guam, which the United States took over at the end of the Spanish–American War in 1898. Many military personnel and their dependents live on Guam, and other Americans have settled there to do business. Even so, the local government is mainly in the hands of the native Chamorros, who also retain considerable real estate in the southern part of the island.

In the Trust Territory a large immigration of civilians and military personnel occurred during the Japanese period, principally to the larger high islands. They were all repatriated at the end of World War II, when the United States took over the area, and only a few have returned. Non-Micronesians have not been allowed to purchase land during the U.S. period, and until the 1960s most foreign nationals were prevented from entering the area on grounds of military security. In the Gilbert Islands, controlled by the British, there has been little foreign immigration; the islands have a small land area and are densely populated. Some foreign coconut plantations were established at the turn of the century.

Migration

Substantial movements of indigenous population have occurred within Micronesia. In general, people from the low islands have tended to move to the nearest high islands because of population pressure at home and because of the schools, missions, scientific medical treatment, salaried employment, and other advantages of the high islands. The island of Saipan in the Marianas has a substantial population derived from the atolls of the central Carolines over a long period. Ponape has a substantial number of immigrants from nearby atolls and from the Mortlock atolls to the south of Truk. These and other internal movements dissipate the local cultures of the migrants but at the same time work toward a generalized Micronesian culture.

Government schools and mission schools are important factors. Some kind of elementary school is available in every sizable community in Micronesia. Secondary school education is available for selected children in the main population centres. There is now a junior college on Ponape and a small university on Guam. Many students from the Trust Territory have received scholarships to study at the college level in Hawaii or the mainland United States, and a few have gone on to graduate study. Trade schools are relatively undeveloped in the Trust Territory, one exception being the Ponape Agricultural and Trade School established by a Jesuit missionary. Medical practitioners in the Trust Territory and the Gilberts are trained at a British school in Fiji. Peace Corps volunteer teachers provided much assistance in teaching English in the elementary schools of the Trust Territory, although their number was reduced following friction with the territorial government and with some local governments. Attending school takes up time the children formerly spent in learning their indigenous cultures, although while they are in school, the adults and older children are freed from taking care of them. Perhaps the most important single effect of the schools is to teach the pupils enough English to be able to watch motion pictures, listen to English-language radio programs on their transistor sets, talk to visitors, and read English publications—including mail-order catalogs, comic books, main-

tenance and repair manuals, and the Bible. Local radio stations are numerous, and there were television stations on Guam and Saipan in 1971.

Since the U.S. government relaxed its military security checks on entry into Guam and the Trust Territory, an effort has been made to develop tourism as the most promising source of external income for Micronesia. Commercial air service has also been improved. Many Japanese tourists have visited Guam and Saipan, and American and Japanese tourists have gone to the Carolines in increasing numbers. Tourists visit mainly the larger islands. The development of tourism is a controversial issue in Micronesian politics, since some leaders fear that its growth will relegate Micronesians to menial positions while the managerial positions will be held by non-Micronesians.

Decline of traditional political authority

Effects of modernization. All the Micronesian communities acknowledge the de facto authority of the administering powers: Great Britain for the Gilberts and the United States for the Trust Territory and Guam, with Nauru the only independent island. The power of the traditional local political systems has everywhere declined, but more so in some communities than in others. In the Marshalls, for instance, the ownership of land by many of the traditional chiefs has been officially recognized by foreign rulers; the bulk of the people on these islands are tenants of the chiefs. In contrast, on the high island of Kusaie, immediately to the west of the Marshalls, no traditional chiefs remain, even in name, and local authority is entirely in the hands of elected officials, church leaders, and family elders. With the decline of traditional political authority and the increasing availability of imported alcoholic beverages in the main population centres, there has been an increase in violence and petty crime.

Kin ties outside the immediate family appear to have weakened. The distinction between matrilineal and other relatives has also weakened in areas where it was formerly important. The restrictions on marriage between remote matrilineal relatives are being challenged and on some islands (Kusaie, Mokil) completely ignored. Kin groups are being replaced by the police as the agency for settling physical injuries and other personal wrongs.

Micronesians still speak of social relations in the idiom of kinship and attempt to trace or establish fictive kinship with close associates. But kinship is now often invoked to support relationships first established on some other basis, while formerly it was more common for an already existing kin group to decide to cooperate on some enterprise. In the modern situation it is an advantage to be able to trace kin relations equally in all directions: through father, mother, spouse, and spouses of close relatives.

Practically all Micronesians speak a Micronesian language as their first language. Many of them also speak one or more other Micronesian languages as well. English is now widely spoken by schoolchildren and young adults, and it is used in conversations between speakers of different Micronesian languages as well as in talking with resident Americans and outside visitors. Japanese is spoken by many older adults in the Trust Territory who learned the language during the period of the Japanese mandate.

Christianity and traditional religion

In religion most Micronesians are professing Christians, although traditional local beliefs in spirits and magic continue to be held concurrently in many areas. The people of the Marianas are predominantly Catholic; those of the Carolines, Marshalls, and Gilberts are a mixture of Protestant and Catholic. Protestants were the first successful missionaries in the latter area, beginning about the middle of the 19th century, and are still numerically dominant in many communities. Elements of the traditional local religions survive widely in traditional medical practice, which often attributes disease to supernatural causes. The traditional religions are practiced as major alternatives to Christianity in the central and western Carolines. The *modekngei* movement in Palau appears to be a modern revival of traditional Palauan religion, with strong emphasis on health and medicine.

Perhaps the greatest infusion of Euro-U.S. and Japanese culture has taken place in the arts and recreation. In the main population centres of the Trust Territory and in Guam, baseball is the most popular sport, having been introduced in the Trust Territory by the Japanese. Traditional music is heard much less than hymns and popular music of Western origin. Knowledge of traditional sports, dance, and music remains in most places, but these are performed more often at the instigation of curious outsiders than for members of the local communities. The prospective rise of tourism seems unlikely to encourage an authentic revival of traditional arts, crafts, and recreation. In their self-image, as reflected in their expressive culture, the Micronesians already stand with one foot in the modern world.

BIBLIOGRAPHY. Readable accounts of particular Micronesian cultures include, for high-island cultures, HOMER G. BARNETT, *Being a Palauan* (1959); and LAURA M. THOMPSON, *Guam and Its People* (1947); and, for low-island cultures, MARSTON BATES and D.P. ABBOTT, *Coral Island* (1958); and WILLIAM A. LESSA, *Ulithi: A Micronesian Design for Living* (1966). JOHN L. FISCHER, *The Eastern Carolines* (1957), surveys briefly a number of high- and low-island cultures. The role of outside administrations is presented for the Gilbert Islands a British possession, by ARTHUR F. GRIMBLE, *Return to the Islands* (1957); for the Trust Territory of Micronesia in the Japanese period by TADAO YANAIHARA, *Pacific Islands Under Japanese Mandate* (1940); and in the early American period by ROBERT TRUMBULL, *Paradise in Trust: A Report on Americans in Micronesia, 1946–1958* (1959). For additional references, see HUZIO UTINOMI, *Bibliography of Micronesia* (1952); and for some more recent work, the section on Micronesia in CLYDE R.H. TAYLOR, *A Pacific Bibliography*, 2nd ed. (1965).

(J.L.Fi.)

Microscope

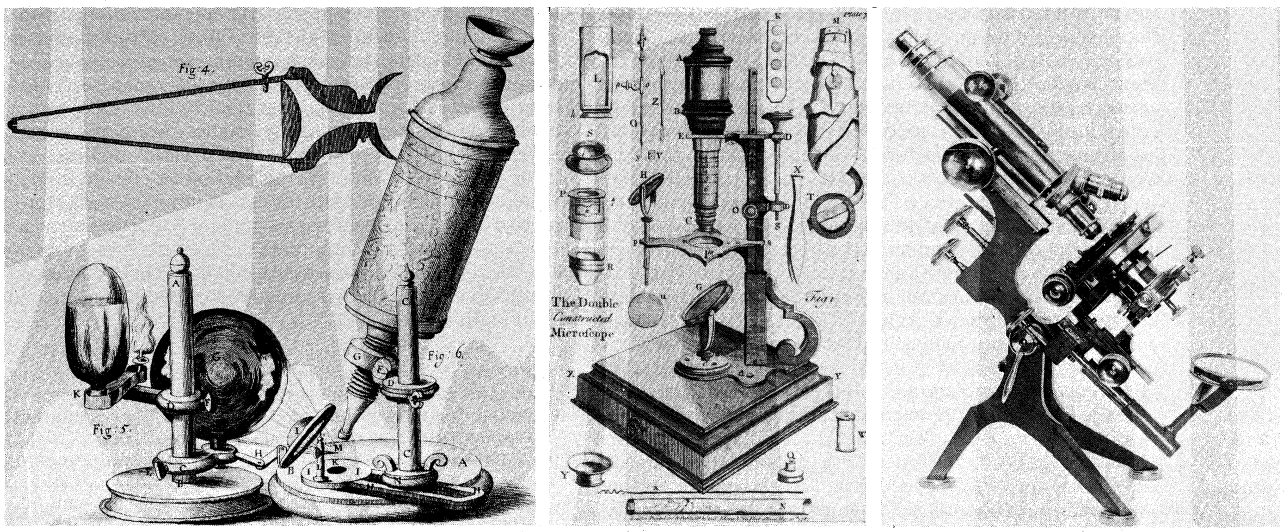
The microscope is an instrument for producing enlarged images of objects that are too small to be seen unaided. Such images may be viewed directly or with a viewing screen or photographic apparatus. Objects for microscopy may be either transparent or opaque. The compound light microscope (see below *Compound microscope*) can be used for viewing objects as small as bacteria, and the electron microscope (see below *The electron microscope*) extends the range to small viruses and large molecules.

HISTORICAL DEVELOPMENT

Simple microscope. The simple one-lens microscope may date from the middle of the 15th century, when low-power magnifiers were used for the examination of insects. From these early beginnings, glass lenses of increased power were developed, culminating in ground lenses of remarkable excellence produced by a Dutch microscopist, Antonie van Leeuwenhoek (1632–1723). In the 18th century, simple microscopes were produced in a cheap, portable form known as the screw-barrel type, which used a screw thread to focus the lens upon the object. Further developments of the simple microscope included perfection of mounts and microscope stands intended for specialized studies in botanical and other fields. After the introduction, in the 1830s, of perfected achromatic lenses, which rendered images free from colour distortions, simple microscopes lost favour except for very low power work, such as in dissection.

Compound microscope. The basic idea of the compound microscope, in which two or more lenses are arranged in such a way as to form an enlarged image of an object, probably occurred independently to several people at about the same time, in the period 1590–1609, in The Netherlands; Hans Jansen, his zon Zacharias, and Hans Lippershey have all been cited at times as its inventor. Shortly after that, microscopes made by opticians in Italy and in England were circulating widely.

Though no real advances were recorded in the development of the optical parts of the microscope during the 18th century, some mechanical variations were introduced. Brass superseded wood, cardboard, and leather as a material for construction of the stand, and better mechanical means were introduced for focussing.



(Left) Hooke's microscope from *Micrographia* by Robert Hooke, 1665. (Centre) Cuff microscope from the second edition of *Essays on the Microscope* by George Adams the Younger, 1787. (Right) Watson microscope made to the specifications of Henri van Heurck by W. Watson & Sons, London, 1904.

By courtesy of (left) the Wellcome Trustees, London, (right) the Medical Museum of the Armed Forces, Institute of Pathology, Washington, D.C.

The image produced by the lenses of early compound microscopes suffered considerably from chromatic aberration; *i.e.*, objects in the field of view appeared to be surrounded by fringes of colour. Experiments with telescope lenses had shown that images free from colour could be produced by combining lenses of different dispersions. Dispersion is the variation of refractive index with wavelength, that for blue light being greater than for longer wavelengths in the red region of the spectrum. The telescopes used a combination of a convex (outward-curved) lens made of optical glass of low dispersion (crown glass) together with a concave (inward-curved) lens of flint glass with a high dispersion.

Achromatic microscope lens. As glass technology advanced, application of the principles used for telescopes to microscopes became possible; the first achromatic microscope lenses were constructed in The Netherlands at about the end of the 18th century. Many more achromatic lenses of low power (and often of dubious quality) were produced in the period between 1810 and 1820. High-power achromatic lenses proved much more difficult to construct, because the accumulation of spherical aberrations or distortions caused by curvature in the lens glass often negated the effects of achromatism.

At that time lenses were still being designed empirically, and introduction of the use of test objects, such as the markings on the wing scales of moths and butterflies, enabled some progress to be made with the higher powered lenses. It was with the work of the English microscopist Joseph Jackson Lister (1786–1869), published in 1830, that construction and design of achromatic microscope lenses was placed on a theoretical basis. Lister showed how two separate achromatic lenses could be combined to act as a single lens so that the image rendered would be completely free from chromatic aberration and from distortions caused by curvature of the lens. This discovery opened the way to the construction of high-power microscope lenses in the next 50 years.

Abbe's contributions

For both his contributions to the theory of the instrument and his practical designs, a German physicist, Ernst Abbe (1840-1905), stands out as the foremost figure in the development of the microscope in the latter part of the 19th century. Abbe placed the theory of microscopic image formation on a sound basis by emphasizing the importance of the light diffracted by the object in resolving (making distinct) its fine detail. He introduced the concept of numerical aperture as an expression of the resolving power and light-gathering capacity of a microscope lens, thus permitting its direct comparison with other lenses. Abbe also developed oil-immersion lenses, in which a drop of oil was placed between the object and the front element of the lens. Very soon this type of lens, with its much improved resolving power and high-quality images, was an essential part of the microscopist's equipment. Later, in conjunction with the German firm of Carl Zeiss, Abbe introduced a whole new series of excellent, highly corrected lenses, which rendered the highest quality images yet obtainable. These distortion-free (apochromatic) lenses allowed full practical realization of the theoretical resolving power of the optical microscope by the end of the 19th century.

Developments so far in the present century have included introduction of larger stands with built-in illumination systems and higher standards of accuracy of construction. Especially notable is phase-contrast microscopy (see below *Phase-contrast microscope*), introduced in 1934, which made it possible to examine living cells and tissue specimens without staining them. Other recent trends are the development of flat-field objective lenses for photomicrography and of methods of using computers for the analysis of microscopical images.

Interest has continually centred on the increase of resolving power of the microscope; the work of Abbe showed that this was possible only by increase in the numerical aperture of the optical system (which was not feasible beyond a certain limit) or by the use of a shorter wavelength for the illuminating light. This latter possibility led to the use of ultraviolet radiation as an image-forming agent, but there were technical difficulties with this type of microscope. Little progress in the use of shorter wavelengths was made until the development of the electron microscope (see also OPTICS, PRINCIPLES OF).

THE SIMPLE MICROSCOPE

A simple microscope, a single positive lens or lens combination, is often known as a magnifier, or loupe. Supplied in a mount furnished with a handle, such a lens is commonly called a reading glass. Magnifiers are often furnished in a hinged mounting between two cover plates, so they may be conveniently carried in the pocket. Occasionally, they are mounted on an arm above a simple specimen-support stage so as to serve as a dissecting microscope. Magnifiers supplied in a short tubular mount, designed to be held in place in front of the eye by contractions of the facial muscles around the orbit, are used for fine instrument work and are called watchmaker's eyeglasses. The apparent size of an object viewed directly by the eye depends upon its distance from the eye. As an object is brought nearer to the eye, the angle which it subtends at the eye increases, and so the retinal-image size (and hence the apparent object size) also increases. This process continues until the object has been brought

so close that the eye no longer forms a clear image. This point is generally termed the nearest distance of distinct vision (in a young adult, about 25 centimetres, or 10 inches), and further movement of the object toward the eye will cause the image to be blurred. If, however, a convex lens is placed between the object and the eye, the sharpness of the image may be restored. If the object is at or within the distance between the lens and the point at which parallel rays passing through the lens converge, or focus (focal length), then light rays entering the lens from the object are bent, or refracted, toward the optical axis (to the centre). If these ray directions were prolonged backward, they would meet at a point in space behind the object. It is here that the brain imagines an enlarged image of the object to be located. Termed a virtual image, this image is magnified and erect.

Field of view. Maximum apparent magnification is reached when the virtual image of the object lies at the nearest distance of distinct vision. The area of the object that is seen through a magnifier (the field of view of the magnifier) is related directly to the diameter of the lens and inversely to the distance between the lens and the eye. Thus, to achieve the largest possible field of view with any given lens, the lens should be held as close as possible to the eye. The position of the object should then be adjusted to obtain the sharpest image. In practice, simple magnifiers suffer considerably from aberrations that affect the peripheral parts of the field of view; curvature of the field, distortion, and lateral chromatic aberration (different colours produce different image sizes) are principally involved.

The simplest form of low-power magnifier is a single biconvex (convex on both sides) or plano-convex (convex on one side and flat on the other) lens. Such lenses have a wide field of view; the eyeball rotates in its socket in order to scan the whole field. Spherical aberration, coma (variation in magnification over the area of a lens), and lateral chromatic aberration are not especially troublesome at the low magnifications involved, but there is appreciable curvature of the field and pronounced pincushion distortion (variation in image magnification radially outward from central point of image). In high-power magnifiers (up to 20 magnitudes, or 20 ×), the field of view is limited, and the eyeball does not rotate to scan it. The chromatic and spherical aberrations of this type of magnifier must be very carefully corrected. Distortion is often troublesome, because the distance between the eye and the lens is no longer small in comparison with the focal length of the magnifier (as it is in lower power lenses).

Types of magnifiers. Four principal types of high power magnifiers are recognized.

Stanhope lens. The Stanhope lens (Figure 1A) consists of a solid block of glass with one polished plane (flat) surface lying in the focal plane of the spherical convex upper surface of the lens. To be viewed, the object must be placed in contact with the plane surface; such lenses suffer from a large degree of distortion, and there is considerable lateral chromatic aberration.

Coddington lens. The Coddington lens (Figure 1B) has two convex surfaces but is characterized by a circular groove cut around the periphery of the lens in such a way as to act as a diaphragm. This improves the image quality, and these lenses are superior to the Stanhope type.

Wollaston doublet. The Wollaston doublet (Figure 1C) uses two plano-convex lenses, with their plane faces pointed inward. The lenses are separated by a distance calculated so as to allow a good correction of the lateral chromatic aberration and the spherical aberration and are fitted with a central diaphragm. This type of lens, when used as a magnifier, suffers from an inconveniently short working distance and a restricted field of view.

Steinheil and Hastings magnifiers. Achromatic doublets or triplets are now almost universally used for practical magnifiers; the best known are due to Steinheil and to Hastings. Lenses of this type generally consist of a biconvex (convex on both sides) lens of crown glass cemented between and enclosed by two lenses of flint glass

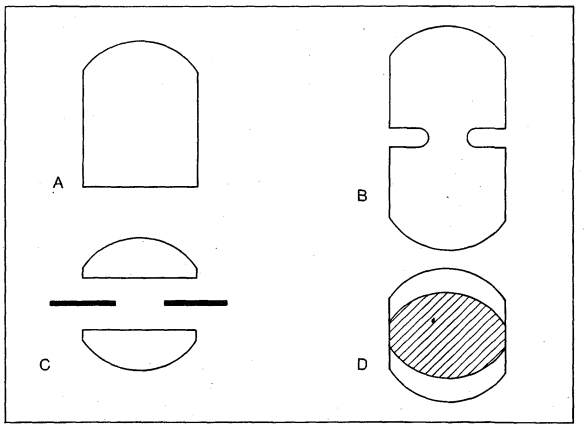


Figure 1: *Various magnifiers.*
(A) Stanhope lens. (B) Coddington lens. (C) Wollaston doublet.
(D) Steinheil achromatic triplet.

(Figure 1D). These cemented triplets are well corrected for spherical aberration and coma (for rays up to 25° from the optical axis), and their distortion is low. At the same time, they produce a very bright image and have a conveniently long working distance.

THE COMPOUND MICROSCOPE

A compound microscope has an objective lens (the lens closest to the object being viewed) and an eyepiece so mounted that they may conveniently be moved relative to the specimen to focus the image. An illumination system, usually added, consists of a condenser-lens system to focus the light onto the object and either a mirror with an external light source or, as is more common today, a built-in light source (commonly a low-voltage tungsten bulb).

In the compound microscope, magnification depends on the objective lens and the eyepiece. The objective lens produces a real image at some finite distance behind its back focal plane, a distance which is known as the optical tube length.

Compound microscopes may be provided with a single body tube for viewing with one eye; such monocular instruments are used for teaching and for work that is not extended or too exacting on the eyes. The more elaborate microscopes used for research purposes or for long periods of study are generally binocular (two eyepieces fitted with some means of distributing the light between them). A typical monocular microscope would normally be provided with three objective lenses with focal lengths of 16, 4, and 2 millimetres, giving initial magnifications of about 10, 40, and 100 magnitudes, respectively. If two eyepieces were provided with a magnifying power of six and 10, for example, the range of total magnification would extend from 60 to 1,000 magnitudes.

Monocular microscopes

Body tube. The body tube, with its associated focussing mechanism, is carried on the arm of the microscope, which is in turn supported by a horseshoe or tripod base. Attached to the arm, at a right angle to the optical axis, is the stage on which the object is placed. Below the stage is the substage fitting, which carries the condenser lens (used to focus light on the object) and either the mirror (if fitted) or the built-in light. The simplest microscopes have no substage. Instead, they have a plain object stage and a body tube fitted with only a coarse focussing movement. A more elaborate instrument has, in addition, a substage with its own centring and focussing movements. The body tube has both coarse and fine focus and a revolving turret at the bottom fitted to carry several objective lenses so that any one may quickly be brought into the optical axis. The largest and most elaborate research stand is of much sturdier construction and has an illuminating system built into the base of the instrument. A mechanical stage carries the slide and allows it to be moved easily and precisely. A binocular tube is provided for viewing with, possibly, a third, monocular tube permanently fitted to carry a camera. In the past, microscopes had their focus controls mount-

ed on the upper part of the arm. The current trend is to place these, together with the stage-movement knobs, at a low level so they may be operated without moving the hand and arm from the table.

For many years the horseshoe form of foot was popular, though the tripod foot was preferred in England. With the introduction of the built-in light source, both of these have given way to construction in which the arm and foot are solidly joined together, the foot being in the form of a box.

The body tube is mounted on a slide or slides to allow it to be moved for focussing. The coarse focus is actuated by a rack-and-pinion mechanism, while the fine focus, working on a slide with a much more restricted range of movement, is worked either by a screw of very fine pitch that has its motion converted and translated to the slide by a lever or by some cam or bell-crank arrangement. The fine focus must have a very free action but one that operates without any appreciable backlash. In some modern stands, in order to achieve maximum rigidity, the coarse- and fine-focus mechanisms act on the stage instead of on the body tube. Such an arrangement is easily adapted to the provision of low-level controls and to the fitting of concentric coarse- and fine-focus control knobs.

Specimen stage. The specimen stage holds the object under examination and must be fixed accurately at right angles to the optical axis of the microscope. In many microscopes, a simple stage with a central hole and spring clips to hold the slide suffice, but research instruments are often provided with either a detachable or a built-in mechanical stage. This device holds the slide and allows it to be moved in two directions by screw threads. Such stages allow the slide to be moved very precisely and systematically over its whole area, and, as they are often fitted with graduated vernier scales, they enable the position of objects to be located and recorded precisely. For work with polarized light and photomicrography, it is an advantage to have a mechanical stage that also allows rotation of the specimen around the optical axis of the microscope.

The objective lenses of the microscope are fitted to a disk of metal known as the nosepiece, which is accurately centred to the microscope axis. Rotation of the nosepiece allows any of the lenses to be brought quickly into use, and the accurate alignment, or centration, ensures that the object will still remain approximately in the centre of the field of view. If more accurate centration is required, it is usual for each objective lens to be mounted on its own individual carrier, which is provided with centring screws.

The substage

The substage carries the condenser and the mirror, if fitted. It slides in a rackwork mechanism to allow adjustment of its position. In some of the cheaper instruments, however, the substage is fixed to the underside of the stage and consists of a helical mount in which the condenser slides. An iris diaphragm, a device which opens and closes to control the size of the illuminating cone of light, is often carried on the substage, rather than on the condenser itself, although this latter arrangement (which is found in the more expensive instruments) is preferable. One essential feature of the substage of any instrument intended for serious work is the provision of screws that allow the condenser to be centred with respect to the optical axis of the microscope.

The built-in light source of large modern microscopes is often mounted in the base of the instrument and comprises a small low-voltage bulb provided with collector lenses and arrangements for focussing and centring the light source. The light is reflected into the condenser either by a front-surface mirror or by a prism. The binocular body employs an arrangement of prisms to present the light from the microscope objective to each eye.

Objective lenses. The value of a microscope lies in its ability not only to magnify objects but also to make their fine details visible. These two features depend for the most part on the quality of the lens that forms the primary image. Such an image must be sharp; the aberrations that reduce the sharpness must therefore be reduced to a minimum. Microscope objective lenses originally were single biconvex lenses that suffered severely from aberrations; nowadays all objective lenses, even those of the lowest power, are corrected as fully as possible. Chief corrections are for chromatic aberration (which produces colour fringes around the edges of the image of an object) and for spherical aberration (which is caused by the curvature of the lens). Spherical aberration causes a hazy, indistinct image, lacking in contrast. Chromatic aberration may be corrected by combining a concave negative element made from glass of a different dispersive power from that of the original convex lens. At the same time, the curvatures of such a combination may be calculated to correct part or all of the spherical aberration as well. Though such corrections are relatively easy with a lens of low magnification, satisfactory correction requires increasingly complicated lens design as power and numerical aperture (discussed below) increase. This usually involves addition of more elements. The design and construction of lens systems of high power and aperture are quite difficult, and these still remain the most costly parts of any microscope.

In addition to the primary aberrations just mentioned, coma, astigmatism, distortion, and curvature of field also have to be corrected by the designer of the lens. Coma causes magnification to differ according to the zone of the lens through which the rays pass, so that portions of the image axis are badly distorted. Astigmatism, if present, results in different appearances in different focal planes. A pinhole would appear to be a radial line at one focus and a tangential line at another. This aberration may cause misinterpretation of the shape of objects if severe at the centre of the field of view; a slight amount of astigmatism is tolerable, however, at the periphery of the field of view. Curvature of field causes the image of a plane object to appear on a curved surface. This means the centre of the field of view may be in sharp focus while peripheral areas are blurred, or vice versa. This defect increases in severity with increase in the power of the objective lens. Though not an intolerable drawback in visual work, curvature of field is undesirable when photographing through a microscope. It may be largely corrected by the use of a special eyepiece, but a more satisfactory solution involves the use of specially designed flat-field objective lenses, which are now commercially available (see also OPTICAL ENGINEERING). Empirical studies in the 19th century led microscopists to realize that the ability of their microscopes to resolve fine details was in some way connected with the angle of the cone of light rays accepted by the objective lens. This was termed the angular aperture. Ernst Abbe, in 1873, showed that this was not entirely accurate, and he was able to demonsrtate that the detail resolved was proportional not to the angular aperture but to the sine of half this angle (α). Abbe gave the name numerical aperture (N.A.) to this mathematical relationship and showed that in the case of a dry lens—*i.e.*, one in which the cone of light rays from the object passes through air—the numerical aperture = sine α. In an immersion system, in which oil or sometimes water or glycerin is placed between the front lens of the objective and the specimen, the numerical aperture is increased and the value of sine α has then to be multiplied by the value of the refractive index of the immersion medium.

Numerical aperture

The numerical aperture of a lens thus is a direct measure not only of its resolving power but also of its light-gathering capacity. Light-gathering capacity, however, is not of great significance in microscope objective lenses, and high numerical apertures are used for the consequent gain in the resolving power that accrues. A further advantage of the use of the concept of numerical aperture is that the fine-detail-resolution performances of two lenses may be directly compared by the use of their respective numerical apertures.

Microscope objective lenses may be classified according to either their focal length or their type and degree of corrections. The most common focal lengths are 75, 50, and 25 millimetres for low-power lenses, 16 and eight millimetres for medium power, and four and two millimetres for high powers. Lenses may also be classed as achromats, semi-apochromats, or apochromats, in as-

cending order of accuracy of corrections; flat-field objective lenses may be produced in any of these grades. Some of the features of the common lens types are summarized in the table.

Characteristics of Various Types of Microscope Objectives

type	focal length (mm)	numerical aperture	initial magnification	working distance (mm)
Achromatic	50	0.11	2.7 ×	50
dry	25	0.18	5.5 ×	16
	16	0.25	10 ×	5.0
	4	0.65	40 ×	0.5
Achromatic	4	0.85	40 ×	0.35
oil	2	1.25	100 ×	0.1
Apochromatic	16	0.30	10 ×	4.85
dry	8	0.65	20 ×	0.75
	4	0.95	47 ×	0.18
Apochromatic oil	2	1.40	90 ×	0.06
Semi-apochromatic	16	0.30	10 ×	4.0
dry	8	0.60	25 ×	0.54
	4	0.75	40 ×	0.33
Semi-apochromatic	3	1.25	63 ×	0.65
oil	2	1.30	100 ×	0.12

Achromatic objective lenses. The commonest types of microscope objective lens, achromatic lenses, are corrected so that rays of two widely differing wavelengths (usually in the red and the blue) are brought to a common focus. This correction is usually carried out for an intermediate zone of the lens; if it were done for the marginal zone at the edge of the lens, then the centre would be only poorly corrected, and vice versa. A suitable compromise, therefore, is achieved by correcting for the intermediate zone, which then results in minimal aberration in both the centre and marginal zones. As only two colours are fully corrected, there is at times some remaining colour, known as the residual colour or the secondary spectrum. This appears as a fringe around the edge of objects with sharp contrasts, but it is usually not prominent.

Residual colour

Achromats are also corrected for spherical aberration and coma. This type of objective lens is much used on microscopes intended for teaching and for routine visual observations.

Apochromatic objective lenses. The invention of new types of optical glass in the 1880s, together with the use of the natural mineral fluorite in lens making, enabled Abbe to develop a new type of microscope objective lens in which the colour correction was greatly improved but at the cost of a considerable increase in lens complexity. An apochromatic lens usually possesses a greater numerical aperture than an achromat of equivalent focal length. Care is needed in the use of apochromats because the high power lenses have very short working distances. Expensive, they are ideal for critical work at the limit of resolution of the optical microscope and for photomicrography in colour.

All apochromatic lenses (and semi-apochromatic lenses, which are more common) suffer from a kind of error known as lateral chromatic difference of magnification. This means that the image will be larger in blue light than in red. Thus, dark objects with sharp contrasts at the periphery of the field of view appear fringed with red or orange on their outer margin and with blue on their inner sides. This may be eliminated by the use of special compensating eyepieces that have been computed to introduce a compensatory error of similar magnitude but of opposite direction. This cancels out the lateral chromatic differences of magnification and thus gives the best possible image.

Semi-apochromatic objective lenses. Sometimes called fluorite lenses, these are a compromise between the achromats, on the one hand, and the apochromats, on the other. This compromise allows a reduction in complexity and, hence, in cost. Fluorite and the new optical glasses are used in their construction, but there are fewer elements than in an apochromatic lens. The colour and spherical corrections are better than those of an achromat, the secondary spectrum being much less obvious and, since the number of points of meeting between air and glass (air–glass interfaces) in such a lens is fewer than in an apochromat, the image contrast is often surprisingly high. Again, the numerical aperture of a semi-apochromat is usually 15–30 percent higher than that of an achromat of equal focal length. The lateral chromatic difference of magnification noted above is present in this type of lens, so it is used with a compensating eyepiece to obtain the best possible results. Recently, makers have been deliberately introducing lateral chromatic differences of magnification into their achromatic lenses (which do not normally show this particular fault) in order to eliminate the need for two separate types of eyepiece.

Spherical aberration and the coverslip. With some high-aperture dry objective lenses the protective piece of glass, or coverslip, overlying the mount may introduce a surprising amount of spherical aberration. Though the effect is negligible with low-power objective lenses and may be disregarded, high-aperture apochromatic lenses will give a satisfactory performance only if this factor is taken into consideration. Lenses are normally computed for a definite thickness of coverslip (often 0.17 millimetre), but extra thickness of mounting medium and variation in the thickness of the slip itself may cause wide divergence from this figure. Errors introduced in this way may be compensated for by altering the tube length of the microscope, shortening it if the cover is too thick and vice versa. Alternatively, most high-aperture apochromats are provided with a rotatable collar on the lens mount that varies the separation between the elements of the lens and so adjusts the correction of the lens for spherical aberration. The performance of an oil-immersion lens is, of course, not affected by such variations in coverslip thickness, since there is no air between it and the front element of the lens.

Lenses intended for metallographic work and for studies on opaque objects are normally computed so as to work without a coverslip. In the majority of cases, these lenses are also designed to work with vertical illuminators and must be provided with special antireflection coatings at points of contact between air and glass.

Metallographic lenses

Flat-field objective lenses. All objective lenses of the achromatic, semiapochromatic, and apochromatic types suffer to some extent from curvature of the field, which becomes more marked as the power increases. Such field curvature is troublesome in photomicrography (photography through a microscope), and new lenses have been developed in which this fault has been largely removed. But this has made it necessary to increase the complexity of the lens. In one manufacturer's series of flat-field lenses, the curvature of field produced by the objective lens is cancelled out by a lens in the microscope tube that interposes an equal but opposite curvature of the field. The system is so designed that the same negative lens is common to all the objective lenses of that series, and, hence, it may be fixed in the microscope tube itself. Both apochromats and achromats are available in flat-field versions.

Objective-lens mounts. All microscope objective lenses are fitted into metal mounts that are attached to the nosepiece. To protect the optical elements from damage resulting from the lens inadvertently being brought into contact with the microscope slide, most modern lenses are in resilient spring mounts that compress on light contact with the object slide. Most objective lenses of any one maker are fitted into mounts in such a way that very little adjustment of focus is needed on changing from one power to another. Such lenses are said to be parfocal. Mounts are normally engraved with the magnification of the lens, its numerical aperture, and, usually, the focal length, the tube length at which the lens operates, and the thickness of coverslip for which the lens is computed. A coloured band around the barrel assists in rapid identification.

Eyepieces. The eyepiece forms a further enlarged final image from the real primary image furnished by the objective. The eyepiece may also carry pointers or trans-

Graticules for expediting measurements

parent measuring scales (graticules) and correct for residual aberrations in the image produced by the objective lens. This is important in the compensating eyepieces used with apochromatic and semiapochromatic lenses.

Huygenian eyepiece. Two types of eyepieces are in common use. The Huygenian eyepiece (Figure 2A), sometimes called a negative eyepiece, has a plano-convex field lens nearer the objective lens with a second, smaller plano-convex eye lens above it. The field lens is so called

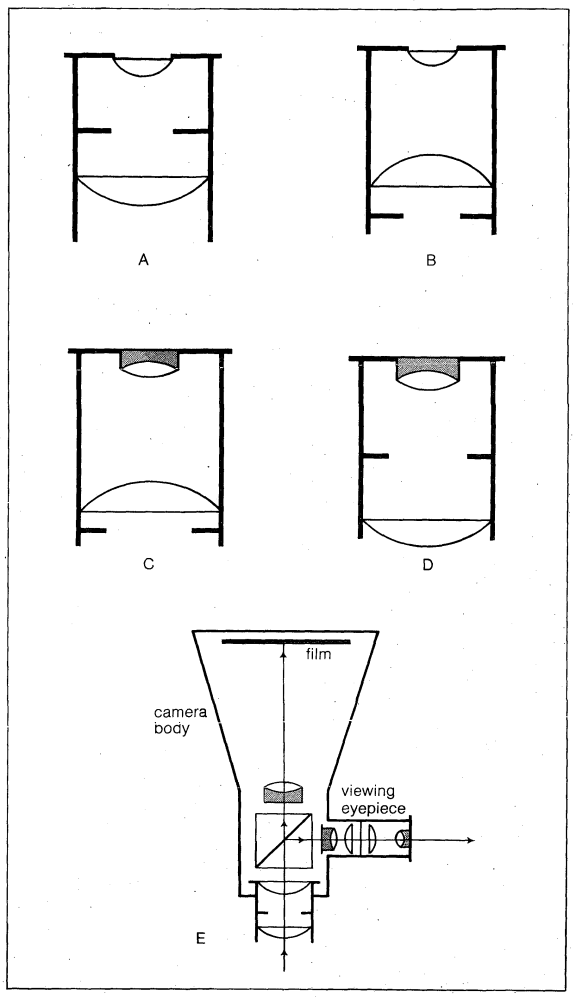


Figure 2: *Various eyepieces.*
(A) Huygenian. (B) Ramsden. (C) Achromatized Ramsden.
(D) Compensating. (E) Photomicrographic.

because it bends the pencils of rays from the objective towards the optical axis of the microscope without altering their individual convergence or divergence, thus increasing the angular field of view. After leaving the eyepiece, all the rays pass through a small circular area above the eye lens; this is known as the Ramsden circle, or eyepoint. Its diameter should approximate the diameter of the pupil of the observer's eye in order to obtain the largest field of view of maximum overall brightness. In the Huygenian eyepiece, the focal length of the field lens is about twice that of the eye lens, and the two lenses are separated by a distance roughly twice the focal length of the eye lens. It is usual to place a diaphragm between the lenses, which limits the field of view.

Ramsden eyepiece. The Ramsden, or positive, eyepiece (Figure 2B) has two plano-convex lenses of approximately equal focal length, arranged with the plane side of the field lens towards the objective lens and the plane side of the eye lens towards the observer. These are separated by a distance of approximately two-thirds their focal length. The focal plane of the eyepiece lies below the field lens, and the diaphragm is placed at this point. The Ramsden eyepiece is especially suited to carry measuring graticules and pointers. Both Huygenian and Ramsden eyepieces have uncorrected lenses, but the spherical and chromatic errors of each lens are arranged to cancel out.

Kellner eyepiece. The Kellner (achromatized Ramsden) eyepiece (Figure 2C) resembles the Ramsden very closely but has an achromatized doublet as an eye lens.

Compensating eyepiece. The compensating eyepiece (Figure 2D) used with apochromatic lenses also has an achromatized doublet for its eye lens; the corrections are matched to a specific objective lens. Formerly, eyepieces were available with adjustable compensation. These could be used with any apochromatic objective lens simply by adjusting the separation between the field and eye lenses of the eyepiece.

Photomicrographic eyepiece. For photomicrography, the eyepiece shown in Figure 2E is now much used, in conjunction with a miniature camera. A device called a beam-splitting cube produces two identical images, thus allowing the user to focus and observe the object up to and at the moment of exposure. Demonstration eyepieces allow a student and teacher to examine the same microscope field simultaneously by means of a similar type of beam-splitting arrangement. The addition of a pointer that is visible to both viewers is common in this type of eyepiece. A comparison eyepiece presents the images from two separate microscopes side by side, so that a detailed comparison of them may be made. Such a devise is of great value in many industrial applications and in forensic microscopy (*i.e.*, the study of material intended for use as evidence in legal proceedings).

Condensers, illumination, and resolution. Although very low power objective lenses may be used on the microscope with the illumination provided from the light cone of a concave mirror, all higher power lenses require a special system of lenses below the stage to provide a cone of light rays of sufficient numerical aperture to match that of the objective lens in use. Such a lens system is called a condenser. It is usually in a mount that can be moved along the optical axis in order to focus the condenser and that has spring-loaded screws to centre the condenser to the optical axis. Other functions of the condenser are to provide sufficient light on the object to allow the use of high magnifications and, in some cases, to provide special types of illumination. Examples of this are the oblique illumination and hollow cone illumination, sometimes used to obtain contrast in unstained or transparent preparations.

For low powers, a simple convex lens (sometimes called a spectacle lens condenser) may be used. This is preferable to the use of a concave mirror, because the condenser is provided with a diaphragm that allows some control over the working aperture of the system. The commonest type of condenser is a two-lens system known as the Abbe condenser. It is cheap, simple, and much used for visual observations with routine and student microscopes, though it suffers very badly from chromatic and spherical aberration, which often make it impossible to utilize the full aperture of illuminating cone provided by the condenser, which in turn makes it impossible to use the full aperture of the objective. An improved three-lens Abbe condenser has slightly better spherical corrections. For certain kinds of careful work, an achromatic condenser is essential. Such condensers are corrected for both spherical and chromatic aberration.

Abbe condenser

Illuminated field. The size of the illuminated field is a function of the focal length of the condenser. Thus, when low-power objective lenses are used, it is advantageous to employ a long-focal-length condenser to obtain a fully illuminated field of view. The same end may sometimes be achieved, however, by removing the top lens of a standard condenser, which then works at a reduced aperture but with an increased focal length. A pancratic (zoom) system, which permits change of focus from a large illuminated field at low power to a small area at high power may be employed. Alternatively, an auxiliary lens may be fitted in the light path to achieve the same end.

If the condenser is arranged to provide a hollow cone of light rays which are of such an obliquity that they cannot

enter the objective lens directly, then only the light diffracted by the object enters the lens and participates in image formation. The effect of this is that the specimen appears bright upon a dark background. Dark-ground microscopy gives images of high contrast but requires an intense light source and an accurately centred optical system.

Objects may be examined with either transmitted or incident light. Most of the work in biology is with transmitted light; the opposite is true for materials such as metals and stone. Either transmitted- or incident-light methods may be adapted to give a reversed contrast effect (the object appearing bright against a dark background).

Light source. In all modern microscopes, the light source is an electric lamp of some form. The condenser of the microscope is used to focus an image of the surface of the lamp into the plane of the object. As the light source is in focus at the same time as the object detail, this method of illumination is called source focus; the term critical illumination has also been used. It is essential that the light source be fairly large and homogeneous to obtain an evenly illuminated field of view. A valuable refinement involves the addition of an iris diaphragm in front of the lamp to provide some control over the area of the objective that is illuminated. To minimize glare, this area should not be much in excess of the visual field. In most research microscopes, the illumination is arranged for the instrument in such a way that an image of the light source is formed in the back focal plane of the microscope condenser by a corrected condenser on the lamp house. The lamp is provided with a variable field stop (similar to the aperture diaphragm on a camera), and the microscope condenser is focussed so that this stop is seen in the plane of the object. When this has been achieved, the evenly illuminated lamp condenser serves as the apparent light source; hence, an evenly illuminated field of view results from the use of a small, intense, but nonhomogeneous light source. This method of illumination is called Köhler illumination, after the name of its inventor. For optimum results, the image of the lamp filament should be large enough to occupy a large portion of the area of the aperture of the back focal plane of the condenser. In practice, this means that the filament should be composed of several parallel coils—the so-called solid-source filament—and that the lens of the lamp-house condenser should be of a suitable focal length. The built-in systems of illumination that are now common comprise a small low-voltage bulb, collector lens train, and field stop.

Source focus

Illumination of an opaque specimen from above the stage is often accomplished by some form of vertical illuminator. In its simplest form, this is a thin glass slip inserted into the objective lens holder at 45° to the optical axis. Light is admitted through a small aperture in the side of the lens mount and is reflected down into the object through the lens, which thus acts as its own condenser. The light reflected from the object then passes back through the objective lens and glass slip to form the image. Many of the incident illuminators now available are designed for dark-ground work. In such cases, the incident light is reflected downward by an annular (ring-shaped) mirror around the objective and is then concentrated onto the object by a ring condenser mounted in the housing. An image of the specimen is formed by the objective lens, which is inserted into the centre of the illuminating housing. The reflected light enters the microscope tube by passing through the central hole in the mirror. Such dark-ground systems often give greatly enhanced contrast, as compared with standard vertical illumination.

Resolving power. Because of diffraction effects, the image of a point source of light can never be perfect, even assuming that all the imperfections in the lens are corrected. The image of a luminous point thus appears as a bright circular disk surrounded by a series of concentric light and dark rings. This pattern is the Airy disk, after Sir George Biddell Airy (1801–92), an English astronomer who first studied it mathematically.

Any image may be considered as being composed of innumerable overlapping diffraction patterns, each corresponding to a given point in the object. Two point sources that are close together will produce Airy disks that appear side by side. If the separation between the objects is reduced, then the diffraction patterns begin to overlap; if the overlap is considerable, then the two points are not seen as separate; *i.e.*, they are not resolved.

The best resolution should be realized when the full aperture of the objective lens is used and the wavelength of the illuminating light is decreased. This has been shown to hold in practice up to a point, beyond which diminution of contrast causes an apparent loss of resolution, in that the visibility of object detail is impaired. This loss of contrast is, however, distinct from the resolution proper. Modern expressions of the theory of microscopic resolution take into account the degree of coherence of the illuminating light.

If an optical instrument is resolving two closely situated points in the object, this fact will not be apparent to the observer unless the magnification of the instrument is such that the final image points subtend an angle of at least one minute of arc at the observer's eye. There is thus a minimum magnification below which the detail will not be perceived. In practice, this is usually taken as a value 500 times the numerical aperture of the lens. Conversely, excessive magnification is useless, since no more information will be presented to the eye, and the diffuse nature of the separate Airy disks making up the image will become apparent, giving an impression of poor image quality. A magnification, therefore, of 1,000 times the numerical aperture is regarded as the upper limit, values in excess of which will constitute so-called empty magnification.

SPECIALIZED TYPES OF MICROSCOPE

Stereoscopic binocular microscopes. In contrast to the binocular image of the ordinary microscope, these produce a truly stereoscopic image; they are invaluable for the examination at relatively low magnifications of many objects, for the assembly of very small electronic circuits and solid-state devices, and for dissecting and detailed surgical operations. An early form of this type of instrument was the Greenough binocular, which consisted of two complete microscope systems with their axes inclined at an angle of about 16°, so that they converged in the plane of the object. More stereoscopic binocular microscopes are provided with prisms to give an erect image for more convenient examination and manipulation of the object. Though the objective lenses may be duplicated, the tendency today is to use a single large objective lens with a device called an aperture-splitting system arranged above it that presents the images to the two eyepieces.

Aperture-splitting system

Ultramicroscope. In 1903, the ultramicroscope was developed as part of a project to study colloidal particles. These tiny particles, suspended in a liquid and too small to be seen in an ordinary microscope, are illuminated from the side with a powerful beam of light at right angles to the microscope's optical axis. Because the particles scatter light, their positions and movements may be seen as flashes against a dark background. Particles as small as five to ten millimicrons (about one-tenth the wavelength of visible light) can be observed.

Metallurgical microscope. This instrument is designed for studying the structure of opaque materials, particularly metals. It may be inverted so that a flat surface of the polished and etched specimen can be placed face down on the stage. It has a vertical illuminator and is usually equipped for photomicrography.

Phase-contrast microscope. Contrast in the image rendered by a microscope is usually due to variations in the light-absorbing ability (optical density) of the specimen. Staining the specimen enhances this effect for certain wavelengths. Living cells and other transparent specimens, however, are very difficult to study, and inducing contrast by staining is frequently impossible. The phase-contrast microscope provides an invaluable tool for the examination of such materials, since contrast is introduced by optical means alone and without any manipulation of the specimen. It does, however, suffer from some

optical defects, the most notable being that accurate measurements of the image are almost impossible to make.

Interference microscope. With the interference microscope, contrast is introduced into transparent objects just as in the phase-contrast instrument but without optical defects inherent in the phase-contrast system; in addition, measurements are relatively easy to make with this type of microscope.

Ultraviolet and fluorescence microscope. Invention of the electron microscope was preceded by continuous effort to increase the resolution of the microscope. One approach was to use shorter and shorter wavelengths, moving into the ultraviolet region of the spectrum. The first ultraviolet microscopes, using ultraviolet light to illuminate the specimen and special quartz lenses, showed a twofold increase in maximum resolving power, as compared with an instrument of the same numerical aperture using green light. Actual operation was tedious, however, largely because of difficulties in focussing the image and the necessity for recording all of the results by means of photography. Recently, the development of television cameras sensitive into the ultraviolet region of the spectrum has much simplified the recording and display of the image produced by the ultraviolet microscope.

With the development of the high-pressure mercury arc burner, a source rich in shortwave radiation became easily available. Similarly, the development of glass filters that transmit light between accurately known limits of wavelength in the ultraviolet portion of the spectrum have combined to make practical use of ultraviolet light for excitation of fluorescence (conversion of ultraviolet light to visible light) under the microscope. The light is focussed onto the specimen, using either a normal or a dark-ground condenser. Certain compounds absorb this energy and re-emit it as fluorescence of a longer wavelength in the visible region of the spectrum. This primary fluorescence is often of characteristic colour and may be of value for the identification of chemical compounds. The objective lens and eyepieces of a fluorescence microscope are similar to those normally used, except that it is desirable to choose lenses in which the cement between components does not itself show appreciable fluorescence. In order that any ultraviolet light that has not been absorbed by the specimen should not reach and damage the eye of the observer, a barrier filter is usually fitted between the objective lens and the eyepiece. The image is formed by the light emitted from the specimen itself, which thus appears luminous against a dark background. Ultraviolet microscopes find increasing use today in biology and medicine.

Fluorescence

Reflecting microscope systems. Interest in ultraviolet microscopy stimulated the redevelopment of reflecting systems, which had been tried at various times since the 18th century. The use of mirrors rather than lenses for ultraviolet is advantageous because of the complete freedom of mirrors from chromatic aberration and because, once focussed in visible light, the same focus applies to ultraviolet and infrared wavelengths. Coupled with these advantages is the fact that the working distance of a mirror system is much longer than that of a conventional lens system of comparable numerical aperture. Drawbacks to the use of mirror systems are that at least one of the reflecting surfaces must be aspherical (*i.e.*, with a nonspherical surface, which makes them difficult to manufacture) and that the front surface mirrors of such a mirror system do not have a very long life because of atmospheric attack and tarnishing. The reflectors used in this instrument consist of a large, aspherical, primary, concave mirror and a smaller, secondary, spherical convex mirror arranged axially with it. Only a few microscopes of this type have actually been constructed.

Reflecting systems have primarily been employed as additions to conventional optics. One ingenious system is added to a lens in order to obtain a long working distance. The mirrors are arranged to project an image into the front focal plane of a lens. Such devices are valuable for examining objects heated in a small vacuum furnace on the stage of the microscope or for carrying out the dissection of living organisms under the microscope.

Infrared microscope. By the addition to the microscope of an image converter (a device which converts an image formed by invisible light, such as infrared or ultraviolet, into a visible image), the longer wavelengths that constitute the infrared region of the electromagnetic spectrum may be made visible. This device attached to a microscope has been used to investigate the infrared transmission or reflectivity of objects and to render visible details in objects which may be invisible in visible light. Though applications are numerous, one particularly valuable field is microscopic investigation of forgeries. Another expanding use is detection of surface defects in silicon crystals used for transistors and other semiconductor devices.

Polarizing microscope. The polarizing microscope is used by geologists and crystallographers as well as by biologists. Ordinary light consists of waves vibrating in all directions. Plane-polarized light waves vibrate in only one plane. A polarizing microscope uses plane-polarized light to illuminate the specimen, which appears bright against a dark ground. In the course of passing through the microscope, the polarized light splits into two beams with different velocities. Their relative difference when they leave the specimen is called the retardation. When measured in millimicrons and divided by the thickness of the object measured in the same units, it gives a quantity termed the birefringence, a numerical value of help in the identification of minerals under the microscope.

Birefringence

Television microscope. The application of a television camera to a microscope is useful for recording and displaying the image to large audiences and for teaching purposes. More recently the scanning inherent in the formation of a television image has been used in measuring portions of the image and in automatic processing of the data obtained. An example of this use is found in the Quantimet image-analyzing computer, in which a microscope image is projected into the photocathode of the tube of a television camera; the output from the line scan passes after amplification not only to a monitor screen for visual inspection but also to a detector that responds to changes in the output voltage as the scanning spot passes over features of varying optical density. Specialized computer circuitry then enables the measurement of the area of specific features and permits the number of items in the field to be counted. Various other parameters may also be measured. Not only does such automation save operator time, but it also increases the degree of accuracy of the measurements.

Laser microprobe. A laser is a device for producing a very intense beam of light that can be sharply focussed. In the laser microprobe, a tiny beam of light from a pulsed laser (which produces laser light in pulses rather than continuously) is focussed onto the specimen from a vertical illuminator system that uses the objective lens of the microscope to focus the beam and allow observation of the area to be analyzed. The intense heat generated by the pulse of laser light vaporizes part of the sample. In one commercial form of this instrument, this vapour short-circuits an auxiliary gap between two electrodes positioned just above the specimen, and it is thus raised by the ensuing electrical spark to a temperature at which its spectral (light) emission may be measured by means of a device for analyzing light (spectrograph). This instrument has been used to analyze traces of metals which occur in living tissues with a minimal amount of damage to the tissue and has applications in the metallurgical, ceramic, and geological fields.

Holographic techniques. The coherence and intensity of laser light can be used to record the microscope image as a hologram, or three-dimensional representation in the form of an interference pattern. The hologram freezes the specimen at one instant in time, and it may then be reconstructed and viewed at leisure.

X-ray microscopy. Theoretically, the use of shortwave X-rays should provide a great increase in the resolving power of the microscope; the practical difficulties, however, are great, and predicted improvements in resolution have not been achieved. X-rays are not refracted by solid glass lenses, so that conventional methods will

not serve for the formation of an X-ray image. A few experimental X-ray mirror microscopes have been built; aberrations are high, however, and, therefore, resolution suffers.

In practice, X-ray microscopy is used not to achieve high resolutions but to analyze the X-ray absorption characteristics of the specimen. For this purpose, contact microradiography may be used. In this process, the object is placed in close contact with a fine-grained photographic emulsion and exposed to a beam of X-rays. After the plate has been developed, the X-ray image may be examined with a normal optical microscope.

Point-projection X-ray microscopy

An alternative technique is point-projection X-ray microscopy, in which an electron gun and condenser system mounted in an evacuated column direct an intense beam of electrons onto a metal cathode, which forms the end window of the tube. A beam of X-rays thus generated emerges and radiates from the foil. If a specimen is placed close to the target foil, its shadow will be cast onto a fluorescent screen or photographic plate placed some distance away. With a focal camera with a length of 10 centimetres and with the object about 0.1 millimetre from the X-ray spot, an initial magnification of 1,000 magnitudes will be achieved. Such an instrument suffers in that exposure times are long because there is difficulty in getting sufficient current into a very small electron spot, and the X-ray output is thus very low.

The X-ray projection microscope has the advantage that its depth of field is very large; stereoscopic photographs may be taken to show the organization of the structures in thick specimens. The image contrast is largely due to differences in the absorption of X-rays, and it is often possible to identify substances present by measuring the X-ray absorption at different wavelengths.

THE ELECTRON MICROSCOPE

Fundamental research by many physicists in the first quarter of this century suggested that cathode rays (*i.e.*, electrons) might be used in some way to increase microscope resolution. Louis de Broglie, a French physicist (1892–), in 1924 opened the way with the suggestion that electron beams might be regarded as a form of wave motion. De Broglie derived the formula for their wavelength (see ELECTRON), which showed, for example, that for electrons accelerated by 60,000 volts, the effective wavelength would be 0.05 angstroms; *i.e.*, one hundred-thousandth that of green light. If such waves could be used in a microscope, then a valuable increase in resolution would result. In 1926 it was demonstrated that magnetic or electrostatic fields could act as lenses for electrons or other charged particles, and this discovery initiated the study of electron optics. In 1931 a two-lens electron microscope was devised. It produced enlarged images of the electron source. Later, in 1933, a true electron microscope was built. A year or so later, the resolution of the optical microscope was surpassed, but the specimen was severely damaged by the heating effects of the electron beam, a problem that was not solved for some years. In 1935 the construction of the first commercially built electron microscope was begun in England, and it was followed soon by production microscopes in Germany and the U.S. Progress in the construction of electron microscopes was rapid, with a great impetus to the achievement of high resolution being the introduction, in 1946, of a device called a stigmator, which compensates for astigmatism of the objective lens.

Electron microscopes are now available that are easy to operate and reliable and that give point-to-point resolutions of less than three angstroms. Such high resolutions, coupled with suitable advances in the methods of specimen preparation, now permit the direct visualization of large molecules. Recently, the transmission electron microscope has been supplemented by the introduction of a scanning electron microscope, in which a beam of electrons is scanned over the surface of an object and used to build up an image of the details of the surface structure. Another development is the electron probe microanalyzer, which allows analysis of the composition of materials to be made by using the incident electron beam to excite the emission of characteristic X-radiation by the various elements composing the specimen. These X-rays are detected and analyzed by X-ray spectrometers built into the instrument. Such probe microanalyzers customarily have facilities to produce an electron scanning image so that structure and chemical composition may be easily correlated.

Differences of types. There are many differences between optical and electron microscopes. The electron microscope requires that the electron beam be in a vacuum, because electrons cannot travel any appreciable distance in air at atmospheric pressure. The column of the electron microscope is evacuated by pumps, and the specimen, photographic plates, and any other necessary apparatus are introduced into the vacuum by means of airlocks or seals. Unlike the optical microscope, in which the lenses are of fixed focus and the distance between specimen and objective lens is varied, the electron microscope has variable-focus lenses, and the distance between specimen and objective lens and separation of the lenses remain constant. The magnification is determined mainly by the value of the current (for magnetic lenses) through the intermediate and projector lens coils. The image is focussed by changing the current through the objective lens coil. A high degree of electrical stability is necessary, because the current in a magnetic lens controls its focal length and magnification. Another difference is that the optical microscope is usually operated (for visual observations, at least) so that the image is a virtual one, and in the electron microscope the final image is invariably real and is visualized on a fluorescent screen or recorded for study on a photographic plate.

In the optical microscope the image is formed by absorption of light in the specimen; in the electron microscope the image results from a scattering of electrons by atoms in the specimen. A heavy atom is more effective in scattering than one of low atomic number, and the presence of heavy atoms will increase the image contrast. The electron microscopist may incorporate more heavy atoms into his specimen for this purpose.

Lenses. Electron lenses may be either electrostatic or electromagnetic. The simplest form of electrostatic lens consists of two coaxial cylinders with voltage potential difference applied between them. Such a lens is not of practical use in the electron microscope. A more convenient arrangement is the unipotential or Einzel lens. In it, three coaxial cylinders or apertures in plates form the lens. The outer two are grounded, while the middle one is at a high negative potential.

Electromagnetic lenses, which are used for the condenser system and for the projector of the electron microscope, differ from the objective lens only in details. The manufacturing and performance tolerances of condenser-system and projector lenses are much tighter than they are for those used as objective and intermediate lenses.

Efforts to improve the resolution of the electron microscope have tended toward production of a single-field condenser-objective lens of low aberrations. In such a lens, the first part acts as a condenser and the second as the objective; the specimen is inserted into the centre of the lens, where the axial magnetic field (the field along the axis of the instrument) is at a maximum.

Electrostatic lens aberrations

Electrostatic lenses show spherical aberration, distortion, coma, astigmatism, and curvature of field. Electromagnetic lenses, because of the rotation of the electron beam during its passage through the lens, also suffer from other less important aberrations. In addition chromatic aberration due to variation in the wavelength of the electrons may be of considerable importance in both types of lens. Such changes of electron velocity may be due either to variations in the high-voltage supply to the electron gun or to energy losses from collisions of electrons with atoms in the specimen. The first of these may be controlled by careful stabilization of the high-voltage supply, but variations in the second cannot be reduced with the electron sources presently in use. For the very thin specimens and the electron energies commonly used, such variation due to energy loss in the specimen may be neglected. The resolving power of the microscope is se-

verely limited by the spherical aberration of the electron lenses. It is not possible to correct this aberration in the same manner as in the optical microscope, because magnetic electron lenses are always convergent lenses. Careful lens design may help, but at present the only practical answer to this problem is reduction of the numerical aperture of the electron lens.

Astigmatism in the electron microscope is largely due to asymmetries in the axial component of the magnetic field of the lens and is the result of imperfect construction of the lens or of the interaction of the electron beam with residual gas molecules in the column. This leads to deposits on the specimen that introduce asymmetries into the magnetic field when they charge up under the influence of the beam. Some astigmatic images may cause great errors in interpretation.

Transmission electron microscope. The transmission electron microscope has three essential systems: (1) an electron gun, which produces the electron beam, and the condenser system, which focusses the beam onto the object; (2) the image-producing system, comprising the objective lens, movable specimen stage, and intermediate and projector lenses, which focus the electrons passing through the specimen to form a real, highly magnified image; and (3) the image-recording system, which converts the electron image into some form perceptible to the human eye. It usually consists of a fluorescent screen for viewing and focussing the image and a camera for permanent records.

Additional requirements are a vacuum system, comprising at least two pumps and their associated valves, and electric-power supplies with essential stabilizers and control equipment.

The electron gun and condenser system. The source of electrons, the cathode, is a heated, V-shaped tungsten filament with its apex, from which the bulk of the emission occurs, towards the condenser lenses. The filament is surrounded by a shield, sometimes called a Wehnelt cylinder, with a central aperture arranged on the axis of the column; the apex of the filament is arranged to lie at or just above or below this aperture. The cathode and shield are at a high negative potential and are insulated from the rest of the instrument. The final electrode of the electron gun is the anode, which takes the form of a disk with an axial hole. Electrons leave the cathode and shield, accelerate toward the anode, and, if the stabilization of the high voltage is adequate, pass through the central aperture at a uniform speed. The control and design of the electron gun are critical in ensuring satisfactory operation.

The intensity and angular aperture of the beam are controlled by the condenser lens system between the gun and the specimen. A single lens may be used to converge the beam onto the object, but, more commonly, a double condenser is used. In this the first lens is strong and produces a reduced image of the source, which is then imaged by the second lens onto the object. Such an arrangement is economical of space between the electron gun and the object stage and is more flexible, because the reduction in size of the image of the source (and hence the final size of illuminated area on the specimen) may be varied widely by controlling the first lens. The use of a small spot size minimizes disturbances in the specimen due to heating and irradiation.

The image-producing system. The specimen, mounted on a copper grid, is carried in a small holder in a movable specimen stage. The operator can control it by a system of micrometer screws and bell cranks working through vacuum seals, so that a large area of the specimen can be examined systematically. The objective lens is usually of short focal length (one to five millimetres) and produces a real intermediate image with a magnification of 20–200 magnitudes. An area of this image is rendered in turn with further magnification by the projector lens or lenses. A single projector lens may provide a range of magnification of 5:1 and by the use of interchangeable pole pieces in the projector a wider range of magnifications may be obtained. Modern instruments employ two projector lenses (one called the intermediate lens) to permit a greater range of magnification and to provide a greater overall magnification without a commensurate increase in the physical length of the column of the microscope. Such a system may be operated so that each stage in turn provides magnification, or the intermediate lens may be used to reduce the image produced by the objective and the projector set to magnify part of it to give the final image.

For practical reasons of image stability and brightness, the microscope is often operated to give a final magnification of 1,000-50,000 magnitudes on the screen. If a higher final magnification is required, it may be obtained by photographic enlargement. The quality of the final image in the electron microscope depends largely upon the accuracy of the various mechanical and electrical adjustments with which the various lenses are aligned to one another and to the illuminating system. The lenses require power supplies of a high degree of stability, and stabilization is provided electronically for a satisfactory standard of resolution.

Image quality

The gun, specimen chamber, and column and camera are evacuated so that the electron beam can pass freely. At the place where the specimen and photographic plates are inserted, air locks provide that only a minimal amount of air is introduced into the column; this reduces pumping time before the microscope is again operational.

Image recording. The electron image must be made visible to the eye by allowing the electrons to fall on a fluorescent screen fitted at the base of the microscope column. Such a screen is satisfactory for quick observations and for focussing and aligning the instrument. A low-power optical microscope fitted outside the column allows the image on the screen to be inspected at a magnification of about ten magnitudes, but the image suffers from poor contrast, lack of brightness, and the inherent grain of the screen. By allowing the electrons to fall upon the emulsion of a photographic plate, the latent image can then be developed and fixed exactly as in ordinary photography, and prints at any degree of further enlargement may be obtained from the negative. Such further enlargement allows all the detail present in the image to be resolved by the eye and also allows some control over the contrast.

Image-intensification systems have been developed using some form of television tube, with the final image presented on a cathode-ray-tube monitor. Image intensifiers permit electronic amplification. Beam currents that would not register any image on a fluorescent screen may be used to permit examination of materials which are susceptible to radiation damage, such as certain types of plastics.

Limitations of the transmission electron microscope. The modern transmission electron microscope has tremendous resolving power, so that large molecules can be visualized. Several limitations may affect interpretation of the image: (1) The object must be placed inside the vacuum of the column; this means that living material cannot be examined, and the biological organization of a specimen may be altered when it is subjected to the preparation necessary for electron microscopy. (2) The penetrating power of 60-kilovolt electrons is severely limited, so that specimens for study in conventional transmission electron microscopes must not be much thicker than 0.1 micron. Elaborate techniques must be devised for slicing bulk specimens or for separating the constituents of a bulk phase. (3) The use of thin sections, in turn, imposes a two-dimensional limit on the micrograph. Stereoscopic techniques that are technically possible by tilting the specimen between two exposures are, therefore, only of limited value. The structure in three dimensions may be difficult to ascertain, and recourse often has to be made to elaborate serial sectioning and reconstruction to obtain this information. (4) The fact that the electron image is formed by the heavy atoms in the specimen scattering the electrons may be a limitation in biological work. The materials of which biological specimens are composed are intrinsically of low scattering ability, and, therefore, heavy metals or their salts must be used in the preparation of such

material in order to introduce the necessary contrast into the image. These reagents may make it difficult to obtain any information about the chemical composition of the specimen.

Current researches are surmounting some of these drawbacks: in particular, study of cell metabolism with the electron microscope is now possible, and chemical information can now be obtained at the ultrastructural level.

Special types of electron microscope. *High-voltage instruments.* Several different types of instrument have been constructed to operate at accelerating voltages in excess of the 100 kilovolts normally used in the conventional transmission instrument. Such high-voltage microscopes are in commercial production for accelerating voltages of up to 1,200,000 volts.

These instruments possess several advantages over the conventional microscope: (1) The theoretical resolving power of the instrument increases as the accelerating voltage increases. (2) Thick specimens are easily penetrated by the beam. (3) Chromatic aberration caused by energy losses and scattering in the specimen is decreased. (4) Specimen heating and irradiation damage are decreased. (5) The resolution of electron diffraction patterns is improved.

Although the gain in resolution is theoretically about two and a half times that of a 100-kilovolt instrument, such increases have not as yet been achieved in practice. It is primarily for the increased penetrating power and consequent reduction in chromatic aberration that the high voltage instruments are used. Technical difficulties still abound in this field, largely because of problems of the adequate stabilization of the very high voltages and of lack of contrast in the material when examined with such high-energy electrons.

X-ray dangers

The operator must have adequate protection from the X-rays that are produced when the electron beam impinges on objects in its path. Progress has been made toward solutions to these problems, and the high-voltage microscope is proving invaluable in the study of metals and alloys. It is possible that biological materials may also be examined if they are enclosed in microchambers inserted into the path of the beam. Bacteria have been examined in this way, but many problems remain with this technique.

The scanning electron microscope. The scanning electron microscope, designed for studying the surfaces of objects directly, utilizes a beam of focussed electrons as an electron probe that is scanned in a regular manner over the specimen. The action of the electron beam stimulates emission of secondary electrons from the surface of the specimen. These are collected and passed to a scintillator crystal (a crystal that produces tiny flashes of light when bombarded with electrons). The light flashes in the crystal operate a photomultiplier tube, which amplifies the electrical signal. The resultant amplified current is used, in turn, to modulate the brightness of a display cathode-ray (television) tube that has its line scan driven in synchronism with the probe beam in the microscope column. There is a point-by-point correspondence between the brightness of each point on the display tube and the number of secondary electrons emitted from any point on the surface of the specimen. In this way, a picture of the surface topography of the specimen is built up. The magnification of the image may be controlled by changing the ratio between the relative sizes of the area covered by the scanning probe and the picture on the display tube.

No elaborate specimen-preparation techniques are required for examination in the scanning electron microscope. Large and bulky specimens may be accommodated. It is desirable that the specimen be rendered electrically conducting, otherwise a sharp picture will not be obtained. Conductivity is usually achieved by evaporating a 10–20-nanometre-thick layer of metal onto the specimen in a vacuum (such a thickness does not materially affect the resolution of the surface details); or, under certain circumstances, an aerosol antistatic spray will achieve the same purpose.

The scanning electron microscope has the valuable feature that the image contrast may be changed by differences in potential across the surface under study. This facility is used widely in the semiconductor field and in studying microcircuit devices while they are operating. Recently, scanning instruments have been used for transmission electron microscopy, with the advantage over the conventional transmission instrument that very low magnification (of about 20 magnitudes) may be used. At the same time, large fields of view are obtained, and very thick sections may be studied; this latter allows easy three-dimensional reconstructions to be made.

The electron-probe microanalyzer. A limitation of the conventional electron microscope is that it provides no chemical information. Electron-probe microanalyzers have been developed since 1947 to carry out nondestructive elemental analysis at resolutions approaching those of the transmission electron microscope. This is done by measuring the intensity of the characteristic X-rays emitted by a specimen when a focussed electron beam impinges on it. Early instruments used a static electron probe, but now scanning is the usual mode of operation, and a scanning X-ray image is built up on a cathode-ray tube. Such an instrument is called an X-ray microprobe analyzer.

X-ray probe microanalyzer

The X-ray microprobe analyzer is so valuable that X-ray spectrometer attachments are now provided for existing transmission electron microscopes. Combined instruments are specially designed with both transmission-electron-microscopy and X-ray microprobe analysis facilities.

The electron-probe microanalyzer has applications in mineralogical and metallurgical fields and is also used in biological work.

Field-emission microscope. In this instrument a fine wire with a rounded tip is mounted in a cathode-ray tube. Electrons are drawn from the tip by a high electrical field and travel toward the screen on which the image is formed. Only heat-resistant metals, such as tungsten, platinum, and molybdenum, can be examined in this way, because the high current density at the metal tip has a great heating effect. The magnification is proportional to the ratio of the radius of curvature of the screen on which the image forms to the radius of the metal tip.

A variant of the field-emission microscope, called the field-ion microscope, is used in metallurgy for direct observation of defects in the crystalline lattice and for assessment of the effects of radiation damage and chemical reactions on the metal surface.

The proton-scattering microscope. This instrument uses a beam of protons from an ion source or gun. The beam is focussed and accelerated onto a crystal of a metal such as copper or tungsten, which serves as the target. The pattern of scattered protons is recorded on a fluorescent screen; the blocking patterns provide information on the orientation of the crystals.

PREPARATION OF SPECIMENS

Optical microscopy. The surface of bulk materials can be studied at relatively low powers with incident light by simply mounting the specimen for ease of handling. Often, more information may be gained if the surface is flattened by grinding and polishing. In the case of metals, it may be necessary to etch the specimen in acids to allow the grain boundaries to be easily differentiated. With smaller amounts of material, such as powders and dusts, the sample is usually mounted on a slide in a resin medium of known refractive index and covered with a coverslip. For examination by incident light, the specimen material may be simply strewn on a slide.

Biological material, both plant and animal, generally requires more preparation. Organisms must be reduced to pieces which are small enough for study under the microscope. If unicellular organisms or small invertebrates are to be studied alive, they are mounted in a suitable fluid medium. Permanent slide preparations may be made of material that is first killed and then treated with chemical reagents that react with the protoplasm and prevent any further change in it. Contrast is obtained by soaking the specimens in dye solutions. Finally, the water in the objects may be removed by treatment with a series of alco-

hols of gradually increasing strength before the preparation is cleared in an organic fluid and mounted in a synthetic resin. This sets hard, fixing the coverslip firmly onto the preparation. When large masses of tissue are studied, the individual cells are separated from one another by the process of maceration (softening by soaking), or, more commonly, the object is embedded in a supporting medium, such as paraffin wax, which is then cut into very thin slices. The resulting sections are attached to glass slides, the embedding material dissolved away, and the section stained in various ways to obtain the required contrast. Dehydration, clearing, and mounting in resin complete the preparation of the slide. In special cases, the tissue is made rigid enough for sectioning by freezing it solid.

Maceration

Rocks and other hard specimens may be cut into thin sections for microscopical examination by using a diamond saw or impregnated wheel. The section is reduced to its final thickness by grinding it down with an abrasive powder.

Electron microscopy. For transmission electron microscopy, the specimen must be in very thin slices, so that the electron beam may penetrate it. Biological material is fixed and dehydrated essentially as for optical microscopy but is embedded in methacrylate or epoxy resin, which sets much harder than the wax and nitrocellulose used in optical microtechnique and allows thinner sections to be cut. The machines used for slicing (ultramicrotomes) usually operate with a knife broken from plate glass and use thermal expansion of the support to move the block forward in the small increments required for each successive section. The thin sections, as they come off the knife, are floated onto the surface of a liquid (usually distilled water) and then picked up on a copper grid; after drying, this grid, bearing the section, may be inserted directly into the microscope.

Contrast in material of biological origin is usually introduced at or just after the fixation stage by the use of osmium tetroxide, which is believed to change chemically after its introduction into the material to a substance (perhaps a lower oxide of osmium or the metal itself) that then provides the required electron density. Further increase in the contrast may be obtained by so-called electron staining of the finished sections when they are on the grids. This is done by floating the grids onto solutions of lead or uranyl salts for short periods of time; after washing and drying, the sections may be examined in the microscope.

Electron staining

For the examination of surfaces in the transmission microscope, it is necessary to take a replica of the surface in plastic, coat the replica with a film of evaporated carbon for strength, and finally evaporate a thin layer of metal onto the surface of the replica at a low angle in order to introduce contrast. Particulate material may be supported on a thin film of plastic carried on a copper grid. Contrast may be added to the biological material by shadowing it in a vacuum with heavy metals at a low angle. This shows up the surface detail just as the setting sun shining along a road shows up the irregularities in the road surface in clear relief. Alternatively, a negative staining technique may be used. In this, the specimens are surrounded by a matrix of the salt of some metal that is opaque to the electron beam. When examined in the microscope, a negative image results with the shape of the particle outlined in sharp relief. This technique is used in studying viruses and in molecular biology.

Metals may be prepared for the electron microscope by obtaining them as foils and then thinning them down further electrolytically to the point at which they will transmit the electron beam. Plastics and rubber may now be sectioned for examination in the electron microscope by freezing and then slicing.

Specimens for study in the scanning electron microscope simply need to be mounted on a specimen-stub holder and rendered electrically conductive either by an antistatic spray or by evaporating onto them a very thin layer of metal.

BIBLIOGRAPHY

History: S. BRADBURY, *Evolution of the Microscope* (1967), a general account of the history of the microscope from early days to the present.

Optical microscopy: B.O. PAYNE, *Miscroscope Design and Construction*, 2nd ed. (1957), a survey of the basic principles of the mechanical design of the instrument; L.C. MARTIN, *The Theory of the Microscope* (1966), an advanced mathematical treatment of the physical theory of the microscope; A.L.E. BARRON, *Using the Microscope*, 3rd ed. rev. (1965), a discussion of various types of instruments and their application; G.H. NEEDHAM, *Practical Use of the Microscope* (1965), a comprehensive treatment of the practical aspects of the use of the microscope and its accessories; W.H. BURRELLS, *Industrial Microscopy in Practice* (1961), a discussion of the industrial applications of various microscopes; G.W. WHITE, *Introduction to Microscopy* (1966), a general treatment of the instrument and its uses; V.E. COSSLETT, *Modern Microscopy* (1966), a survey of the various types of microscopes and their limitations.

Special types of microscopes: K.F.A. ROSS, *Phase Contrast and Interference Microscopy for Cell Biologists* (1967), a simple treatment of phase and interference microscopes and their biological applications; M. FRANCON, *Progress in Microscopy* (1961), an advanced treatment of phase contrast and interference microscopy; W. KRUG, J. REINITZ, and G. SCHULZ, *Contributions to Interference Microscopy* (Eng. trans. 1964), a technical treatise on the interference microscope in all its forms; A.F. HALLIMOND, *The Manual of the Polarizing Microscope*, 3rd ed. with appendixes (1970), a treatment of the microscope emphasizing its application in mineralogy and crystallography; A.H. BENNETT *et al.*, *Phase Microscopy* (1951), a detailed monograph largely devoted to the theory of phase contrast microscopy; P. GAY, *An Introduction to Crystal Optics* (1967), a work on crystallography dealing with the application of the polarizing microscope in the identification and characterization of minerals; S. TOLANSKY, *Interference Microscopy for the Biologist* (1968), a discussion of the applications of interferometry in studies of surface contours of biological materials.

X-Ray microscope: V.E. COSSLETT and W.C. NIXON, *X-ray Microscopy* (1960), a detailed treatise on the X-ray microscope and its potential.

Electron microscope: P. GRIVET *et al.*, *Electron Optics*, rev. ed. (Eng. trans. 1965), a detailed treatment of the theoretical background of electron optical devices; M.E. HAINE and V.E. COSSLETT, *The Electron Microscope* (1961), a comprehensive discussion of the design of electron microscopes; C.E. HALL, *Introduction to Electron Microscopy*, 2nd ed. (1966), a general text on the theory and applications of electron microscopy; E. RUSKA, in *Advances in Optical and Electron Microscopy*, vol. 1 (1966), a survey of possible approaches towards higher resolution in the electron microscope; P.R. THORNTON, *Scanning Electron Microscopy* (1968), a thorough treatment of the theory of the scanning electron microscope and its applications in material science; D.G. BRANDON, in *Advances in Optical and Electron Microscopy*, vol. 2 (1968), a general treatment of the theory and applications of the field ion microscope; E.W. MULLER and T.T. TSONG, *Field Ion Microscopy* (1969), a work on the theory, practice, and applications of field-ion microscopes; T. MULVEY, "The History of the Electron Microscope," in S. BRADBURY and G.L.E. TURNER (eds.), *Historical Aspects of Microscopy* (1967); G.A. MEEK, *Practical Microscopy for Biologists* (1970), a good practical account of the instrument and its operation.

Techniques: C.E. MCCLUNG (ed.), *Handbook of Microscopial Technique*, 3rd ed. (1950), a survey of optical microtechniques; C.P. SHILLABER, *Photomicrography* (1944), a general discussion of basic microscope theory, with detailed treatment of photographic methods for various types of specimens; D.F. LAWSON, *The Technique of Photomicrography* (1960), a work on the practical aspects of this subject; D.H. KAY (ed.), *Techniques for Electron Microscopy*, 2nd ed. (1965), a detailed survey of techniques for electron microscope use; D.C. PEASE, *Histological Techniques for Electron Microscopy* (1960), a discussion of the basic techniques in the preparation of biological materials for electron microscopes.

(S.Br.)

Middle Ages (c. 395–1500)

The term middle age (*medium aevum*) was first used in the later 15th century by Humanist scholars as a description of that period of western European history between the collapse of Roman civilization in the 5th century AD and the revival of civilized life and learning in which the Humanists believed themselves to be participating. The

purpose of this article, however, is to describe some more positive achievements of those centuries, which saw the emergence of Europe as a cultural unit and the rise and decay of a distinctive civilization within it.

Foundations of medieval civilization

The materials from which this civilization was molded were essentially threefold: the inheritance of classical antiquity, the Christian tradition, and Germanic and Scandinavian social patterns. Classical antiquity, which set the standards of learning, culture, and government by which medieval no less than Renaissance scholars measured their own achievements, passed into Europe by several routes. Over part of Europe, most notably Italy, Spain, and southern France, the Germanic invaders entered a society in which Roman social and political organization, urban life, and even local government continued, much enfeebled but never totally interrupted. In northern Gaul, always more thinly Romanized, this was much less true; in Britain little but the roads and crumbling walls survived as witness to the secular presence of the empire. If, however, Germanic tribes overran the empire, then the Roman and Catholic church rapidly overran the tribes. Most of the invaders were either pagans or followers of the heresiarch Arius, who had denied the true divinity of Christ. This resulted in some persecution of the orthodox in Spain, Italy, and North Africa; but the Catholic baptism of the Frankish king Clovis, soon after 493, was the first of a series of conversions by Germanic rulers that led to the triumph of orthodoxy. Thenceforward, the church was able to play an essential role in preserving literacy and even some classical learning in its liturgy and literature, in maintaining some of the forms of public administration in its diocesan government, in perpetuating the tradition of corporate responsibility for peace and the relief of want, and perhaps most of all in creating a new universal society to replace that once provided by the fallen empire. It was ultimately the Latin Church rather than the Roman imperial tradition that determined the frontiers of modern Europe.

Classical antiquity, however, was also mediated through two other sources. While the empire in the West slowly ebbed away, that of the East, well protected from the north by the Bosporus, survived in Constantinople (or Byzantium). Although sea communications with the West were frequently interrupted and Byzantine culture became progressively more Greek than Latin, the influence of a continuing Roman Empire was great. Byzantine architecture and painting, the forms of imperial government, and the incomparable wealth and size of the greatest of Christian cities continued to exercise a powerful attraction. Ecclesiastical and political differences, culminating in the sack of Constantinople by crusaders (1204) and the growing self-confidence of Western rulers and scholars, much reduced the empire's influence by the 12th century. By then it was not from the Christian Byzantines but from the Muslim Arabs that the West was recovering much of the Greek tradition in mathematics, metaphysics, and the natural sciences, though in interpolated and corrupted translations.

The Germanic influence upon medieval society derived from a long series of migrations, invasions, and settlements, which began in the 3rd century AD and continued into the 12th. Their impact varied greatly, according to the conditions of their entry into the territory of the former empire, but they introduced a quite different pattern of social organization and lordship, one that took on some of the forms of Roman and Christian tradition while wholly altering their emphasis and function.

This article is divided into the following sections:

I. Chronological outline of medieval history
II. Structures and components of society
  The rulers
  The aristocracy
  Other social groups
III. The operation of the medieval world
  Economic patterns
  Forms of lordship
  Church government
  Royal government
IV. The world of the senses and the mind

I. Chronological outline of medieval history

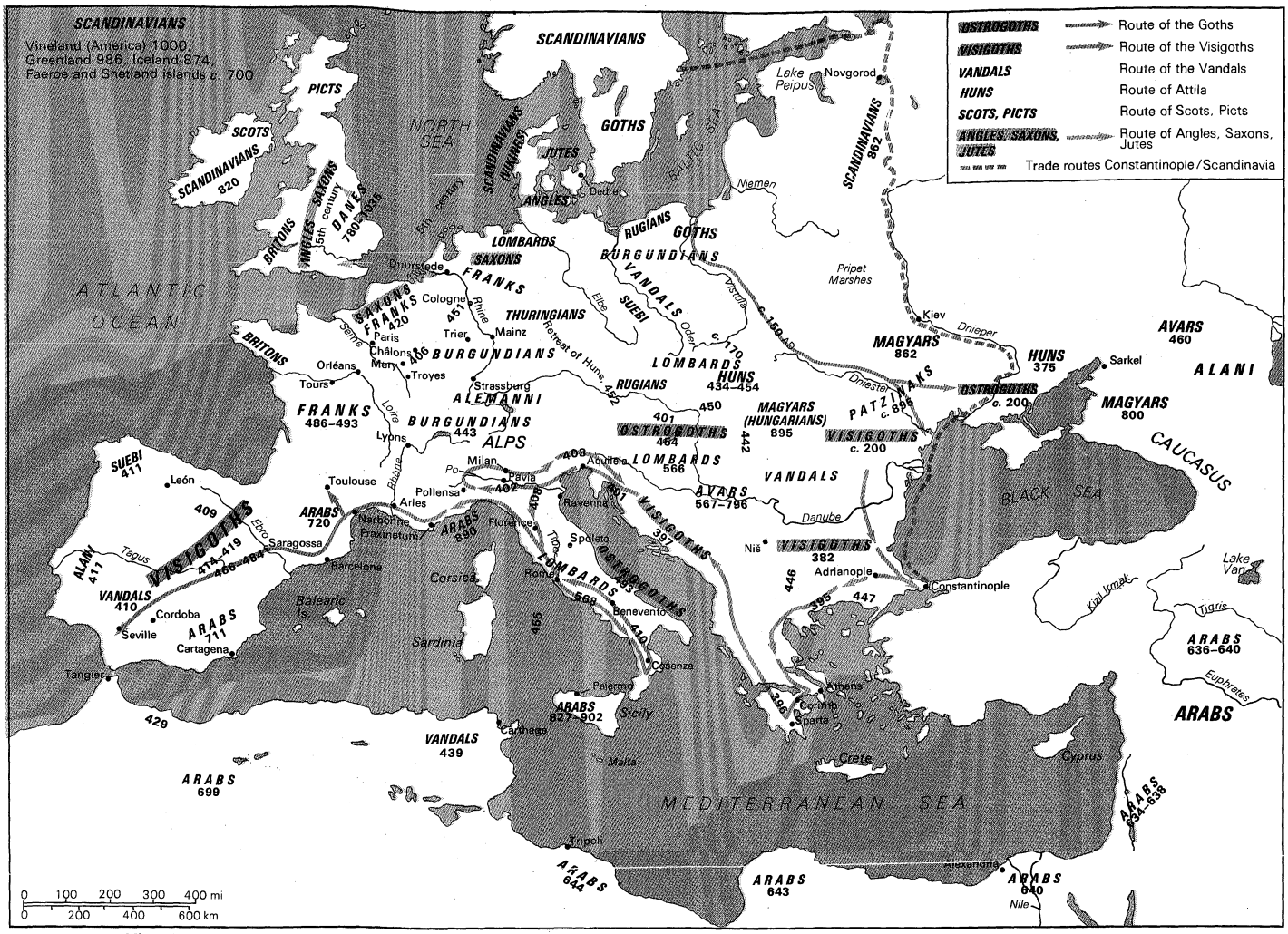
From the time of the emperor Hadrian (died AD 138), the northern frontiers of the empire had rested upon the Danube and the Rhine. In the course of the 3rd century these came under heavy pressure from Germanic tribes —Burgundians, Franks, and many others on the Rhine and Goths and Vandals on the Danube. The Goths, once settled on the plains north of the Black Sea, came, like their fellows on the Rhine, under pressure from the raiding nomadic horsemen of the Steppe—the Alans and Huns. The emperors, often desperately short of men and resources, had long been accustomed to taking some of these neighbours into their pay as allies to defend the frontier against their fellows, quartering them upon the local population and granting their leaders exalted posts in the imperial government. In the 5th century this frontier gave way along its whole length.

Collapse of the Roman frontier

In 410 the Visigoths under Alaric plundered Rome itself, while a Vandal army was ravaging southern Gaul before crossing into North Africa. By 455 the whole province west of Tripolis was in Vandal hands; soon thereafter, the Burgundians occupied the area around Lyons and the Visigoths established their kingdom over southern Gaul and, later, Spain, while Italy fell to an Ostrogothic army under the Romanized though Arian Theodoric after 487. In northern Gaul a Frankish force under Clovis appeared as the dominant power at the end of the century. Under the rule of Clovis' dynasty (the Merovingians), Frankish hegemony was asserted over the Burgundians, the Visigoths of Provence, and later the pagan Germans of Thuringia and Bavaria, where Irish missionaries such as St. Columban (died 615) helped to consolidate the conquest. A sprawling and much divided kingdom thus straddled the earlier frontier of the empire, with its centre in the Rhine Basin. The amalgam of North German peoples who conquered the southern and eastern parts of Britain between 450 and 600 found little and preserved less of the more shallow-rooted Roman society there, though they were later converted by Irish missionaries from the north and by a mission under the monk Augustine (later St. Augustine of Canterbury) sent from Rome in 597.

Only the Frankish and Anglo-Saxon kingdoms lasted long. Under the emperor Justinian (reigned 527–565), Byzantine armies destroyed the Vandal and Ostrogoth kingdoms, but at great cost and to little purpose. From the north the Lombards entered Italy in 568 and created a new kingdom, with its capital at Pavia, that extended southward to Benevento and Spoleto, leaving Byzantium with a precarious hold on the coasts, Rome itself, and the far south. Meanwhile, from Mecca the Muslim Arabs passed rapidly westward, capturing all North Africa and leaving only tiny pockets of Christian resistance in Spain, in Asturias and the Pyrenees. The northernmost point of their expansion was marked by the defeat of a Muslim army outside Tours in 732, though Sicily was to fall to them in the 9th century and there was an Arab stronghold in the Alps until 972. Frankish power was reasserted by Pepin the Short and his son Charles the Great (Charlemagne), under whom the frontier was pushed beyond the Elbe in the east and the Tiber in the south; again the missionaries, including the Anglo-Saxon St. Boniface, followed closely behind the army and consolidated its work. In 800 Charlemagne was made emperor and crowned by the Pope in Rome, founding the medieval empire in Europe out of a fragment of its Mediterranean parent.

The Scandinavians in the late 8th and 9th centuries expanded rapidly. They moved eastward through the Russian river systems to found the great trading centres at Novgorod and Kiev and even to besiege Byzantium. Scandinavian traders and pirates travelled southward to attack Ireland, England, Gaul, and the coast of Italy. Outside Russia they founded permanent settlements in Iceland, Greenland, Ireland and the Scottish islands, eastern England, and Normandy, and some reached North America. In the east the Magyars, a second wave of nomadic Asian horsemen, threatened Germany and Italy before settling eventually in Pannonia—the germ of the modern Hungary. Before these threats the descendants of



Migrations and conquests, 2nd to 10th century.
From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

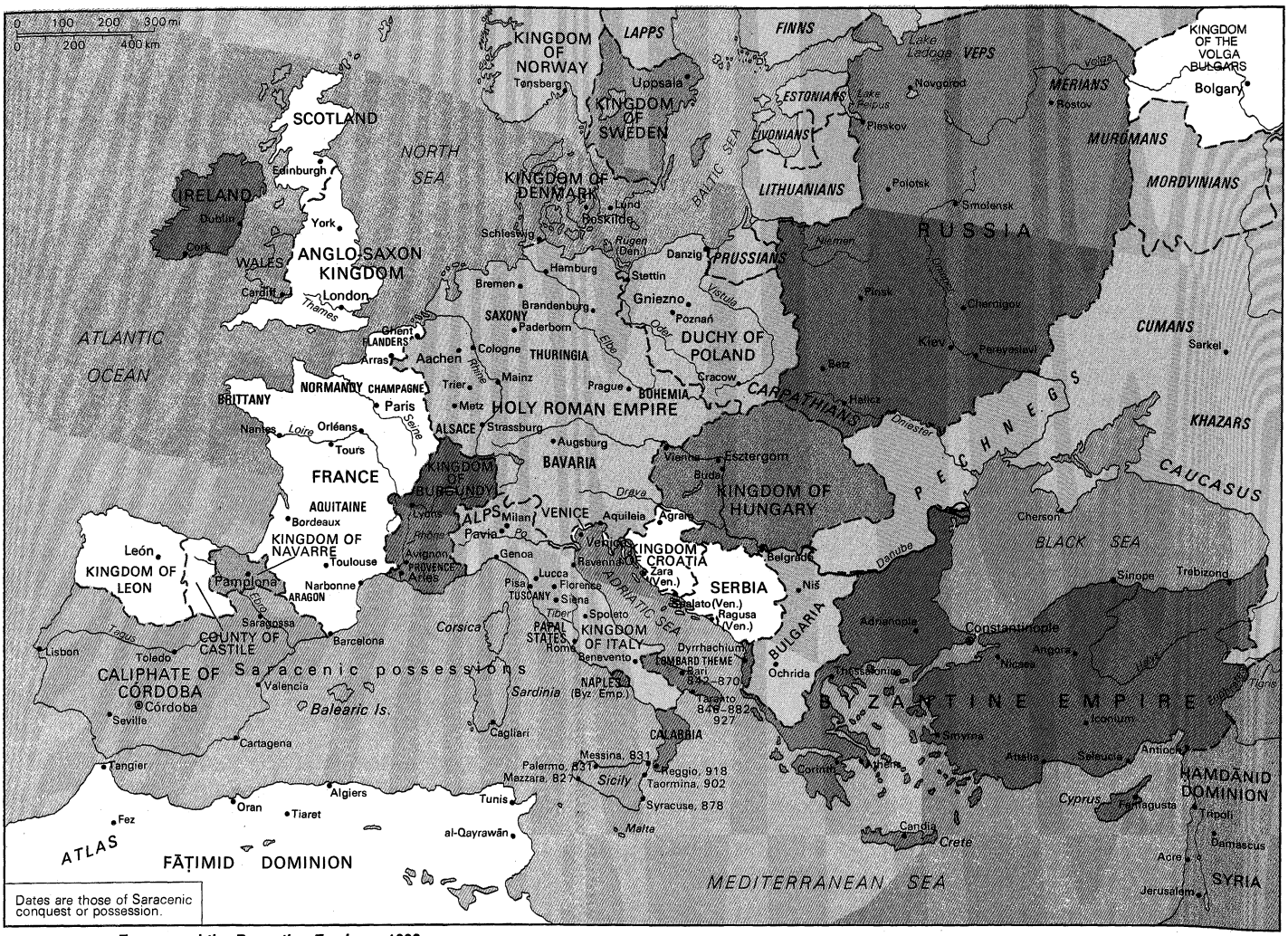
Charlemagne (the Carolingians) survived as kings of the Western Franks until replaced in 987 by their greatest subject, Hugh Capet, but effective political authority had long passed to the local princes. A new Saxon dynasty under Otto I defeated the Magyars and united northern Italy with Germany under the imperial crown in 962. In England, Danish invaders had destroyed all but the solitary kingdom of Wessex of King Alfred, under whose descendants the country was united for the first time.

New territorial and intellectual expansion

As these dangers receded, a new period of territorial and intellectual expansion began. German settlers pressed eastward across the Elbe into former Slavonic territory as settlers and missionaries. In Spain the kings of Leon and Aragon began the long process of reconquest, which was not completed until the fall of Muslim Granada in 1492. In Italy, Norman adventurers expelled the Byzantines from Calabria and the Muslims from Sicily. In Scandinavia, Poland, Bohemia, and Hungary the native princes still ruled, but now as Latin Christians. External missionary advances were matched by an internal revival of ecclesiastical life—first in the monasteries of Burgundy and Lorraine, later in the secular church. Beginning in the later 11th century, under Pope Gregory VII and his successors Rome took the lead in this movement and challenged the emperor and all secular rulers for pre-eminence within Christian society.

Widespread economic revival, reflected in an increasing volume of trade, accelerated internal as well as external colonization, and renewed urban growth, beginning in Italy, was accompanied by a fresh interest in law, both canon (ecclesiastical) and civil. Canon law provided the weapons in the armory of the church in its struggles against such rulers as Frederick I of Germany and Henry II, king of England, lord of Ireland and western France; it also encouraged the growth of local church government and a wide variety of institutions, all under the authority of the papacy and its evolving central bureaucracy. Civil and secular law provided materials for new theories of secular dominion and encouraged the development of more sophisticated forms of government in such states as Norman Sicily, Angevin England, and, later, Capetian France. All these developments supposed and encouraged a desire for more education for clergy and laity, a trend best illustrated by the rise of the first universities (12th century).

The course of the later 12th and the 13th centuries was marked politically by the division of Europe into two political types—that of Germany and Italy, where imperial authority broke up in favour of the Italian city states, the papacy, and the German princes, and that of France, Spain, England, and Scandinavia, where there was a steady advance in the progress of central authority, though on varying principles. Economically, it was a period of rising population, widespread land clearance, and the appearance of a powerful and wealthy merchant class, particularly in Flanders and even more in Italy. In 1231 Frederick II struck the first gold coins of commercial importance to be seen in western Christendom since the failure of the old imperial Roman coinage in the 8th century, and his example was soon followed in Venice and Florence. Intellectually, it was a period of elaborate syntheses of the traditional doctrine of the church with the new learning of Aristotle—first in logic and dialectic, later in natural science and metaphysics. The harmony of rea-



Europe and the Byzantine Empire c. 1000.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

Synthesis between church doctrine and Aristotelian learning

son and revelation celebrated in the theological systems of St. Thomas Aquinas (died 1274) and his fellows matched the legal coherence and widespread acceptance of a doctrine of papal supremacy dramatically expressed by Innocent III in the Fourth Lateran Council, in 1215.

The 14th and 15th centuries were beset by difficulties that overthrew most of the assumptions of earlier government and compelled the rise of new ones. Climatic change, famine, and a series of disastrous plagues, beginning with the Black Death of 1347–50, reduced the population in some areas by as much as half and dealt a grave blow to the traditional patterns of agriculture, which were already under strain. Destructive war and recurring pestilence provoked the first extensive peasant revolts—the Jacquerie of France (1358) and the English Peasants' Revolt (1381). In the 15th century similar rural grievances gave rise to revolts as far apart as Catalonia and Swedish Finland and formed an important element in the nationalist risings of Engelbrekt in Sweden and the Hussites of Tábor in Bohemia. Yet for the survivors of the plagues there were distinct advantages: woolgrowing in Spain and cloth working in England, cereal raising in eastern Europe, and many industries rose to new heights of prosperity in these years, which saw the fortunes of such imperial towns as Nürnberg at their zenith.

Politically, the fragmentation of Germany into innumerable petty sovereign states was consummated by the end of the 15th century, though on its eastern border Poland, united with Lithuania under Władysław II Jagiełło, and Hungary, under Matthias I Corvinus, rose to great prosperity. The stability of both France and England was threatened by their long struggle in the Hundred Years' War and the feuds of the greater magnates, among whom the dukes of Burgundy almost succeeded in carving out a new state along the frontiers of France and Germany. Under Charles VII and Louis XI, the Valois monarchy emerged as masters of all France, with a large measure of absolute power similar to that later achieved by the Tudor Henry VII in England. In Spain the fall of Granada and the union of the crowns of Aragon and Castile under Ferdinand and Isabella laid the foundations for Spanish hegemony in the next century.

The papacy, even when in exile in Avignon (1309–77), continued to centralize the government of the church on an ever more elaborate structure of administration, but it faced formidable difficulties because of the rise of nationalist and monarchical feeling, especially in France and England; because of the rise of an educated and often anti-clerical laity; and, above all, because of the Great Schism of 1378–1417, brought about by the election of two rival popes, one for Avignon and one for Rome. Though the 15th-century popes were able to overcome the challenge to their authority presented by a series of councils, they were ill-prepared either by this struggle or by the difficulties of Italian politics to give an effective lead to the demands for ecclesiastical reform increasingly audible by the end of the century.

Intellectually, the 14th century saw the synthesis formed by the 13th-century theologians begin to break up; the dominant school of Nominalist metaphysicians, asserting an unbridgeable gulf between the truths of reason and revelation, pursued epistemological enquiries of great daring but limited application, while at the other extreme the mystical way of Thomas à Kempis and many others

had few academic or institutional consequences. The dominant new current was to be the classical Humanism of the 15th century Italians, with its more restrained parallels north of the Alps. What was new in this movement was not admiration for the classical past but a determined effort to recover the secular and educational values of classical civilization and cast aside the prolonged medieval attempt to adapt pagan learning to a Christian purpose. Armed with the new weapon of the printed book, rapidly diffused in the later 15th century, the Humanists were able to create an autonomous secular education of the kind Europe had lost with the fall of Rome.

Classical Humanism

II. Structures and components of society

For most of the medieval period in most of Europe, the structure of society was determined by the difficulties of providing an adequate and continuous supply of food and raw materials. The proportion of grain reaped to seed sown was generally low; acute difficulties of transport made agricultural specialization hazardous and prices local and unstable. Hence, particularly between the 8th and 11th centuries, the proportion of the population freed of all agricultural tasks was low, and the materials of luxury or warfare were rare and highly prized.

The aristocracy that could be supported by such a society was both in form and origin a blend of Roman and Germanic elements. Late Roman society was marked by the existence of an aristocracy of wealthy landowners whose estates were worked by slaves either maintained in the household or housed on small plots about the great house. Other dependents, of rather higher status, might be freedmen or formerly independent cultivators, such as the *coloni*, whom war or financial distress had brought under the landowner's protection. Well before the formal end of the empire in the West, such landowners had been accustomed to deal out a usurped justice even over their free clients and had maintained armies of *bucellarii* (bodyguards) in their own defense, though rarely placing a high value on military accomplishments. With difficulty they retained their contacts with the towns and with the traditional urban education of their class.

Earlier Roman government had depended upon these towns, which were almost universally in decline by the 5th century. Vandal and, later, Muslim piracy disrupted the vital sea routes to Africa and the East; on land the impotence of local government made communications dangerous; and ever heavier taxation crippled trade. Long-range commerce and the more local urban industries declined, until even wheel-turned pottery became scarce or vanished. The retreat toward economic self-sufficiency and a barter economy was reflected in the near collapse of Western coinage. Over this society, rapidly receding into a pre-Roman localism, there presided, with ever-feebler effect, an emperor or emperors who attained office at the head of an army and according to no settled principle of succession. Once in power the emperor exercised a near-absolute authority over the army, the courts, and the administration, limited only by prudence and the likelihood of mutiny or assassination. Emperors were hailed as themselves divine before the conversion of Constantine early in the 4th century; thereafter, even a Christian emperor who enforced the precepts of the church readily took on a quasi-priestly character.

Decline of the towns

The Germanic invaders who settled in or along the frontiers of this empire lived very differently. Not all seem to have had kings, at least permanently, but the majority, whether as allies or invaders, entered the empire as warrior bands led by chiefs who depended for power on their continued capacity to win battles and to retain followers with gifts of present wealth and hope of more. Kings were ring givers, holders of great feasts, lords of men. Yet their authority rested on other grounds too—their descent from the gods and a belief sometimes current that divine favour was assured the people through the royal lineage. Some of the pagan kings (such as the Ynglings of Sweden) acted as priests, and a king's subjects usually followed him to the font for baptism as they followed him to war. Very special measures, including the earliest recorded Frankish consecration by a bishop, were required to legitimize the succession of a king from outside the ruling dynasty.

The politically active element in the people was represented by the warriors—the freemen—who formed the army; when this was gathered it represented the people for war or peace, and it was the army that acclaimed each new king by raising him upon their shields or enthroning him upon some sacred stone. Among these freemen there was a group of special importance and standing, set apart by their birth and their lord's favour—the *comitatus*, or *gasindi*, already known to the Roman historian Tacitus in the 1st century AD. For them, loyalty to their lord was the supreme virtue, celebrated in epic verse from *Beowulf* on. From their ranks were drawn the barbarian officers of post-Roman and Germanic society, though the highest such posts seem to have been monopolized by a restricted number of noble kin groups, distinguished by their descent and deeds rather than by any set title to property.

THE RULERS

Kings. In the societies in which these disparate patterns blended, beginning in the 8th century, the medieval kings were raised to office by a combination of ritual acts that revealed the elements of their power. Until the principle of male primogeniture became dominant in the late 12th century, the king would first be "chosen" in an assembly from among the kin of the last ruler, whose designation would carry great weight in cases of doubt; this choice was often not complete until the new king had travelled through his kingdom in what has been called a continuous election. After the consecration of Pepin the Short as king of the Franks by St. Boniface in 751, it became increasingly common for the king so acclaimed to be consecrated with a liturgy carefully modelled upon the Old Testament precedents of Saul and Solomon and to be invested with such insignia of kingship as crown, sword, helmet, or sceptre. Before or after this ceremony, the leading men of the kingdom came in to declare their allegiance, often performing symbolic acts of domestic service at the coronation feast. In these ceremonies the king secured both rights and duties. As the anointed of the Lord, he had a special claim on the obedience of the church and a measure of physical security; the murder of Canute IV of Denmark in 1086 horrified Europe, and the violent deaths of kings continued to be regarded as striking at the whole fabric of the divine and human order. This process, whereby the church sharply distinguished the king from the chieftain, in part explains an important transition of the 9th and 10th centuries. Thenceforward, the multiplication of kings, common among the descendants of Clovis or Charlemagne, ceased; only a formal act of the papacy under rare conditions was thought capable of legitimizing such later kingdoms as Hungary (999), Sicily (1139), or Portugal (1143).

Medieval notions of kingship

The church, which played so large a part in creating such a king, was active in prescribing duties; the coronation oaths and prayers insisted upon his obligation to protect the church, the defenseless, and the poor, to make war upon the heathen in the service of Christ, and to ensure that justice was done. It was chiefly in church councils that the king's duty to the whole people was emphasized against his relations with his warriors.

Even where, as with the Capetian dynasty between 987 and 1316, son succeeded father in unbroken descent, it was conventional to refer to the king as chosen by his people, and this became an active principle when there was no obvious claimant, as at the end of the Saxon or Hohenstaufen dynasties in Germany or in such exceptional cases as the founding of the Latin kingdom of Jerusalem in 1099. Among the larger kingdoms of the 13th century, only Germany, beset by frequent changes of dynasty and by the papacy's hostility to a hereditary empire, was still in practice an electoral state. Charles IV's Golden Bull of 1356 formally defined the procedures of election that remained in force in the empire until 1806, although after 1437 the throne was monopolized by the family of Habsburg. In Denmark, Sweden, and Poland the kingship was also formally and often prac-

tically elective, though hereditary right became absolute in France, England, and Spain.

Emperors. Theoretically, after 800 the emperors—the heirs of Charlemagne—stood above the kings. The appearance of a second Christian empire caused endless difficulties with Byzantium, and the nature of its pre-eminence was always diffuse and uncertain. It could be considered a recognition of facts—the attribution of a supreme title to the most powerful of the Latin kings—and many saw Charlemagne or Otto I in this light; but none of their successors enjoyed the same authority as did these founders of empire. The more general view, dominant by the end of the 10th century, was that the empire was specifically Roman, but in several possible senses. Because it was Pope Leo III who had taken the lead in recognizing Charlemagne as the first emperor in the West in 300 years, and, because Leo's successors had the undoubted right to consecrate later emperors at Rome, it could be claimed that the empire was an office within the Christian community, or *ecclesia*. The chief duty of such an emperor would then be the protection of the church and the enforcement of ecclesiastical discipline. When the emperor Henry III summoned the synod of Sutri of 1046, which ratified the withdrawal of three rival popes, he fulfilled one interpretation of this role; when Pope Gregory VII declared Henry IV deprived of his kingship in Italy and Germany for offenses against the church in 1076, he was employing another. Rome was, however, the city of the Caesars as well as of St. Peter, and some claimed that the emperors were the heirs of Augustus and the pagan emperors as well as of Charlemagne, with a universal authority that owed nothing to the church. This view was asserted notably by the apologists of Frederick I Barbarossa (reigned 1152–90), supporting their claims upon the revived study of the Roman civil law; modified by the doctrines of Aristotle, it survived in the writings of Dante and among the courtiers of the emperor Henry VII (reigned 1308–13).

Chief duty of the emperor

Without effective control in Rome, all of these views lacked substance, and such control was made impossible because of political stress in Germany under Henry IV and, later, the rise of the Lombard towns (which inflicted a serious defeat on Frederick I at Legnano in 1176) and the papacy's determination to establish an independent state in central Italy under Innocent III and his successors. After the death of the emperor Frederick II (reigned 1215–50), who had united the crowns of Germany, Italy, and Sicily in his own person, prolonged succession disputes on both sides of the Alps ruined the foundations of central authority in Italy and Germany. After the reign of Otto I (German king, 936–973; emperor from 962), the king of Germany had an exclusive title to the empire, though he could only secure it through coronation at Rome. The universal and ecclesiastical pretensions enshrined in this act had become a fiction long before the last imperial coronation in Rome, that of Frederick III in 1452, the bulk of whose reign was spent trying to maintain a foothold even in his own duchy of Austria.

Popes. If the imperial authority became progressively more confined, that of the popes made great advances, partly at the empire's expense. The foundation of papal authority lay, first, in the claim to inherit all the powers conveyed to St. Peter by Christ and, second, in the special position of the bishop of Rome. Until the 8th century, papal pre-eminence was complicated by the existence of jealous patriarchates of the East, not least the new Rome at Constantinople, whose emperor remained the nominal and sometimes the effective overlord of the city of Rome. By 700, however, most of the other patriarchs were subject to Muslim rulers, and Byzantine military weakness had forced such popes as Gregory I to take on most civil responsibility for the city of Rome. Dogmatic disputes in the Byzantine Empire over the veneration of icons (725–843) hastened a process of estrangement between the Greek Eastern and Latin Western churches that was to deepen eventually into near permanent schism. Rome then was left in solitary eminence in the West. Missionaries from Anglo-Saxon England, which had been converted by Roman agents, came to exercise a dominant influence upon the Frankish kings of the early 8th century; under Pepin the Short, Charlemagne, and the latter's son Louis the Pious, strenuous efforts were made to enforce a single liturgy, canon law, and monastic observance, the authenticity of which was guaranteed by its use at Rome.

Until the 11th century this confirming and legitimizing of action begun elsewhere was the predominant if not exclusive role of a papacy largely under the control of a series of such local Roman noble houses as the Crescentii and counts of Tusculum. The papacy favoured rather than led movements of active reform in the monastic and secular church. With the accession of Gregory VII, however, the papacy assumed the leadership of a movement that saw the inertia of the bishops and archbishops as the chief obstacle to ecclesiastical reform and viewed this situation as a consequence of excessive lay influence on their appointment and conduct. Accordingly, Gregory and his successors pursued a policy aimed at reducing the authority of the secular princes over the church (so provoking the long and bitter Investiture Controversy with the emperors Henry IV and Henry V) and at replacing secular authority with an ecclesiastical government based upon the written canon law and directed from Rome. By the pontificate of Innocent III (reigned 1198–1216) substantial successes had been achieved. The Fourth Lateran Council of 1215, attended by more than 400 bishops and 800 abbots from the whole of the Latin obedience and by representatives of all the greater princes, disposed of secular and ecclesiastical business on the widest scale. The difficulties that beset a full realization of the papal vision of a Christian and priestly monarchy were both internal and external. Internally, the process of centralization of authority made for slow and expensive decisions and was thought to demand a political independence in Italy that, in turn, involved long and costly wars with the Hohenstaufen and their successors north and south of the Papal States. Both centralization and the need to levy taxes on the church at large aroused the distrust and resentment of the local hierarchy, which could be exploited by kings (such as Philip IV of France in his contest with Boniface VIII) who were determined to secure their own right to tax their clergy and to submit them to the royal law.

Papal leadership

The rise of national consciousness that accompanied this resistance as well as the exile of the popes at Avignon made the pope's universal role as arbitrator in secular or ecclesiastical affairs harder to sustain. The double election of 1378 created a schism and weakened the position of the papacy. Because neither contestant would give way, the schism could only be ended by an external agency—a general council of the church held at Constance in the presence of the emperor Sigismund between 1414 and 1418. The energies of the abler reformers of the next 30 years or so were largely diverted to disputing the rival claims to authority of a monarchical papacy and a general council. With the failure of the most ambitious of these at Basel, the papacy emerged from the struggle again in Rome and again enjoying the plenitude of power; but the chief beneficiaries of the struggle were the secular princes, who had secured valuable concessions of control over their clergy as the price of their support.

THE ARISTOCRACY

The greater aristocracy. By 1100 a greater aristocracy had evolved over almost all the Latin West, marked by a combination of three elements. First, its members were normally the lords of a number of men bound to them by an oath of fealty and an act of homage, while they were themselves so bound to a king, a prelate, or the pope; the obligations such bonds involved were various, but the performance of military service was the most widespread and characteristic. Second, this aristocracy commonly exercised a large measure of judicial, financial, and administrative authority over its own men as lords and often over others as royal officers. Third, this was a landowning aristocracy that rewarded its dependents with grants of lands (fiefs), much as its own wealth had been enhanced by grants from kings, and maintained large

Elements of aristocracy

households from the proceeds of the estates retained in the lord's own hand—the domain.

To this combination, or some elements of it, the term feudalism is commonly though inconsistently applied; Marxists sometimes further apply it to a particular form of agricultural exploitation, the manorial system. But none of these elements necessarily supposes the others: in England it could be said that all land was received from the king, yet public authority remained unusually concentrated in his hands; in Germany the greatest princes came to hold rights of jurisdiction from the emperor rather than land, and their reciprocal duties were very slight; in southern France and Italy tenure for homage and service was rare but political authority was exceptionally fragmented.

Within the ranks of the greater artistocracy, certain distinctions came to exist. Originally, there seem to have been a limited number of very great families in Germanic society, defined by birth and sometimes uniquely described as free, but this division slowly gave way to others whose titles derived from functions rather than social rank. The term duke was widely applied to the lords of certain areas, which were incorporated in the Frankish empire without wholly losing their identity—*e.g.*, Bavaria, Burgundy, Brittany, Aquitaine, Saxony—and also to the lesser lords of the loosely constructed kingdom of the Lombards; the Scandinavian jarls or Anglo-Saxon earls after the reign of Canute (died 1035) were similar. In origin the word duke (*dux*) meant military commander, and it was on the basis of military need that the majority of 10th-century duchies emerged; for similar reasons, there appeared the margraves of Italy and Germany, lords of border land with a wider military authority. In Germany duchies multiplied in the 12th century, as in the creation of Austria in 1156, when the title implied membership of the highest rank in the social and tenurial scale; in 13th-century France and 14th-century England the term was only revived for great lordships (appanages) created for the cadets of the royal family.

The most widespread of such titles originally denoting public office and only later social rank or landed wealth was the Latin *comes* (German *Graf*, Anglo-Saxon *ealdorman*), or "count." *Comes* originally meant companion or member of the king's household of specially trained warriors; the appointment of members of the Frankish king's inner circle to the earlier administrative districts of *gau* or *pagus* caused the word to be applied to an officer of Carolingian government, then to a hereditary holder of office and rights, and finally to the head of a noble landed family. The terms duke, marquis, and count were themselves no necessary guide to relative wealth or prestige because some duchies were almost empty titles while the powerful lords of Flanders or Champagne were only counts. The institution of the *Reichsfürstenstand*, the class of imperial princes in late 12th-century Germany, was conceived as defining the highest rank of lay and church princes standing immediately about the throne. While the French institution was to become essentially an empty honour under the monarchical government of later kings, the German class was buttressed with important privileges that gave it enduring importance.

Recognition of rank

The origins of this greater aristocracy and its privileges were diverse. Some Roman senatorial families survived or absorbed their invaders; some descendants of independent chieftains of war bands that did not secure lasting kingdoms entered the Frankish, Anglo-Saxon, or Scandinavian nobility; the descendants of some specially favoured royal companions were able to hold onto the gains of their ancestors. Birth was from the outset probably an essential element of both exalted rank and access to royal favour, but important modifications to this took place between the 9th and the 12th centuries; slowly, the emphasis upon patrilineal descent and primogeniture became ever more absolute, and the extended kin group gave way to the dynasty. The passing of public offices into hereditary possession, the growing importance of military considerations, and a concentration of wealth and influence upon a limited number of indivisible castles all contributed to this effect. Whereas the great Carolingian families of the 9th century derived their names from an ancestor, their successors named themselves after the area of their rule or their principal stronghold.

Entry into this later aristocracy was possible by a variety of routes. Throughout this period the enduring hazard of central authority was that those appointed to maintain its interest would succeed in converting this representation into a right. In Germany even the unfree class of *ministeriales* sometimes succeeded in establishing such claims to the imperial fiefs and castles entrusted to their care and thus forced their way into the aristocracy. In France and England in the later Middle Ages, royal servants were commonly enriched by the opportunity to deprive or buy out those outside the circle of royal favour. There, too, in the 14th century a certain number of great merchants in the service of the crown were able to enter the ranks of the nobility; in Italy, however, the incompatibility of merchant adventure with noble birth had broken down much earlier. In later medieval Europe the emergence of more professional armies (especially in Italy and France during the widespread disorders of the 14th century) provided opportunities for successful soldiers of modest birth to rise into the ranks of an aristocracy coloured by ideals of an elaborate knightly code.

Lesser nobility. It was of the essence of high social rank that the aristocrat should have a large following of men who were themselves of free birth. Every great man, like every king, was made great by his capacity to maintain a retinue and reward his followers. An estate was valuable less for its revenues in money and goods than for the men whose services could be secured from it. The noble retinue was held together by gifts of gold, horses, weapons, or hawks, while those who served their lord well might hope for a gift of land. The numbers of such followers, or knights, were swollen by freemen who were driven by need to surrender their own land to the lord and to receive it back as a conditional grant by him; in return they received a protection modelled upon that extended to the lord's own blood kin. With increasing specialization of warfare, free birth tended to give way to the skills of cavalry warfare as the essential qualification for noble service; these knights (French *chevalier*, German *Ritter*) might live as retainers in the lord's household or hold land from him—a knight's fee, or fief—coming only at an exceptional summons to his feasts, courts, or wars. Many enjoyed a large freedom in the government of their estates and came to form a lesser nobility, bearing arms, using their own seals, and living in more or less fortified manor houses. As the apparatus and conventions of knightly warfare became more elaborate and the influence of the courtly romances more pervasive, membership in the nobility came to depend on knighthood.

Knights

The rise of the knight to this more exalted status was in part a function of the dissolution of the tight bonds of dependence his land tenure would once have imposed. This was a general phenomenon of the later Middle Ages, for the complexity of tenurial obligations made them increasingly inadequate either for defining status or for providing the lord with his honorable retinue. The later medieval principalities therefore owed their legal constitution to the powers of the lord in the exercise of his sovereign rights rather than to his personal claims on the services of his vassals; by a parallel development, the lord's household and following were commonly then paid for their services. Even at the lower levels of society, the same movement can be seen at work. In the 9th century a great estate would normally have a proportion of household slave labour; by the 13th century almost all the household servants would receive some wages, though food, lodging, and security still formed the bulk of their payment.

The church hierarchy. The greater aristocracy throughout Europe included a number of churchmen, for the sees of bishops and many monasteries had received wide grants of land, over which they often exercised more extensive rights than did their lay fellows, both as lords of tenants and as royal officers. Under Charlemagne and later princes, the state intervened to enforce universal acquiescence in the form of spiritual government over the

whole body of the laity, and this could be burdensome or even oppressive. Set aside from their fellows to a different and a higher calling, this aristocracy and its subordinate officers were also, in theory, recruited on quite different principles. In practice, however, high office in the church remained almost wholly the prerogative of the aristocracy and later of the knightly classes. Between the 8th and the 15th centuries it has been calculated that not much more than one-third of the 2,000 bishops appointed to German bishoprics came from non-noble families, and only five are known to have come from the dependent peasantry that formed the great bulk of the population. Certain monasteries and colleges of cathedral canons were explicitly reserved to those of the most carefully authenticated noble birth. Yet access to the highest church offices was in part dictated by other considerations; from the 11th century onward, royal or papal service and mastery of the disciplines of theology and canon law provided the means for men to rise on their ability alone. Suger—abbot of Saint-Denis, a monk of modest origins who became chief adviser to Louis VI of France, and regent for Louis VII—and died one of the most powerful men of Europe, or Thomas Wolsey, a butcher's son who rose to be cardinal and archbishop of York and chief minister to Henry VIII of England, illustrate the advancement that the church could provide, however rarely. Opportunities for this upward mobility were closely related to the existence of an effective royal administration capable of excluding the local magnates from a monopoly of patronage; the hope of modest advances provided many obscure but devoted clerical servants for the infant bureaucracies of 12th-century kings.

Access to church offices

Bishops. The original nucleus of church organization had lain in the bishoprics, groups of which, from the end of the 8th century at least, formed provinces normally presided over by an archbishop or metropolitan. The special interest of the church in political concord, which derived not merely from its dedication to charity but also from the vulnerability and wide extent of its property, made it for long the natural ally of kings and emperors; the anathemas of the bishops were regularly employed to reinforce the sword of the Lord's anointed. Being, if aristocratic, at least not hereditary magnates, the bishops were the allies upon whom the Saxon and Salian emperors of Germany and the early Capetian kings of France largely depended as a counterpoise to their greater secular subjects. Episcopal estates were treated as royal estates whenever the king journeyed about his kingdom, and the burdens of military service on German ecclesiastical estates were often unusually heavy; the royal administration was staffed and directed by the clergy, and the church's endowments were used to reward them. In return the bishops secured great privileges; the archbishops on either side of the Rhine—at Reims, Sens, Mainz, Trier, or Cologne—were at once powerful landowners and retainers of a wide measure of delegated royal authority, and these were only among the more visible examples of a movement in progress throughout Europe. Theoretically, the success of the Gregorian reform movement in the 11th and 12th centuries had stemmed from the distinction it drew between the bishop's absolute obedience to canon law and his subordinate obligation to the prince. But in practice, the bishops continued to depend on royal authority for defense against grasping lay neighbours, for the enforcement of ecclesiastical discipline (particularly in matters of heresy), and, in the later Middle Ages, for protection against the growing anti-clericalism of some of the educated laity and even the financial demands of the papacy.

Secular clergy. From the 5th century onward, the original constitution of the diocese, which rested upon the bishop and a small group of clergy living with him at a central point, altered substantially. Subordinate centres grew up within the diocese, some large and early ones served by groups of resident clergy but the great majority caring for only a small parish and served by a single priest. Recruitment to these was wide; some houses of canons and wealthy parishes provided revenues for men of high birth or influence, while others were served by men drawn from the peasant population, who were little more than domestic servants of the landowner who had built the church. The widespread grant of parish churches to monasteries and religious communities (as many as one-third of the parishes of a diocese could be so granted) and the prevalence of absentee clergy, such as pluralists holding several benefices or scholars engaged in study at the universities, meant that many churches were served by substitutes—vicars, who enjoyed only a proportion of the revenues of their office, often a quite inadequate one. Besides these, most parishes also provided some employment for a floating population of clerks in minor orders. With a rapid decline in the creation of new parishes in western Europe after the 12th century, it was the endowment of chantries, the setting aside of money and buildings exclusively for the saying of mass for the soul of the founder, that provided the greater number of new, if modest, benefices, though these were often attached to other parochial or charitable obligations.

The determined efforts made first by the early Carolingians and then by the reformers of the 12th and 13th centuries to provide an efficient system of supervision and control over the local pastoral work of the diocese ultimately produced an administrative hierarchy parallel to, though partly distinct from, the pastoral one. The rise of the bishop's formal jurisdiction saw the appearance of a host of legal and financial agents, ranging from the powerful archdeacon to the humble summoner, whose task was to enforce attendance at the church courts. The increasing centralization of church government at Rome in the 13th and 14th centuries required the appearance of other officers, proctors of bishops and abbots at Rome, collectors of papal taxes, and, later, such disreputable figures as the itinerant peddlers in papal indulgences and dubious relics—the pardoners.

Administrative hierarchy

Regular clergy. By the mid-13th century there existed beside this hierarchy another one, the regular clergy, living under a formal and corporate rule more demanding than the minimum canonical requirements of the seculars. Between the 5th and the 11th centuries this sector had been overwhelmingly monastic, composed of autonomous houses ruled over by an abbot with near absolute powers and devoted to the maintenance of the regular liturgical cycle; there was one pre-eminent rule, that of St. Benedict, which envisaged a life of corporate self-sufficiency as the norm. In the 11th and 12th centuries a number of variations on this rule appeared, designed to heighten the emphasis upon manual labour, corporate poverty, or solitary contemplation; in a more striking departure, some houses of canons under the flexible rule of St. Augustine were, it was hoped, to combine monastic rigour of life with active parochial work, while military orders founded in the course of the Crusades in the Holy Land and Spain and on Germany's eastern frontier combined the yet more disparate profession of war with monastic ideals. All these were property-owning corporations, as were the houses of nuns and canonesses that lived under similar conditions; their presidents, especially the abbots of the more well-to-do Benedictine abbeys, were often wealthier than many bishops and were drawn from a similar social background. Only among the Cistercian lay brothers was entry into the monastic life readily open to men of humble origin. The original followers of St. Francis (died 1226) were often drawn from a much wider variety of classes, but by the end of the 13th century both they and the followers of St. Dominic, originally intended to be educated preachers trained to combat heresy, had become orders of friars with a strong academic tradition, excluding all but the most talented of the very poor. These men renounced all property (even corporate), were especially engaged in the work of preaching (notably in the towns), and travelled constantly. By 1300 the multiplication of religious orders had come to a virtual standstill.

Clerks. Beyond the regular and beneficed secular clergy was a large body of clerks, in minor orders or in none, who had the crowns of their heads shaved in the tonsure without proceeding further, thus securing some of the legal immunities of the clergy with few of their

duties. The clerks of the royal or imperial chapels who formed embryo civil services represented an exceptional group of such men, but many more were to be found at the universities of the 12th and 13th centuries, and a still larger number yet were loosely attached to a parish church or joined a vagrant body of "hedge-priests" that canon law struggled for centuries to regulate or abolish.

Educated laymen. An important test for determining membership in the clergy was literacy. But in the 13th century (and occasionally earlier) there appeared in Europe a body of educated laymen of a kind that had not been seen there since the collapse of Roman secular education; lawyers and administrators in the royal service might now have an entirely secular career, and most merchants were necessarily literate. The 13th century produced two celebrated author kings, the emperor Frederick II and James I of Aragon (though even in the 14th century Edward III of England could barely write a few painful letters).

The growth of literacy

Between 1300 and 1500 the proportion of educated laity rose steadily; in part, the change was manifested in the growing importance of lay patronage in architecture, poetry, and the production of devotional manuscripts, the latter only the most striking example of a much more general and more active participation of the laity in the devotional life. On the other hand, the secular agents of government were jealous of ecclesiastical claims to privilege and immunity. Anti-clericalism in western Europe was as old as the Christian state, but the lawyers who staffed the sovereign governments of the 14th century had an alternative theory of the state to set against that of the church, and clerks such as Marsilius of Padua, William of Ockham, and Wycliffe were providing a theoretical justification for lay intervention in the affairs of the church as far-reaching as the claims of Pope Innocent IV for papal authority in the secular world had been a century earlier. At the opening of the 14th century the French chancery was directed by two lawyers, Pierre Flote and Guillaume de Nogaret, who launched a virulent and largely successful attack upon the universal claims made for the papacy by Pope Boniface VIII (1294–1303). Though even the revolutionary Hussite movement in Bohemia (1419–36) failed to subvert the authority of the church, the autonomy of ecclesiastical government was under pressure throughout Europe, largely because there now existed an educated laity able to perform the tasks hitherto reserved to a uniquely literate clergy.

OTHER SOCIAL GROUPS

Townspeople. The urban society of the Roman Empire in the West was breaking down long before the deposition of the last emperor in the late 5th century, and it continued to fade with few interruptions until the 10th century. The estimated population of Rome fell from more than 1,000,000 in the 1st century AD to 40,000 in the 7th; nevertheless, it long remained the largest city of the Christian West, although tiny compared to the greater Muslim cities or Constantinople. Many Roman towns continued to be occupied, often because they remained, or became, the seats of bishoprics or abbeys and so centres at least of consumption, or because their walls offered some shelter from a long series of invaders. Commercial activity continued in those ports of Italy still in touch with Byzantium and perhaps on a modest scale in some of the Lombard towns. It even increased in such northern ports as Duurstede in the Rhine Delta, the Viking entrepôt at Hedeby (in Schleswig), and the Russian cities trading through the Black Sea. By the end of the 11th century the recovery of urban life was more general. Merchants trading over long distances began to appear as an urban aristocracy, wealthy men anxious to secure a larger measure of control in the government of their towns. In Venice this development was already clearly visible, and the fleets and trade of this port stimulated a revival of town life throughout northern Italy. The origins of these early capitalists are much disputed: some may well have been fortunate peddlers, but more can be shown to have been members of the lesser aristocracy, or tenants or officers of great churches, with capital to invest in goods and transport. In Italy the participation of the landed aristocracy in the life and trade of the towns was shown by the presence of clusters of their tall stone towers within the city walls rather than on outlying hill tops.

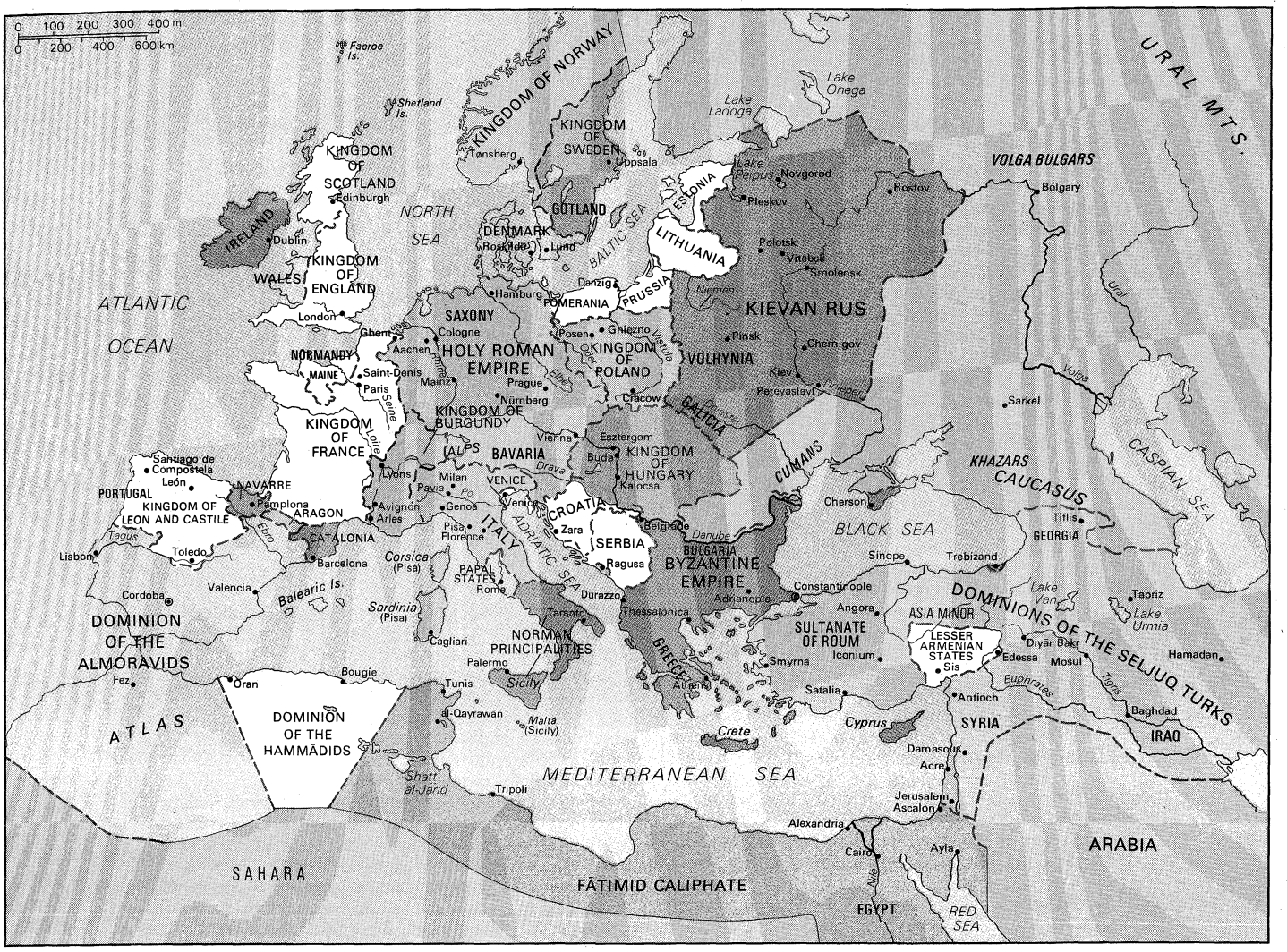
Recovery of urban life

Increasing political stability on land and contact with the Levant by sea saw the rapid increase of such trading communities in size and influence. The Italian cities of Genoa, Pisa, Lucca, and Siena rose to challenge the earlier dominance of Venice and Milan. In Germany the Rhineland towns were already a political force to be reckoned with in the civil wars at the beginning of the 12th century, but they came to be overshadowed by the great prosperity and activity of the Baltic towns of the Hanseatic League, such as Lübeck, Hamburg, and later Danzig (Gdańsk), and later by the prosperity of such South German towns as Augsburg and Nürnberg. In Flanders textile manufacture, based in part on imported wool, produced in the 12th century a precocious industrial community with a large unstable proletariat concentrated in such towns as Ypres, Ghent, Bruges, and Arras; alone of the northern cities these could compete in influence with the greater Italian centres such as the manufacturing town of Florence (with a population of perhaps as high as 200,000 in the 13th century, largely supported by its textile industry) or the equally large Milan, celebrated for its metalwork and, most notably, its armories. Paris by the end of the 12th century was already acquiring most of the qualities of a capital city for the fast-growing area of Capetian power, with a population estimated at 80,000.

By the 13th century the Italian towns commonly contained a group of aristocrats by birth or by wealth long possessed whose political dominance was challenged by the more substantial of the merchants. Below these were the retailers and masters of the smaller crafts and then the wage labourers, apprentices, and beggars, who were normally without a political voice but whose grievances sometimes broke out in violence and even bloodshed. The urban nobility of Italy was without true parallels north of the Alps and it was the greater merchants, or merchant adventurers of the later phrase, who formed the directing elite in most towns. Although even at the end of the medieval period the towns' share in the wealth and population of Europe was still very limited (over most of the Continent perhaps not above one-tenth of the total), their influence was much greater. Since the 14th century, access to the towns had been an important agent in social change, and the continuing or growing prosperity of some 15th-century towns was in marked contrast to the widespread decline of agricultural production and profit. Politically, the merchant interests of the towns were of the first importance, for it was only they who could provide the large sums of ready cash with which the kings of the later 15th century established royal authority over the magnates. To the degree that the king could continue to enjoy the taxes of the burgesses of his towns, he could maintain or extend his authority; without them he was reduced to competing with his own magnates on little better than equal terms.

Agricultural society. The countryside during the Roman period was chiefly cultivated either by the slaves of the great *villa* estates or by more or less free cultivators, sometimes bound by the government to remain on their land in order to maintain a tax-paying population but otherwise not labouring under grave personal disabilities before the law. In the Germanic societies of the invasion period, a threefold division of the agricultural population was common—the people *par excellence* (the *karl*, *ceorl*, or *bôndhi*), free peasants cultivating their own land, bearing arms, and so participating in the public assemblies of *gau* or hundred; a more obscure group often described as freedmen (*aldiones* for the Lombards and Bavarians, *Leti* among the Germans, Franks, Frisians, and also perhaps in Kent), possibly freed slaves, possibly the survivors of an earlier conquered population who enjoyed limited rights but did not usually play an active part in the courts; and the thralls, slaves either captured in war, condemned to their state by the law, or reduced to it by penury. Such bondsmen might sometimes have their own huts, plots of land, and

Germanic division of agricultural society



Europe and the Mediterranean about 1097.
From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

houses and therefore enjoy a minimal independence.

These forms tended to merge so that over much of western Europe by AD 1000 the characteristic villager (villein) held a substantial plot in the village fields but equally was expected to perform such onerous services on a lord's domain as ploughing, reaping, and carting and was subject to his lord's will in much the same fashion as the earlier landless slave. In addition, the village community would usually contain other men with smaller plots of land (cottagers), whose rents in labour and goods were correspondingly lighter, and also a small number of slaves, in the old sense of the word. All these unfree cultivators could be described as serfs, however diverse their economic conditions. In parts of Europe, notably in northern Germany and eastern England, the earlier free peasantry still formed an important element in the population, and by the 14th century their ranks were swollen by numbers of villeins whom changing economic conditions had allowed to commute their servile obligations for fixed rents in money. The settlers who had been persuaded to take up holdings in the planned land clearances of the 12th and 13th centuries by offers of considerable freedom of tenure had reached the same goal by different means. At the other end of the social scale, the landless wage labourer was now much more common.

The timing of this process, whereby many of the earlier free peasants were first assimilated to the legal if not the economic conditions of Roman slaves and then increasingly recovered their personal liberty to become rent-paying tenants, was extremely uneven. The cycle was complete for much of Italy by 1200 and for most of France and England by 1400, but the free peasantry of Scandinavia, parts of eastern Europe, and Castile was only beginning to feel the presence of a serf-owning class of landlords in the 14th and 15th centuries, at a time when it was fast becoming obsolete elsewhere.

Beggars and outlaws. A society that was at best only just above subsistence level maintained a large body of vagrants. Economic disasters such as plague or famine forced a desperate population to attack the crops of more fortunate neighbours; in the late 12th century, freelance mercenaries roamed Europe in search of employers and booty, while the 14th and 15th centuries saw more organized (and so more formidable) free companies of professional troops whose favoured fields of operation were France and Italy, where their captains traded their services like sovereign princes. All large towns either generated or attracted their share of beggars and petty thieves; in François Villon, 15th-century Paris produced a universal poet to express the pleasures and much more frequent miseries of this vagrant life. Town and country alike lacked any effective police force, and therefore, although the majority of known or suspected criminals could be dispossessed of land and property, even murderers were rarely apprehended. At times in 13th-century England only one in 100 murderers was ever brought to trial and convicted. The rest fled, many apparently into the forest to live like beasts or find an organized band of outlaws, idealized tales of which became widely current in the later Middle Ages in the legends of Robin Hood. The local community protected itself as best it could against such bands, but the assistance of a

The vagrant life

local knight at the head of his retainers could rarely be distinguished from the nuisance it was supposed to abate. The fate of these outlaws was bound up with that of the forests, and both were in decline by 1500. But with the first signs of a recovery of population, vagrancy and unemployment were probably increasing.

Women. A society directed by warriors and celibate clergy was not one in which women would exercise extended rights very often. After *c.* 1100 patrilineal descent was almost exclusively the test of nobility, while matrilineal descent was often the test of serfdom. The property rights of women, though protected by canon and secular law, were confined; it was a cherished freedom for widows to be allowed to refuse a second match proposed by lord or kin. Practice, however, did not altogether conform to this appearance. In barbarian society in the period after the invasions, and sometimes long after, matrilineal descent was often important—the house of Charlemagne traced its origins back to a daughter of Arnulf, bishop of Metz. Although few queens ruled in their own right, many exercised great political authority, as in the minority of their sons; the regency of Blanche of Castile for St. Louis in France was a notable and successful example. The remarkable Queen Margaret succeeded in uniting the three kingdoms of Sweden, Denmark, and Norway under her regency in the Union of Kalmar of 1397, an act that influenced the future of all Scandinavia. Similarly, although debarred from the priesthood by their sex, a number of women played a leading role in ecclesiastical affairs. St. Catherine of Siena and St. Bridget of Sweden played a major role in achieving the return of the papacy from Avignon to Rome in 1377; both were celebrated adepts of the spiritual life, the female contribution to which is also attested by the works of Margery Kempe in England. The influence of women was also felt in their role as patrons, sometimes of Christianity itself, as when a number of the invading chieftains of the 5th and 6th centuries were first married to Christian princesses and then converted. Queen Margaret of Scotland (died 1093) was responsible for the thoroughgoing reform of the church in her kingdom, which brought it rapidly into the mainstream of the Latin Church; another notable lady, Matilda of Tuscany (died 1115), had been the last refuge of the reforming papacy for almost a generation.

As patrons of the arts, and above all of poetry, Eleanor, duchess of Aquitaine and wife first of Louis VII of France and then of Henry II of England, and her daughter Marie, countess of Champagne, were the leading figures of their century. At Marie's request, Chrétien de Troyes translated Ovid's *Art of Love* and also wrote one of the first courtly romances to be based on Geoffrey of Monmouth's Arthurian history. In Marie de France the courtly romance found a skillful female writer—appropriately, because these tales, which evolved from the masculine world of the chanson de geste, reveal the growth of a new social convention in which women had a larger part and a higher function. Foreshadowed in the diffusion of the cult of the Virgin Mary in the 11th century, this new convention was developed in the Italy of Petrarch and Dante.

III. The operation of the medieval world

Diversity and complexity of medieval society

The most striking feature of medieval society was its peculiar diversity and complexity. Although scholars often conceived the world as a hierarchy in which all power was mediated from God according to a single ordered descent, in practice four distinct types of structure coexisted, overlapping and modifying one another profoundly but each with its own laws and objects. The first was the economic structure, essentially a diverse and inefficient agricultural society with islands of commercial activity. The second was the seignorial, the structures by which this economic activity was adapted to provide a surplus for a small class of lords and occasionally for great merchants. The third was ecclesiastical, in theory an autonomous economy of salvation in which all the forms of secular life had their spiritual counterpart. The fourth and for long the most tenuous was the centralized monarchical structure of the sovereign state. The triumph of this last over the earlier claims of church and magnate government marks the end of medieval society.

ECONOMIC PATTERNS

Latin Europe by 1500 covered a great diversity of lands and climates. A first division was that between the heavy soils of Germany, northern France, and Britain and the predominantly lighter soils of the Mediterranean lands of Spain, Languedoc, and Italy, in which vines and olives formed an important adjunct to grain.

Medieval communities. Both main types of agriculture were conducted by cultivators who lived, where possible, in large settlements, with their churches, mills, and barns, the whole settlement often walled, as in the *bastides* of southern France, or at least hedged against animal and human marauders. Both economies contrast with those that existed on their perimeters and in the less fertile or more broken country of the interior. In the Celtic lands of Ireland, Scotland, and Wales, in Spain and the foothills of the Alps, as well as in Hungary and the Latin Balkans, a largely pastoral economy flourished, depending on flocks of cattle, sheep, or goats and producing a quite different pattern of living. As oxen provided the essential power for the cereal farmer's plough, and milk, cheese, salt, and meat were as important to his diet as a minimum of crops were to the most pastoral of societies, so most medieval communities represented a precarious balance of forces between the needs of crops and beasts. Beyond these types of communities there existed various types of highly specialized settlements, such as the fishing communities of the North Sea coast, the salt evaporators of southwestern France, the fenmen and the miners, sometimes solitary and sometimes organized in tightly knit communities, as were the tin miners of Cornwall and Bohemia. With few technological resources and poor communications, all medieval communities were largely conditioned by their environment.

Some of the Mediterranean communities of Italian cultivators had been organized as great estates (*latifundia*) in the late and post-imperial period; each tenant held only as much land as could support him and his family while he worked on the lord's estate. Very early, however, the comparatively widespread use of coinage allowed landowners to abandon such direct exploitation, which was clumsy and time consuming, in favour of a variety of leases for terms of years or lives. By such means the community became one of independent cultivators, each with his own field. Similar tenancies prevailed over much of southern France, where a Roman Law much modified by custom and the decay of the judicature continued to regulate contracts and the functions of the market.

The triangle formed by the Rivers Loire and Rhine contained the chief area of nucleated cereal-growing settlements, though there were notable extensions of this into Britain to the west and into Franconia, Swabia, Bavaria, and the German east. In the Loire–Rhine region, with its heavy soils and wet climate, the earlier small enclosed fields of the Iron Age gave way widely to the characteristic open fields of the medieval vill, where the arable land of the community lay in large blocks cultivated by heavy, ox-drawn ploughs. The mechanical difficulties of turning such an implement and the need for effective drainage produced a characteristic effect of long narrow parallel strips running down contours of the hills. Such cultivation required a regular alternation of crops because there were few means available for restoring the fertility of the soil except allowing either half or one-third of the land to lie fallow each year. Similarly, the other resources of the land in pasture, woodland, water, and common grazing needed careful annual regulation because there was constant tension between the needs of men, crops, and beasts. This required an organism whose common life maintained some continuity regardless of the divisions that accidents of lordship might impose; later, the village was often a legislative body, enacting its own bylaws.

Open fields of the medieval vill

Until the 13th century this open-field cultivation continued to support the classical manorial organization. The lord drew his profit partly from his possession of a share of the vill's arable land, which was exploited for him by

his tenants for some of the week, and partly from the exercise of his right to compel his tenants to use his ovens and mills, at a price, and to control and exploit the use of the surrounding waste land. In practice such communities had always maintained a population of landless men—slaves earlier and labourers later—and there usually were tenants who owed rent or personal services of a less strictly economic variety. (Such free tenancy became widespread with the clearing of outlying lands in the 12th and 13th centuries, especially in the planned colonies along the frontiers.) Again, principally under the impact of increasing internal trade and a greater circulation of money, cumbrous forms of exploitation tended to break down in the later Middle Ages as the domain was leased out and the lord allowed his tenants to commute their labour dues into cash rents. When, in the late 14th and 15th centuries, the relation between land and labour shifted drastically in favour of the tenant over most of western Europe, the economic centralization of the local community focussed on the lord's estate almost vanished, though the need for common action in the agricultural cycle did not. In contrast to these tightly knit, often productive and vulnerable agricultural communities stood the much looser society of the higher lands, where settlement was necessarily dispersed and where social bonds rested much more exclusively upon the ties of the kin or clan than they did in the nucleated villages. Even the arable farmers in a scattered community had more independence than did their lowland fellows. The Icelandic sagas written down in the 13th century, in manuscripts typically preserved in farmhouses rather than in monastic libraries, provide a vivid picture of such communities.

Increasing volume of economic exchange

The increasing economic specialization of the period between 1100 and 1500, which saw vine growing and sheep rearing rise to the status of basic rather than supplementary occupations for whole communities, was made possible by a great increase in the volume of exchanges, itself a function of greater political stability. These exchanges took place pre-eminently in three settings: the markets, chiefly for local produce, held at short and regular intervals; the trading communities of the towns; and the occasional great fairs, annual events to which merchants travelled from the ends of Europe and the proceeds of which enriched such fortunate princes as the count of Champagne, lord of the fair towns of Provins, Troyes, Lagny, and Bar-sur-Aube.

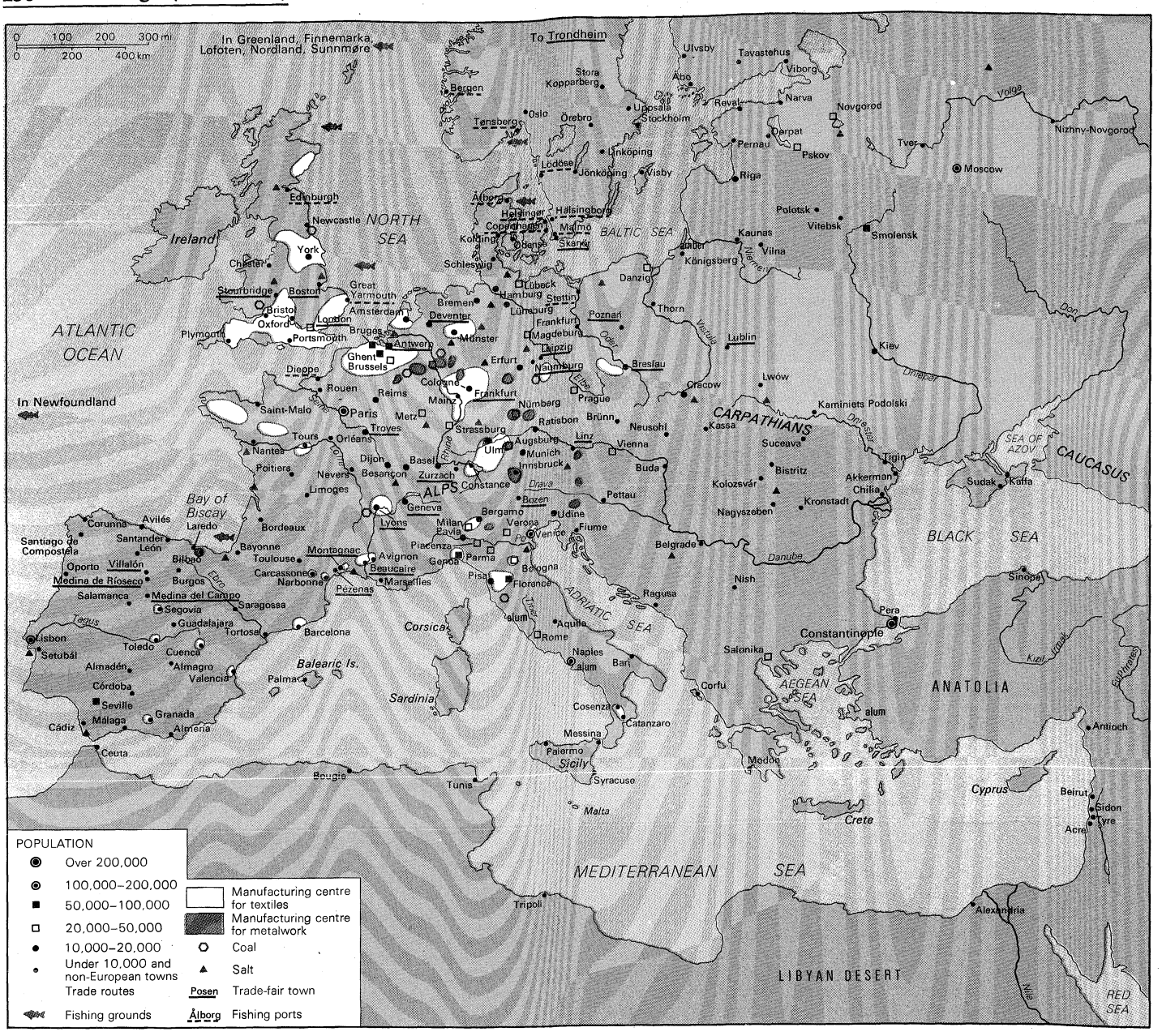
Urban growth. Early concentrations of population in settlements occurred for political, ecclesiastical, or defensive reasons as often as for commercial ones; but the period from the 11th century onward saw the widespread rise of classes of men engaged in the commerce of exchange or manufacture and settled for that purpose in urban groups. Many of these were originally only merchants in a subsidiary sense; and many towns were barely distinct from large villages, being surrounded by the town fields cultivated by the citizens and containing within their boundaries numerous livestock and small patches of cultivated ground. Yet, a common interest among merchants or craftsmen produced associations of considerable importance. Merchants guilds developed—groups of men whose interests extended far beyond the political horizons of the lords of most towns, anxious for the reduction of tolls and in need of their own courts of justice appropriate to the urgent demands of itinerant commerce. It was these who led the struggle for urban autonomy in Europe north of the Alps—a movement which almost always involved conflict with the local lords, especially bishops and abbots, but was often favoured by princes such as the counts of Flanders or Henry the Lion, duke of Saxony, in the German east—and thus the juridical phenomenon of the town as a corporate person emerged, created within a pattern of territorial lordships. Similarly, such towns acted corporately, much like their contemporary landed magnates, seeking to subordinate neighbouring communities to the authority of the city magistrates, to compel the man of the countryside to acknowledge the legal and economic primacy of the urban centre, and to divert profitable trade routes to their own advantage. The 12th-century crusading movement reflected such local hostilities—the Byzantine monopoly enjoyed by Venice was challenged by the direct access to the trade of the East offered by the crusader states, often maintained by the rival fleets of Pisa, Genoa, or Lucca. Long intercommunal wars, such as those between Pisa and Genoa or between Venice and Ravenna and Pola (now in Yugoslavia), saw the rise of a small number of city republics governed by elected consuls combining birth and commercial substance in a unique blend. Within such towns there existed other interests also driven by their common economic interests into corporate action.

In some towns, such as London or Bologna, a civil organization based on a geographical division into wards or quarters existed in potential conflict with the alternative divisions by occupation or wealth. Associations of craftsmen and retailers formed themselves in guilds that had many objects beyond the commercial. Among the earliest known forms of such guilds were those of prayer; and many guilds had their own chaplains, churches, and such distinctive religious pageants as the cycles of plays performed by the various mysteries (or crafts) at Wakefield in England. Many also maintained a form of insurance for their members, for the old, and for the widows. But all came to be devoted largely to the regulation of their crafts—especially entry to them—and to the struggle for their interests against the competing ones of their suppliers or of related trades. The interests of the masters of these crafts were often opposed to those of the greater merchants, thus producing a struggle for control of the machinery of town government. Equally, however, other divisions cut across this classification by trades. Within the guilds there was a struggle between masters and men, between those anxious to reduce competition and those bent on a larger independence. Occasionally, and most strikingly in the industrial towns of Flanders in the 13th and 14th centuries, this produced proletarian risings of wide impact; at Courtrai in 1302 an army largely drawn from the Blue-nails, the hired dye workers, inflicted a defeat upon the forces of the King of France and the Count of Flanders; frequently, the larger Flemish towns acted to free themselves from the power of the count or to maintain their communications with the English wool trade against the pressures of France.

Struggles for the control of town government

Though an urban proletariat of this kind was rare in medieval Europe, conflict between an urban oligarchy of landowners or great merchants (the *popolo grasso* of Italy) and the lesser craftsmen and labourers (the *popolo minuto*) was extremely common; it was at its fiercest where the cities enjoyed great autonomy, especially in Italy, where there was no effective or extensive central authority to maintain order. Hence, there were frequent civil wars in the Lombard towns, complicated by their external relations with each other, with the papacy, and with the empire; and there was a wide variety of forms of communal government, ranging from such attempts at radical democracy as the rising of the Ciompi of Florence in 1378 to an increasingly dominant model of dictatorship, such as that exercised by the Malatesta of Rimini, the Visconti of Milan, or even the Medici of Florence.

The independence of the Italian city-states, which was complete by the end of the 14th century, had rested in part upon the successes of the Lombard League against the empire in the late 12th century. North of the Alps, particularly in Germany and the German-settled lands of the east, towns established similar but rather more restricted liberties against the greater princes and formed into leagues of economic communal interest, of which the most notable was the Hanseatic League of Baltic towns, which secured a near monopoly of the profitable northern trade in timber, furs, wax, and amber. First clearly organized in the late 13th century, by the 14th the league was the dominant commercial force in northern Europe. In alliance with the Teutonic Knights and by an aggressive policy of embargo and blockade, it secured privileges superior even to those of the native merchants in Russia, Sweden, Norway, England, and Flanders. Only in the later 15th century did the league begin to decline, with the failure of the Knights, the increasing commercial ac-



Medieval industry and trade *c.* 1500.
From *Grosser Historischer Weltatlas*, vol. II, *Mittelalter* (1970); Bayerischer Schulbuch-Verlag, Munich

tivity of the western kingdoms, and the decline of the Baltic herring fisheries.

Urban independence on the Lombard or Hanseatic model was exceptional, however; in France, Spain, parts of Germany, and England the towns remained firmly within the framework of royal or princely government, not least because only the king could be relied upon to maintain political and fiscal stability at home and to negotiate commercial privileges abroad. Contingents from the towns formed a numerically formidable element in the royal army at Bouvines (1214), when Philip Augustus defeated a combination of Norman, Flemish, and imperial forces aiming to restore the earlier independence of some of the greater vassals. The protection of merchants had been a duty to which 11th-century popes had recalled kings; it was the financial support of the towns that was of importance in the triumph of the "new monarchies" of 15th-century France, Spain, and Tudor England.

Protection of merchants

FORMS OF LORDSHIP

Upon these economic structures there rested a variety of structures of lordship. The earliest that can be discerned in the centuries after the migrations were extremely complex, being functions of a power that was at once one of property, of kinship, of public function, and even sometimes of priesthood; only in the course of the later medieval period did these elements of authority come to be distinguished and sometimes divided.

Serfs. The economic bases of lordship in its least complex form lay in the ownership of men: the large household of slaves was the lord's property, and they worked his land. They enjoyed no rights against their lord, and he was responsible for maintaining peace and regulating their duties; such households of slaves became rare quite early over most of Europe. Domestic slavery was still an active force in the Iberian peninsula and Southern Italy at the end of the Middle Ages, though only in Sicily and the Balearic Islands were slaves widely engaged in large-scale agriculture. The responsibility of the lord for his own domestic servants remained near absolute, the successors of many household slaves were the serfs who settled on small plots of land; the descendants of other freer men were reduced to a similar status by poverty, war, the burdens of public taxation or even by the justice of the church. Like the Roman slave, the medieval serf had no public rights against his lord; he was *adscriptus glebae*

—so bound to his land that he could not leave it, and a part of its stock to be sold with it. The lord, by virtue of holding the land, was responsible for all police jurisdiction over his unfree men in his own court. His serfs were unable to marry without the lord's consent; in theory, their goods were his to tax at will during their lifetime and to confiscate at their death; their children were born into serfdom; and in all disputes there was no appeal from the lord to any higher tribunal.

Practice diverged at least in part from this grim theory. The agricultural cycle in which these serfs lived was conservative and complex, and custom operated powerfully to maintain it. Correspondingly, the disputes of tenants with each other and with the lord were regulated essentially by local custom, which was proclaimed by the body of the tenants in the lord's court. The rights of the lord to the labour of his serfs through the year and at the chief seasons of ploughing and harvest were also fixed by custom and were rarely (and slowly) altered except at times of sudden crisis. The rights of the lord to the property of the serf were equally confined, so that the serfs' duties were expressed as the obligation to render produce for the great feasts of the year or to take their grain to the lord's mill, their flour to his ovens, or their grapes to his wine press or to make fixed payments at the marriage of a daughter, a father's death, or entry into a peasant holding, whether by inheritance, marriage, or purchase. Any or all of the other services could be commuted for money payments also, and in western Europe, at least, the function of the lord's court in regulating serf labour on his demesne was becoming obsolete by 1500, though its other functions still had a long future.

The bond between serf and lord

In some parts of Europe the duties of the serfs included services that are less comprehensible; and the very intimate bond between serf and lord might contain elements more accessible to an anthropologist than to an economist or historian. In many cases the autocratic power of the lord was mitigated not merely by force of custom but also by close and frequent personal contact. The plans of early manor houses and castles show no extensive private rooms for the lord's use: he dined with his servants in his great hall, where he meted out justice. There was, until the later Middle Ages, little provision for private gardens for the lord's use; all but very great men could be found working at their own harvest. And feasts in the hall brought together the whole village community.

The lord had need of other services beyond those of his house and arable land; the greater Carolingian estates had even carried their complement of carpenters, smiths, potters, and weavers, free or unfree, as well as the usual haywards, shepherds, and beekeepers. Men personally as unfree as the serf might also undertake more responsible tasks; the *ministeriales*, pre-eminently of Germany, acted not only as bailiffs but even bore arms as knights, though in such cases the dignity of the occupation came ultimately to cancel the defect of birth.

Feudal bonds. Men of free birth were found over all Europe, though in varying numbers, who recognized the authority of a lord for some purposes—holding his land in return for rent or services less servile than those of the villein. Most areas of Scandinavian settlement, much of Saxony, and the Low Countries were marked by numbers of such men, who often pronounced judgments in their lord's court or escorted him through his estates, providing a contingent of men-at-arms in war and rent-paying tenants in peace. Lordship of this kind stemmed directly from the ownership both of land and of some of its tenants; the profits of lordship were, however, drawn from much wider sources than the labour of serfs or the rents of free tenants. Either by grant or by usurpation, a great variety of forms of indirect taxation were open to the lord. The levy of tolls on rivers, bridges, or roads was perhaps the easiest source of revenue open to any man; a license to hold a market in a vill would not be sufficient in itself but also involved payments for the holding of stalls and even a tax on individual transactions. Very great men might even mint their own coinage; consequently, the currencies of France, Germany, and Italy were extremely fragmented and often unstable.

The most distinctive form of medieval lordship, however, was that which is often called feudal, from the Latin *feudum*, meaning "fief," which was its central feature. In essence this was a fusion of the earlier *precaria*, a grant of land made for a fixed term in exchange for services or rents, with the very general commendation by which a man placed himself under a lord's protection by becoming his man or vassal. Some of those who had served their lord well would receive from him a benefice of land or revenue as a reward; others in dire need of protection would surrender their own land to the lord to receive it back as a benefice from him. When the tenant's right to his land became hereditary and his tenancy, or fief, ceased to be a reward for past services and became the reason for services to be performed in the future, feudal tenure was fully developed. These processes can already be traced before the end of the 9th century in the Carolingian Empire.

Lord and vassal. By the 12th century such tenure was to be found throughout Latin Europe. It was characterized by a number of symbolic acts. The first was homage, the process by which the man knelt and placed his hands between those of his lord, so putting himself at the lord's disposal and under his protection. The next was the oath of fidelity, sworn by the man to his lord, sometimes sealed with a kiss. Then came investiture, by which the lord handed over some token of the fief to his new man. This sequence was first fully described in the year 1127, but its various elements can be traced or inferred very much earlier, though they were not all necessarily found together. The bond so created was much more than a form of land tenure; it was first a human relationship, in which the lord assumed many of the rights and duties of a father and from which the man could escape only if the lord directly attacked his life or family. If the lord died leaving a child as heir, it was the duty of the tenants to maintain the heir's rights until he came of age; similarly, if a tenant died leaving children under age, their wardship was the lord's. Not the least of the man's obligations might be that of attending the lord at the great feasts of the year, which were at once parties, parliaments, and law courts.

Creation of feudal bond

Duties and rights. The duties of the lord to the tenant were usually only generally stated; he was bound to protect his man in war and peace, in the field, and in the law court. Carolingian legislation sought to ensure that every man had a lord to answer for him; a lordless man was a man in danger himself and a danger to others. The duties of the man to his lord varied enormously. Sometimes, even very early, the terms of his tenure were minutely defined. More often the general phrases of aid and counsel were called on to cover every possibility. As in the case of the servile tenant of the vill however, a general subordination to the lord first became fixed by custom and then often commuted into a money rent. The aid the man owed was primarily military aid, essentially service with the full equipment of a knight—lance, sword, helmet, mail hauberk, and powerful horse. A great man's household usually contained a permanent body of these knights, who were often landless young men of good birth but no fortune; but the obligations of the enfeoffed knights living in the estates they held of their lords were early restrained, by custom at least, to a period of service in the field of 40–60 days and sometimes a limited period of garrison duty or castle guard at the lord's fortress. Not all personal aid, however, was military: there were also the sergeanties—fiefs held on condition of performing other services. These ranged from fulfilling the highest offices in the lord's household, such as steward, constable, or marshal (or even jester), to picturesque or purely honorific obligations, such as providing the lord with an annual goshawk, a leash of hounds, or a pair of gloves. All such services might come to be commuted for money payments; by 1200 it was common for the bulk of the knights liable for royal and even magnate service to settle their obligations by the payment of sums of scutage (literally shield money). Because, by then, service was seen as a burden on land and because this land might be in the hands of a church, a minor, or a woman or might be

Payment of scutage

divided up among many heirs, such payments were a convenience to the tenant. The campaigning season and the high costs of a changing pattern of warfare often made money payment just as attractive to the lord.

Sometimes the lord might need financial aid for less strictly military reasons; again, custom came to distinguish the aids that he could levy from his tenants by right from those extraordinary ones for which the tenant's consent had to be sought. Practice varied, but the four most frequent were the knighting of his eldest son, the marriage of his eldest daughter, his setting out on crusade, or the ransom of his body.

Although in all these cases custom tended to establish the tenants' right in their land, there were three important traces of the lord's continued interest in what his ancestors had once granted: the relief, payable by an heir on entering into his inheritance; escheat, by which the lord recovered control of a tenure for which no direct heir could be found; and forfeit in the case of a tenant's treason or failure to perform his duties. The financial exploitation of these incidents of tenure and of the rights of wardship and marriage was among the most widespread of grievances against kings and great magnates in the 13th and 14th centuries.

The counsel that a tenant owed his lord also acquired a formal sense. By virtue of having tenants, a lord had an honour court, where disputes between tenants or between lord and tenant were heard. The tenant's duty to attend (to perform his suit of court) was of the greatest importance to the lord: it was the number and dignity of the suitors to his court that gave its judgments authority and stability. Around this central institution of the lord's court grew up in the greater honours much of the apparatus of a sovereign state. The greater officers of the lord's household, most notably the steward, or seneschal, played an active part in overseeing the lord's estates, and in the later Middle Ages they often formed a council with regular sessions to audit the accounts of their lord's estates and developed their own code of legal precedents. In the 13th century, treatises upon the customs of such tenurial courts appeared in considerable numbers side by side with studies in Roman or royal government (the *Sachsenspiegel* in Germany or the work of Beaumanoir in France may serve as examples). Only their homage to the king or emperor, with a variable liability to be summoned to his court, distinguished the greater magnates of France or Germany from sovereign princes.

Hierarchy of lordships

Theoretically, society could be conceived as a hierarchy of such lordships, with the sovereign at its head; but the practice was usually very different. In Germany the obligations of vassalage long retained traces of their servile origins, so that great men assumed them only with reluctance. It was correspondingly rare for a man to be the vassal of more than one lord; hence, a hierarchy of homages could appear in the late-12th-century *Heerschild* (a formal definition of social standing according to the number of intermediate lords between a man and the emperor) with the emperor at its head, the greater churches beneath him, then the imperial princes, who had done homage only to the emperor or the church, then counts, then noblemen, and so on. Even here, however, the consent of the other princes (the *Reichsfürstenstand*) was necessary for admission to the highest ranks, and the obligations such homage entailed were relatively very slight. Elsewhere—in France particularly—homage to more than one lord was frequent (the county of Champagne was held of 10 different lords), so that no such organizing principle could operate. What determined the permanence and vitality of tenurial networks was in part political and military circumstance, in part the extent to which some higher authority was able to intervene between lord and vassal, and in part the extent to which this lordship over men or tenants was able to absorb earlier administrative or public authority.

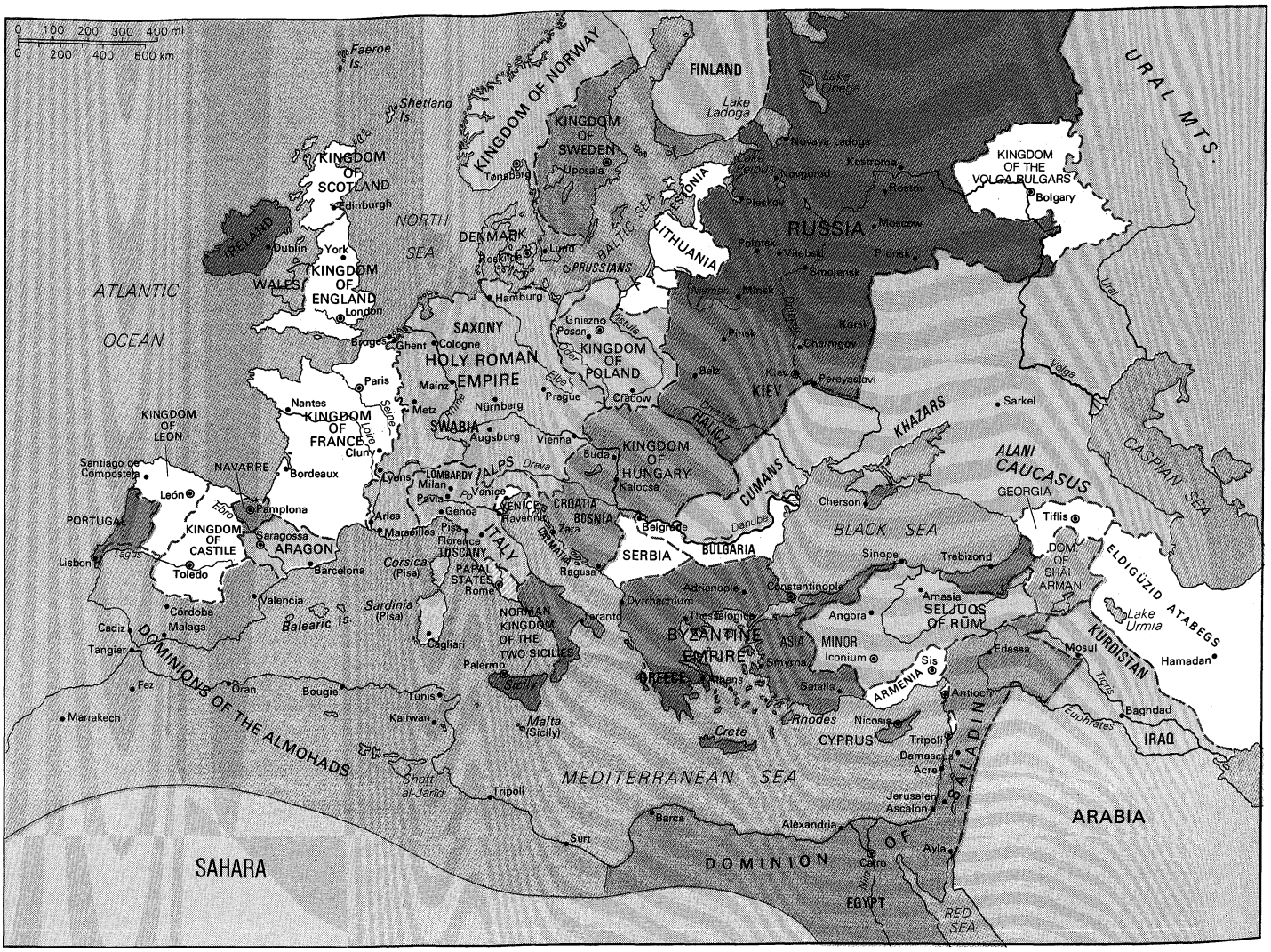
Acquisition of public authority. Some scholars have held that this passing of public authority into the domain of private right was the most essential and characteristic element of a feudal society; it was certainly extremely widespread. In the 9th century there existed two types of local government that might be described as public: first, the residual Roman institutions of Gaul and Italy, with such modifications as the Carolingians had introduced, and, second, the public assemblies in which Germanic and Scandinavian societies had long been accustomed to do justice. In practice, these institutions had largely fused in the period since the barbarian invasions, so that the bulk of Latin Europe could be divided up into *pagi* (German *Gauen*) roughly comparable to an English shire in size and function, each presided over by a count (*comes*, *compte*, or *Graf*) who was the king's representative, most notably at the local court (the *mallus* or *thing*). To this court all free men brought their disputes, the later distinction between civil and criminal justice being as yet unknown. The law of the court might derive from tribal custom, as it had been preserved in either such laws as the Salic, Ripuarian, Burgundian, and Bavarian codes or the collective memory, or, further south, it might be in the Vulgar Roman Law. In any case, it was the suitors of the court who determined the procedure and prescribed the appropriate forms of proof, such as the swearing of solemn oaths in special form, the undergoing of an ordeal (carrying a red-hot iron a prescribed distance, being plunged into cold water, holding a hand in boiling water), trial by battle, or, exceptionally, the scrutiny of sworn or even written evidence. All but the last of these were means of securing God's direct judgment rather than a merely human decision, but every preliminary step was determined by the court and particularly by a group of men learned in the law—*scabini*, *échevins*, *lagamen*, or *Rachinburgen*. Only in the 13th century, when the ordeal was condemned by the church and largely abandoned by the laity in favour of sworn and written evidence, did these lawmen give way to lawyers with special training.

The *pagus*, like the shire, was a military and financial unit as well as a legal one: it was presided over by a count whose responsibility it was to assess revenues, summon warriors for the royal army, and maintain the peace. To assist him in this task, lands were attached to his office, but he was also encouraged to secure the allegiance of the leading men of his *pagus*. In this way, to his authority as a royal representative he added his personal claims over the *castellani*, lords of the principal fortresses of the district whose military importance often allowed them to assume the direction of such subdivisions of the *pagus* as the *centena*, hundred, or wapentake. These, similar in procedure and competence to the courts of the *pagus* or shire, were early and generally in private hands; some may have been from their inception.

Fragmentation of authority

In the widespread fragmentation of authority of the 10th and 11th centuries, the offices of count and castellan became hereditary over most of Europe, so that it was often no longer possible to distinguish the authority of the count from that of the lord or the suitor of the *mallus* from the vassal of the count. Under such conditions, the jurisdiction and procedure of the public court merged with that of the court of vassals. Meanwhile, jurisdiction in the county, having become attached to property, could also be divided; and thus it often passed into the hands of lesser men who held it of the count as absolutely as he held his from the crown. Within the county there now existed liberties, or franchises, islands within which more or less of the count's justice was in the hands of a local lord and from which the count's agents were excluded. The greatest holders of franchises were the churches; the abbot or bishop's advocate exercised a near absolute secular authority that made such offices one of the greatest prizes of magnate competition. While some of these franchises were larger than counties or even extended over several, at the opposite end of the scale, all public authority over an area no larger than a vineyard might pass into the hands of a local knight subject to no restraints except those involved in his homage to a lord.

While the fiscal, legal, and military unity of the county either disintegrated or was reconstituted along the frontiers of a count's personal honour rather than earlier administrative or tribal divisions, there also appeared between the count and the king a number of intermediaries —dukes, marquises, or counts—holding several counties.



Europe and the Mediterranean about 1190.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

In origin, the duchies and margravates of the 10th and 11th centuries had been military commands, conferring on their holders only those rights over the counties of their area that were required by the pressing needs of defense; between the 10th and 13th centuries, these intermediaries fought a desperate battle against their sovereigns, on the one hand, and their neighbouring counts, on the other, to make this military authority (often over a former tribal district) into a more or less universal and exclusive jurisdiction. The results were extremely uneven: some of the German duchies that emerged from the civil wars of the 12th and 13th centuries were coherent units, though they were smaller than their predecessors; the duchy of Normandy was a tightly organized unit until it passed to the French crown under Philip II Augustus in 1202–04; but the dukes of Aquitaine and Burgundy had only a tenuous and formal hold on their outlying vassals.

New forms of warfare. The growth of such lordships had originally owed much to military considerations. The decline of the popular courts and public jurisdiction had been accompanied by the appearance of innumerable private fortresses, often little more than a wooden stockade upon an earth mound (the motte) with a larger enclosure at its foot for stock and dependents (the bailey). These proliferated in the Loire-Rhine triangle in the 10th and 11th centuries, though they did not spread widely in England or Germany before 1100. These fortresses were as important to defensive warfare as the mail-clad rider was coming to be in offense, for the possession of such strong points was a protection against all but the most formal and deliberate assault, which few commanders could keep a force in being long enough to conduct. In the 12th century the increasing number of stone-built castles, a more widespread use of paid troops, and the evolution of improved methods of siege warfare (some perhaps learned from the Muslims and Byzantines during the Crusades) altered the terms of warfare. Successful offense and defense required greater expertise and more extensive resources; correspondingly, the units of military independence tended to be much larger. The early motte and bailey castles could probably be thrown up in a matter of days, certainly weeks, and could be cast down yet more rapidly; but Château-Gaillard, built by Richard I of England to protect Normandy from attack along the Seine valley, took years to build and cost more than three times the whole annual revenue of the duchy: it was still capable of resisting prolonged formal siege with artillery four centuries later. In the 12th and 13th centuries, series of these defenses were being put up by such princes as Frederick II in Sicily or Sancho I in Portugal, but by the 15th century, gunpowder and engineering expertise were making such combinations of fortress and dwelling obsolete.

Private fortresses

A parallel development began to affect mounted warfare around 1200. Although chain mail was expensive, it was far less so than the combinations of plate and mail (or finally plate alone) that appeared in the later Middle Ages; similarly, the horses needed to carry this heavier equipment were scarce and exceedingly costly. The fully armed knight of 1300 was a more formidable figure than his predecessor of 1100, but he was a rarer, a more expensive, and a more limited commodity. The word knight had once meant farm worker; in the 11th century it meant

a mounted and well-equipped soldier, though by no means necessarily one of high social standing. By the 14th century, it was a term for an aristocratic elite, with its own elaborate ethic of love and war. The orders of knighthood founded in the 14th century by kings for their greater nobles and most eminent captains—such as the Order of the Garter established by Edward III or the later Order of the Golden Fleece founded by Philip the Good of Burgundy—reveal the extent to which the forms of chivalry were embraced by the whole range of the aristocracy.

Decreasing military importance of knights

At the same time, however, the military importance of the knight was reduced by the appearance of the more lightly armed men-at-arms or sergeants, and a trained infantry of pikemen and archers—crossbowmen, among whom the Genoese mercenaries excelled, and the longbowmen of English armies in Wales, Scotland, and France. The knights, with their prohibitive cost, increasing vulnerability, and declining flexibility, were becoming obsolete just at the time the ceremonial of knighthood was being most fully developed. The new armies of the 15th century were largely professional; they contained a high proportion of trained and well-armed infantry, which could be kept in the field for a whole campaigning season. The maintenance of such forces, with their extensive infrastructure of transport, victualling, and engineering, strained even the resources of a kingdom, and over most of Europe the near independent lordships of the great princes failed to survive. There were many great, rich, and privileged magnates in 16th-century Europe, but only in parts of Germany and Italy did they retain a complete local autonomy. It was rather by their influence on central government than their independence of it that they maintained political importance. By the 16th century, for the first time since the Roman Empire, centralized public authority covered much of Europe, though now within the frontiers of the national state.

CHURCH GOVERNMENT

Parallel with this structure of secular lordship stood that of the church, in which the opposing tendencies of central authority stemming from a single source and the collective authority of the community existed in a similar tension as in the court of a lord's vassals or in later royal councils.

The papacy. At the head of this church stood the papacy in Rome; after a long struggle with the German emperors of the 11th and 12th centuries, the political and formal independence of the papacy was ensured. After 1059 the popes were elected by the chapter, later called the College of Cardinals according to a carefully defined procedure. The intervention of all lay authority in papal elections was expressly excluded in the revised electoral decree of 1060, and that of 1179 provided for a strict majority principle in all papal elections; two thirds of votes made and still make a pope. Once elected, the pope possessed a widespread and unique authority: he was the final source of appeal; the only source of authoritative pronouncements upon dogma and doubtful points of canon law; the necessary source of the authority to make substantial change in the structure of the local churches, such as the founding of new bishoprics, or the transfer of bishops from one see to another. From the time of the exile at Avignon (1309–77) onward, it was normally the pope who provided, by his own authority, for every bishopric in Latin Christendom, and by a variety of channels he also disposed of a very substantial patronage in lesser dignities.

Growth of papal bureaucracy

Around the pope, conducting an ever-increasing volume of business, there grew up all the institutions of a centralized monarchy. The cardinals, chosen by the pope alone formed the papal council, or the consistory. The 12th, 13th, and 14th centuries saw the rise of specialized departments to deal with legal matters, matters of penance, and finance (the Apostolic Camera); and, most important of all, it saw the rise of an elaborate chancery to direct the issue of the various forms of papal documents (known as bulls from the use of the *bulla,* a leaden seal used to authenticate many of them). A papal chancery had existed from very early times, but the enormous volume of papal business, most of which was necessarily done in writing, gave the forms of the chancery in the later Middle Ages an exceptional importance—even in the shaping of papal policy.

Papal bulls not only contained decisions in legal disputes, valuable privileges for some churches, and the pope's diplomatic and pastoral correspondence; they were also a source of law, to be collected, studied, and compared by the canon lawyers with the writings of the early Fathers and the decisions of early church councils. The earliest centre for the systematic study of canon law was Bologna, but its influence was spread throughout Christendom. The early materials of canon law were the foundation of the reforming work of the papacy between 1050 and 1150; from then on, the rise of canonical science required a steady expansion of the active participation of the papacy in the everyday work of the church. By the 14th century the papal court was attended by an army of proctors who represented the prince and many of the greater churches, so constant was traffic with Rome.

The means of papal action beyond the immediate neighbourhood of Rome were provided by legates, either resident or specially commissioned, who enjoyed most of the sovereign authority of their master to hold councils, legislate, and punish. The penalties of disobedience were excommunication (cutting off specific individuals from all human contact with the Christian community), interdict (the suspension of all the sacramental functions of the church—usually directed against an entire community or state), or, against princes, the launching of military expeditions with a papal blessing. The religious orders directly under papal protection, particularly the friars, provided a means of action and an intelligence service to set beside that provided by the reports of the collectors of papal taxes, often Italian merchants, and by the Inquisition. Founded in the early 13th century for the detection and extirpation of heresy, this institution was presided over by inquisitors drawn from the Dominican and Franciscan orders, but they answered directly to the pope and possessed an extensive staff of men outside the orders.

The bishops. The local organization of the church depended essentially upon the bishops, grouped under an archbishop or metropolitan. The archbishop received a special vestment, a pallium, from Rome as a token of his authority; he was specially charged with the consecration of the bishops of his diocese; he might hold provincial councils or even visit bishoprics to examine their administration; on occasions he could hear appeals from a bishop or take disciplinary action against one. In general, however, his authority suffered from being squeezed between that of the bishop and Rome; the same was even more true of those archbishops who claimed to be primates (*i.e.*, to enjoy the obedience of other archbishops).

The bishops' administration

The bishop, originally elected in theory by the clergy and people of his diocese, was, in practice, appointed by some kind of compromise between the claims of the secular lord of the area, of the clergy of the cathedral church, and of Rome. By the 13th century most bishops possessed a central administration consisting of several offices: a specialized court for dealing with church affairs; a chancellor who kept the bishop's seal by which his formal acts were authenticated and under whose supervision the registers containing copies of these were maintained; a treasury for collecting the revenues—both from his estates and from more specifically ecclesiastical dues, such as a proportion of the tithes of the faithful or charges arising out of the ordinary exercise of his office. All these offices were staffed from the clerks of his household, increasingly to be distinguished from the chapter of beneficed canons, headed by a dean, who served the cathedral church. Although these canons were originally the bishop's household, by the 13th century they formed a compact corporation, sharing the revenues of the diocese with their bishop but often not appointed by him or subject to his strict control.

Within the diocese the bishop had wide responsibilities: he alone could ordain clergy, consecrate churches, receive children into full membership of the church at confirmation, or impose penances for the graver sins. No one

could teach or preach in the diocese without his permission, and he alone could admit novices to religious vows or bless the heads of religious houses. All judicial authority in the diocese was exercised by him either in person or through deputies, and without his sacraments the liturgical life would cease. In a more positive sense, he would teach and regulate the whole diocese as well as act as guardian of all church property, which could neither be received nor given away without his consent.

Some of these duties he could perform in person or, in case of absence, through a regent or vicar general. He could proceed through his diocese, visiting it parish by parish to scrutinize the conduct of clergy and people. He could summon the clergy (and some of the laity) to a synod once or twice annually; there judicial and disciplinary business was also done quite often between the 9th and 12th centuries, though later it was more an occasion for the publication of injunctions either of the bishop's own choice or drawn from recent provincial or papal pronouncements. Some bishops were diligent preachers, particularly to their clergy, and most performed works of charity in the feeding of the poor and the maintenance of hospitals and almshouses.

Much work necessarily remained in the hands of agents. Of these, the chief were the archdeacons, sometimes called the "bishop's eyes" since their duties were essentially disciplinary. Like the bishops, they made visitations throughout their archdeaconries and held courts to which they summoned both delinquent clergy and lay offenders against the comprehensive regulations of the church in marriage, usury, wills, heresy, blasphemy, and the observance of feast days. Their arbitrary procedures and frequent assessment of fines made them widely feared and resented. Most archdeaconries were subdivided into districts presided over by archpriests or rural deans, usually chosen from among the parochial clergy.

Archdeacons

Parish organization. The chief centres of the devotional life of the medieval church were parish churches served by a priest and a variable number of assistants. The priest was normally appointed by the bishop on the presentation of an influential lay patron. The archdeacon was supposed to examine each candidate's suitability, but the surviving records of visitations do not suggest that the task was often performed conscientiously. The parish was maintained out of the land attached to the church (the glebe) and the revenues from the faithful. From the 9th century the payment of tithes, a tenth of the annual proceeds of agriculture and trade, had been obligatory on Christians; although many tithes passed to religious corporations or vanished into the hands of unscrupulous collectors, they remained the largest single element of many parochial revenues.

The parish clergy were principally charged with the salvation of their parishioners; this was to be achieved by the regular administration of the sacraments, baptism, mass, and, above all, the last rites. The priest was also expected to offer an example of good living, expound the basic precepts of the Apostles' Creed and Lord's Prayer, promote concord among his parishioners, admonish sinners, and denounce grave offenders to his archdeacon or bishop. Since the priest was often the only man able to read or write in his community, he would often have to act as an elementary teacher and village notary. At a time when the houses of the peasantry and even most manor houses were modest wooden structures, with slight and crude decoration, the parish church was a striking contrast, an elaborate stone building with walls bearing paintings of the gospel narrative and the Last Judgment and with vestments and church ornaments of considerable splendour. As the building dominated the settlement, so the feasts of the church provided a rhythm of ritual and recreation that in part supplemented, in part reinforced that imposed by the seasons of the agricultural year. Yet there is abundant later medieval evidence to show how widely the older pagan cults survived under a thin veneer of orthodoxy.

The regular clergy. Beside this hierarchy of the secular church stood the regular clergy. The earliest widespread communities of monks had been in theory strictly subordinate to their bishop. In practice, however, widespread recognition of the monastic life as the pattern of religious excellence, the active missionary work of such monks as St. Columban and St. Boniface, and the favour of great families gave the monasteries a position of substantial privilege. In lands under Irish influence, the authority (as opposed to the sacramental power) elsewhere exercised by the bishop was often held by an abbot; but much more widely the greater monasteries enjoyed a large immunity from episcopal influence. In the 11th century and later, such houses—led by Cluny—placed themselves directly under the protection of Rome as a means of excluding the bishop's authority; and for the rest of the Middle Ages, quarrels over the bishop's right to secure an oath of obedience from the abbot, to visit the monastery to scrutinize its observance, and to exercise his office over the numerous churches and clergy attached to such abbeys were almost commonplace. Parallel to the growth of such conflicts was the rise of congregations of regulars, in which houses of a common observance, such as those of Citeaux, Hirsau, or Prémontré, were organized in a hierarchy that allowed the order itself to discipline and correct its members, to legislate for the whole congregation, and to ensure a common rule of life among its far-flung membership. After 1215 all the Benedictine houses not involved in such arrangements were required to organize themselves into local general chapters, though the tradition of autonomy made such provisions unevenly effective. The way of life of the friars, with their dispersed and itinerent congregations, involved a more complex organization modelled upon that provided by St. Dominic in 1216. Under Dominic's organization, the order was divided and subdivided into provinces, ruled by masters, and chapters; elaborate regulations provided for the election of masters and legislation for the order was promulgated in assemblies of elected representatives. Most of these features were taken over by the Franciscans before the end of the 13th century. Relations with the hierarchy of the secular church were sometimes strained because the friars were under the direct protection of Rome and enjoyed a correspondingly extensive, though much contested, immunity from episcopal jurisdiction. In this way, many of the religious orders were involved in a hierarchical system of centralized dependence parallel to that of the secular church; and in some respects they were more responsive to the needs of Rome than were the bishops. Certainly, the religious orders were widely regarded as less identified with ecclesiastical nationalism as pursued in the 13th and 14th centuries.

Direct protection of Rome

Church councils. Although ecclesiastical authority, even in the orders, was exercised on a largely monarchical pattern, the church also contained important elements of corporate authority; it was by definition a society in which there was a distinction of offices but not of persons. The earliest genuine law of the church was substantially contained in the decrees of the ecumenical councils of the 4th and 5th centuries; and such gatherings continued to be held in the East, though their authority was not accepted in the West. After the formal ending of the investiture contest, Pope Calixtus II summoned the First Lateran Council of 1123, claimed as the ninth ecumenical or universal council of the church and the first to be held in the West. The legislation of this and its successors (the last being that of Vienne in 1311–12) was given very wide currency, although according to the general opinion of canonists it was of no greater binding force than the decrees of the pope alone. In the 14th century, opponents of papal authority, such as Marsilius of Padua, viewed the pope as an officer of the whole church subject to its control when assembled corporately in a general council. The double election of 1378 and the prolonged schism that followed allowed such theories a wider currency. It was the Council of Constance (1414–18) that ended the schism, but that of Basel (1431–49) introduced another —in part as a consequence of the belief that the pope was subject to the constant surveillance of the council, standing in the same relation to it as an Italian magistrate stood to the city that appointed him. Such corporate doctrine failed because the council became itself less and

The Lateran councils

less representative; the limitations upon central authority that emerged were those imposed by the enlarged rights of secular rulers over the national churches.

The value of councils of bishops to promulgate decrees and pursue reform for a kingdom or province was generally recognized, but their ecclesiastical authority was, in theory, extremely limited. In practice, the legislation of provincial councils and even synods sometimes allowed local customs or variations from the universal law of the church, but this was by way of dispensation. There was no innate legislative authority resident in the members of the church to set aside or counter that of the Universal Church; and the major function of such councils was rather to select, emphasize, or interpret such aspects of the general law as the decisions of the general councils. The purpose of the participation of many bishops was to give these pronouncements immediate weight, to ensure them a wide circulation, and perhaps to encourage a common understanding of their meaning and importance. Councils, however, were sometimes summoned by the king to assert his divine authority over his turbulent subjects, to enforce his policies with the double sanction of the secular and spiritual sword, or to confront the pope with the consensus of local ecclesiastical opinion as a means of putting pressure on him to abandon a policy. Such gatherings, however, rarely claimed to constitute autonomous local churches; and the nationalist claims made on behalf of the French church under Philip IV the Fair or for the English church under Henry VIII rested on the authority of the sovereign in his kingdom, not the corporate sovereignty of a council of the national church.

The universal claims of the papacy, even if practically confined to the Latin obedience, were constantly threatened by the strong secular particularism of Europe's medieval history. Between the 5th and the 12th centuries the gravest danger was one generally subsumed under the term simony—that is, the constant likelihood that ecclesiastical rights would be subordinated to temporal interests and then fragmented, like them, to become merely aspects of private property. The hereditary lay abbots of the 9th century; the parish churches built as financial investments by their founders and bought and sold amid general acquiescence between the 7th and 12th centuries; and a widespread trafficking in offices, never wholly eradicated, all illustrate the gravity of this danger. These abuses were formally abolished and actually much reduced by the reformers of the 11th and 12th centuries, largely by means of a centralized law. Between the 13th century and the Reformation, the danger was that the political and linguistic boundaries within which the new national states were growing up would also divide the church. Against this danger, the traditional weapons of papal authority proved much less effective, and new ones were not forged before the end of the period.

ROYAL GOVERNMENT

The growth of the centralized institutions of predominantly royal government is a phenomenon that was involved with the rise of certain technical skills in the collecting and organizing of information. Until the 13th century there were no maps of great practical value; the statistics occasionally collected by medieval governments were often inconsistent and made according to erratic principles. Until near the end of the period, the usual methods of accounting and bookkeeping precluded most calculations performed by a modern government. It is a notorious defect of medieval chroniclers that the numbers they attribute to populations or even comparatively small armies are almost invariably the purest fantasy. The technology of transport was equally primitive; the best roads in the 15th century were still the Roman imperial streets, for all their thousand years of neglect, and winter flooding still affected most river systems in spite of some large-scale drainage works that had been undertaken in the Netherlands, Italy, and elsewhere. Under such conditions a central government wholly free of "constitutional" restraints would still be obliged to leave its local agents a large measure of independence, and the custom of the neighbourhood was necessarily the chief regulator of political as well as social and economic relations.

The royal household. The seed from which all medieval institutions of central administration were to grow was the immediate personal household of the king; its members were the only permanent staff he had, and only their intimate personal association with their lord was a guarantee to others that they did indeed represent his views and bear his authority. The essential elements of the early royal household therefore provided the framework of the first departments of state.

The chief elements of the royal household were four—the hall, the chamber, the chapel, and the courtyard, with its horses and stables. The whole household, but in particular the hall, which was at once palace, law court, and dining room, was directly governed by the steward (*dapifer, seneschal,* or *drost*), under whose direction the guests were seated and the feasting conducted, while the wine was under the charge of a butler (*pincerna, Oberschenk*). Under the later Merovingians in Gaul, the mayor of the palace (literally chief of the house) became a figure so powerful as first to overshadow and then replace his king; elsewhere, though less powerful, the steward was the usual chief deputy of the king. Under the Capetians of France he was charged with the annual scrutiny of the accounts tendered by the king's local bailiffs, the provosts; in Normandy, Jerusalem, and elsewhere he was the natural choice as regent in his lord's absence; the butler shared some of the prestige of the steward but had yet few defined functions.

Officers of the hall

The chamber, the room in which the king slept or took private counsel, was also the natural place to store his treasure; hence, the chamberlains were often specially charged with the collection of revenue and handing it out as the king had need. The papal treasury was known as the Apostolic Chamber, and the papal chamberlains were widely entrusted with financial missions.

The chamber

The chapel, containing the royal altar and relics (the term chapel derives from the *capella,* the short cloak of Saint Martin preserved among the chief relics of the Carolingian royal treasury), was served by chaplains, to say the daily mass, assisted by a body of clerical assistants. As the chief and sometimes only literate members of the household, these men were responsible for drawing up such documents as the earlier kings required; among their number and often at the head stood the chancellor, whose special task was the authentication of royal acts, usually with the seal, which he kept.

This royal household was constantly on the move, carried on carts or packhorses, and hence the great importance of the last two major household offices, those of the constable and marshal. The duty of the count of the stable (constable) was closely associated with the organization of the army, and hence the term came to be used of commanders of garrisons as well as the central household officer. The office of the marshal was originally more humble but shared the military fortunes of that of constable. In the 14th and 15th centuries constable and marshal came to a new importance as military commanders and correspondingly acquired judicial competence as presidents of the courts of chivalry, which dealt especially with military discipline, the division of ransoms, and the right to bear a coat of arms.

The growth of a permanent bureaucracy. These early household offices, with characteristically unspecialized duties, changed greatly under the impact of two forces. First, these central offices had a tendency to become hereditary, much as the local offices of count or duke had done; since the domestic service of the king, at least on public occasions, was itself a very great dignity, the most powerful families claimed the right to perform it, so that the office came to be the prerogative of great magnates who were too preoccupied elsewhere to perform their duties in person. Even the chancellorship sometimes became attached to certain archbishoprics—Reims or Mainz, for instance; since the king's domestic needs continued, a distinction evolved between the hereditary dignitaries such as high steward or archchancellor and the men of much humbler rank who actually performed the

routine duties of the household. Second, the increasing volume of business done in the king's name—judicial, administrative, or financial—demanded ever more elaborate records and a more extensive staff; therefore, the offices of government were less mobile, and a physical distinction became common between the constantly itinerant household about the king's person and the more cumbrous (though rarely wholly static) departments of permanent officials. Furthermore, the processes of government, especially in the chancery or (particularly well documented) the English exchequer, became arts or mysteries that demanded a staff of financial or legal experts with a specialized training. Thus were born the chief departments of state; well beyond the end of the medieval period, however, their principal officers were still considered the king's servants in a literal sense. This household character of public office made the distinction between loyalty to the king's person and loyalty to the office extremely hard to draw and frustrated many early efforts at "constitutionalism."

The chancery

Of these departments of state, the chancery was perhaps the most essential, for without a means of transmitting a number of recognizably authentic commands or recording the business already done, no large measure of centralized activity was possible. The first sign of the emergence of a chancery proper is the appearance of stereotyped formulas for the drafting of documents, classified according to their purpose. The papal chancery was the earliest office to develop this skill on a large scale; the rhythm of the text and the forms of authentication for papal bulls were already formalized before the mid-12th century. Similar tendencies are found in England in the reign of Henry II, and in France only a little later. (The urban notaries of Italy and parts of southern France were already using formalized business documents, though for a much more restricted area.) The second sign is the appearance of a substantial collection of records; apart from the financial records, dealt with below though often compiled by chancery clerks, the essential element was the keeping of copies of documents issuing from the chancery, distinguished according to their character. Papal registers had probably been kept from a very early date, and a later copy of the Register of Gregory I still survives; an imposing and ever more complex consecutive series of original registers survives from the early 13th century. These were in the form of books. In England, where the great period of initial expansion lay in the period between 1190 and 1216, and in France, where the early royal archives have suffered much graver losses at the hands of time, the characteristic record was a series of parchment membranes stitched together to form long rolls. These archives evolved rapidly through the 13th, 14th, and 15th centuries, with the preservation of many more classes of document, including informal memoranda. Such records served the purposes of both governor and governed, for they provided precedents and the material for the reform and improvement of administration for the king's servants. They also provided authentic copies of title deeds or privileges for a subject at odds with his fellow or even the king himself. The conventions of the chancery had a marked tendency toward autonomy; efforts at magnate control in England or extensive reform of papal administration in Rome were partly frustrated by the elaboration and conservatism of chancery procedures that might at other times offer a useful defense against arbitrary government.

Among the most immobile elements of medieval government were those connected with the collecting and checking of the royal revenue. Until the 13th century, the only coinage current was the silver penny (or denier), so that large sums could only be transported packed in barrels of great weight. It was therefore natural for royal treasuries to appear as places of permanent deposit, from which the itinerant court or army could be supplied at need. Often the treasury would also be the first site of any permanent archives. The rendering of accounts by the sheriffs, provosts, bailiffs, or seneschals who were local collectors of royal revenue was a matter that could not be performed at a wholly itinerant court. Not only was it necessary for auditors and agents to have a known place and time at which to meet, but all except the least sophisticated forms of accounting required the keeping of records of former debts and present liabilities. Under Charlemagne, a *Capitulare de villis* envisaged a wide enquiry into the estates of the emperor, though only fragments of the returns survive; similar surveys were being made in 11th-century France and Germany on a small scale, but no early text can rival the record of Domesday Book (compiled in 1086/87), a survey of almost all of England, shire by shire and fief by fief, drawn up by the clerks of William the Conqueror. It is still preserved among the English public records.

The growth of permanent records

The rise of the exchequer as a semipermanent office and court for the hearing of accounts and the adjudication of financial claims was an early and striking feature of Anglo-Norman government on either side of the Channel. The exchequer took its name from the checkered cloth used by the royal officials as an accounting device. The cloth was used as a kind of abacus and may have owed something to Arabic skill in mathematics transmitted by the Norman settlements of Sicily. Certainly in existence soon after 1100, its first surviving Pipe Roll of 1130 is the earliest record of receipts of its kind for all of medieval Europe. In France the Capetian Kings for more than a century used the Knights Templar as their bankers, so that it was not until after the withdrawal of his treasure from their hands by Philip the Fair in 1295 that a fully independent Chambre des Comptes emerged, with a comprehensive staff of financial experts.

The cumbersome machinery of such financial bodies was a grave impediment in times of political, financial, or military crisis, although as in the case of the chanceries, due process offered some protection against arbitrary government. The frequent wars of the English kings and their long absences in France, as well as their desire to escape the oversight of great officers approved by their magnates, encouraged the appearance of various financial systems within the household to compete with that of the Exchequer; the wardrobe under Edward I, for example, expanded from its original function as the king's domestic treasury to become the organ responsible for the payment of his whole army or, under his descendants, of the chamber itself.

The king's council

As these departments of state became more fixed, professional, and independent, royal councillors who gave a political direction to the government acquired a distinct existence. From the beginning, the king's household had been of shifting composition, and a number of men enjoying his special confidence were often to be found there even if they were not holding household office. In the early 13th century the king's more or less permanent advisers began to take on a more formal character as the king's council. This council, often given definition by the taking of a common oath, was as omnicompetent as the king it served, and correspondingly, in times of political crisis, magnates or great assemblies sought to impose their nominees—as in England in the crisis of 1258–65 or France in 1356–57 or Aragon under Alfonso III. The direct and personal nature of the council's functions made such efforts almost always abortive. Only in Scandinavia, where the union of the three crowns involved prolonged regencies and uncertainties of succession, did the Råd enjoy a large independence from its king, though many German princes of the 14th and 15th centuries were compelled by their estates to accept a nominated council. In the 15th century the powers of last resort possessed by the king and exercised in consultation with his council allowed the formation of offshoots of this council as prerogative courts of justice, administration, and finance.

Birth of parliamentary bodies. Over the same period, the larger assemblies in which some royal action had long taken place were also taking on a clearer form and defined functions, issuing in the assemblies of estates to be found over most of Europe in the 15th century. Three distinct influences lay at the origin of this development—ancient custom, feudal law, and administrative convenience. The Germanic tribes described by the Roman his-

torian Tacitus in the 1st century AD gathered regularly in arms to determine matters of general importance. In some respects it was the size of such gatherings that decided the size of the kingdoms to which they elected chieftains; such gatherings of the warriors continued well into the period of Frankish rule, even in those areas where the powers of the king were very great. In Scandinavia meetings of the *thing* never wholly passed into the power of magnates or royal officials, though elsewhere they changed radically.

As the old *mallus* or *thing*, with its lawgivers and its president, had become the court of an hereditary count and his vassals, so the assembly of freemen was widely altered in the same way. To the king's great courts his chief tenants were summoned by virtue of their obligation to give their lord counsel; the number of those who came was the critical test of the king's authority. The decline of French royal power in the 10th and 11th centuries or of German kingship in the 11th and 12th centuries can be plotted on maps by examining the composition of their great courts, especially at the chief feasts of the Christian year (or at principal landmarks in the life of the royal family, such as marriages and the knighting of the eldest son).

Since these large sessions of the tenants were also the most solemn courts a king could hold, great issues would be brought to it; since they were usually held at known times and places, those who desired the king's help in securing justice would seek them out. There were, however, further reasons why the king should try to enlarge the attendance at his courts or councils; the bonds of land tenure in most of Europe before 1300 had long ceased to articulate all the elements of society. To regulate the affairs of merchants, for instance, the feudal bond was worthless to secure counsel or consent. More important still, the obligations of feudal tenure no longer provided the forms of taxation needed for the defense of the state, and whatever methods of raising money supplemented them required the consent of the community in a new sense. The princes of the late 13th century claimed to be lords of states, not merely of associations of men; in their conflicts with the papacy, in their assertion of legislative authority, in their claims to the financial support of the whole community, kings required a very general assent. Unless qualified representatives could be gathered, the king's will could not be known or the justification of changes publicized; the rise of the representative assembly is parallel to the rise of royal propaganda.

Representative institutions

In the 13th and 14th centuries these considerations produced a variety of experiments. In France after 1300, two meetings of the estates of magnates, churchmen, and burgesses were held often: one for the provinces of northern France (Languedoil), another for culturally different provinces of the south (Languedoc). Provincial gatherings of this kind were naturally prevalent in the states of Germany, where the dynastic disputes and poverty of many of the princes made them peculiarly subject to the pressures of their estates of knights, burgesses, and clergy; the imperial Reichstag, attended theoretically by the tenants in chief (chief vassals) of the emperor and representatives of the imperial towns, was intended to cover the whole of Germany but was as powerless as its central authority. In the kingdoms of Spain, signs of development appeared very early, representatives of the towns being summoned in 1188 in Leon; by the mid-13th century, the presence of townsmen in representative assemblies was customary throughout the country. In Aragon magnates and knights were separate elements, but elsewhere the more usual pattern of the three estates of nobles, churchmen, and burgesses prevailed. The English gatherings to which the term parliament came increasingly to be applied differed in important respects. From early in the 13th century, the king had summoned his tenants in chief to meet at the same time the central courts of justice and finance were in session; thus, royal officers, innumerable representatives of the shires and boroughs, and the magnates were gathering at one time. By the end of the 13th century, it was becoming common for special representatives to be summoned from the shires and the boroughs to attend these parliaments; by the mid-14th century, their attendance had become the rule, but the clergy had largely withdrawn, except for the great ecclesiastical magnates who sat with the temporal lords, while the knights of the shire and the burgesses of the towns met together as a single body. Hence, there emerged a quite exceptional body with two chambers and a permanent representation of the non-noble landowners to contrast with the much more widespread model of the three estates.

English Parliament

The business undertaken by these gatherings was extremely various. For reasons essentially of convenience, it was customary to publish legislation at such assemblies, as church councils had done for centuries. In Aragon and Catalonia, Scandinavia and much of Germany, the consent of the assembly was required to legitimize such enactments; custom produced much the same effect in England, and the Aragonese kings of Sicily (who replaced the Angevins there during the War of the Sicilian Vespers, 1282–1302) proclaimed the same principle. Yet more frequently they were assembled to assent to taxation, sometimes, though never wholly, saving the prince the difficulty of negotiating with each town or community for a contribution to the common need; they continued to constitute also the most weighty public court for the hearing of great causes. Though summoned essentially to consent to decisions made by the king and his inner council, such assemblies necessarily possessed a potential power to refuse or demand a precedent redress of grievances; and as the powers of central government became greater, efforts to impose restraints on the prince in the name of the community became more widespread. Such restraints had long been recognized in theory, but the rise of these larger assemblies provided a focus for such claims that was more effective than the earlier courts of feudal tenants. Precocious efforts to subordinate royal government to the scrutiny of the magnates in Parliament, and even to require general assent to the appointment of officers of state, occurred in England between 1258 and 1265 and continued with some formal success in the 14th and 15th centuries. It was a successful struggle in the 14th century to require parliamentary consent to all extraordinary taxation that ensured the regular summons of such gatherings and made possible their use as an occasional forum of political discontent or even revolution, as in the reign of Richard II.

The States General in France made similar efforts at political control of the kingship in the crises associated with the capture of King John, between 1356 and 1358 and again in 1413. The Cortes of Spain secured an even larger measure of success; in 1287 the magnates extracted from Alfonso III of Aragon the Privilegio de la Unión, which conferred upon the Cortes even the power to depose an unjust king; here, too, assent to taxation was a prerogative of the assembly, though grants were made only by the burgesses. The weakness of these estates were very like those of the 15th-century councils of the church in their conflicts with the pope. Essentially they were occasions, rather than permanent bodies, which were either summoned by the king or—if gathered by any other means—were unrepresentative. Their control over the monarchy lasted only so long as they were in session, and exhaustion or particularism frustrated their long continuance. The kings could normally exploit either the localism or the diverging class interests of their estates to prevent any continuous supervision. In Catalonia in the 15th century, the Cortes possessed a Diputaçio del General, a standing committee to watch over the government, which not only granted but collected and spent any extraordinary taxation; but the forms of taxation divided the burgesses from the knights and nobles. In England there was less division of class interest but no effective or continuous supervision. In all cases, the resistance of the magnates was a precondition of large claims on behalf of the estates; yet a prolonged and successful magnate resistance destroyed that fusion of local and class interests upon which the estates depended for their vitality.

Law and legislation. The Germanic invaders who occupied the highest offices of post-Roman Europe were for the most part incapable of conceiving of legislation in the

Customary law

classical or modern sense. The law by which all human affairs were regulated was that of the ancestors preserved in custom, an eternal system of right relations. Good law was ancient law, and new law was bad. The function of the legislator, therefore, was to proclaim or interpret the law, not to change it or make it; the laws of King Alfred, which open with the Ten Commandments and continue with a compilation of earlier law, are wholly characteristic. The codes of customary law enacted in the mid-14th century for Poland by Casimir III or for Sweden by Magnus Eriksson about 1350 show the long continuance of such traditions, though these later codes—and, indeed, all written customs for a kingdom—in some sense represented an attack on more local custom. The administration of the law, however, was recognized as mutable at all times; the Merovingians occasionally and the Carolingians quite frequently issued ordinances for the Frankish empire, as did Anglo-Saxon and Danish kings in England. The disintegration of political authority between the 10th and 12th centuries made legislation of wide scope very rare indeed; it was the rise of centralized institutions to enforce one law, written or customary, and of large assemblies in which to publish it that later allowed the multiplication of statutes and edicts of general application.

Outside the towns, where the enforcement of the law was unambiguously the function of the community and its officers, the authority behind such legislation was inevitably that of the king. The earliest general edicts of the post-Carolingian period were therefore royal proclamations; the first extensive example was the *Liber Augustalis*, promulgated by Frederick II for the Kingdom of Sicily in 1231 as the Constitutions of Melfi. Only slowly, and with uneven success, did the view gain ground that such proclamations must be made in large assemblies of fixed or even representative constitution. Where the general edict was a recognition of the limitations of royal authority, as in the celebrated Great Charter (Magna Carta) of King John of England in 1215 or the Golden Bull of Charles IV in 1356, it was in form a royal grant. As a consequence, general legislation, though increasingly necessary in later medieval kingdoms, had formal weaknesses as a limitation upon royal authority that were usually fatal. It was the legal autonomy of the princes, nobility, and towns that set a limit to royal authority, not "constitutional" edicts in the name of the community. Of these the king alone could be the author.

The growth of royal justice

The need for general legislation arose largely from the great expansion of judicial business done by the king and his courts. The right and duty to do justice was so fundamental an element in early conceptions of government that almost every aspect of political authority was exercised in judicial form. The local organs of taxation and military recruitment were the courts of count, prince, or petty baron, to which freemen came to settle their differences; the English exchequer was, both in form and in practice, rather a court than a counting house; the meetings of the estates as Parliament, Reichstag, or Cortes were courts before they were representative assemblies. Similarly, a centralized judicial system involved a centralized government and also a wealthy one.

Early royal government seems to have had little specific legal competence; the king's duty to see justice done was fulfilled by the appointment of honest judges rather than by hearing frequent appeals. Where custom was both decisive and localized, central courts had little attraction; where the modes of proof—by oath, ordeal, or combat—were intended to invoke divine rather than human judgment, there was little room for a theory of appeal; where travel was slow and literacy rare, centralized justice was extremely difficult to enforce. Therefore, the first technique available to the king who sought to enforce his law lay in the dispatch of trusted agents from his court to preside at local assemblies and to reduce the autonomy of counts or royal bailiffs. Under the Carolingians, special *missi*, often churchmen of high rank, performed this duty; in England the itinerant justices formed the model for the earliest French *baillis* in the 12th century. Under the French king Louis IX, in the 13th century, friars were often employed as *enquêteurs* to scrutinize local government and amend the faults of local officers; under Louis's successors the scrutiny was rather of their devotion to the king's interest.

As early as the 12th century such devices were already giving a new importance to the king's central court, for matters of difficulty were often referred by the travelling judges to the king's own hearing; the existence of a body of such judges provided a volume of general legal experience at the centre and a certain degree of consistency of procedure in the provinces that pointed the way toward a single royal law and enhanced the value of a central judgment. At the same time, the king's court as the unique tribunal for disputes between his great tenants and as the last resort of injured litigants was attracting a greater volume of business. Although the ordeals and judicial combat never vanished entirely as methods of proof, they lost ground rapidly in western Europe in the 13th century to the procedures of written evidence and judicial enquiry; this made the theory of appeal easier to accept, and the practice was accelerated by two fundamental characteristics of medieval litigation. The first was that it was an absolute duty to pursue possible rights to their limit; this was pre-eminently true of religious communities that never died and never forgot, for whom failure before the law meant despoiling the saint to whom their church and estates were dedicated, but a secular family was scarcely more forgetful. The second was that judgments, to be effective, had to be accepted by both parties; the object of the court was to secure agreement and the voluntary submission of the unsuccessful litigant. In a society largely devoid of specific offices of enforcement, where those who appeared before the courts were often neighbours, this insistence upon the acceptability of judgments was essential. It made all the more valuable the judgment of a tribunal whose authority could not be refused. Hence, there emerged in all the larger states a variety of central courts.

The French *parlement*

In France the regular gatherings at the king's court produced the distinctive institution of the *parlement*, essentially an omnicompetent supreme court that began in the 13th century as a committee of the king's councillors with some expert legal assessors but that by the mid-14th century had become a fully organized and almost wholly professional body with three principal organs—the Grand Chambre, in which the decisions were pronounced; a Chambre des Requêtes, which scrutinized appeals and distributed them to the appropriate tribunals; and a Chambre des Enquêtes, which carried out the judicial enquiries upon which the final judgment rested. From its position as supreme court, the *parlement* came to exercise a wide supervisory power even over the Chambre des Comptes; in this it continued to resemble the English Parliament as it had in its origins. Unlike the English institution, however, its nonprofessional element tended to diminish, and it achieved a substantial measure of independence in the appointment of its officers. But, having no control over financial supply and no representative or political basis, it was rarely capable of serious resistance to royal autocracy. With the collapse of English and Burgundian power in France in the late 15th century, a number of local *parlements* were created to serve the reunited territories, the earliest being that of Toulouse; in these, however, the crown's right of nomination prevented any serious threat to the legal unity of the realm.

In England the two courts of Common Pleas and King's Bench came to operate as supreme courts, especially for civil cases, while the royal monopoly over serious offenses was chiefly exercised by the itinerant justices of assize; over all these stood the high court of Parliament, which the litigant could petition for redress. The judicial authority of the king and the king's council offered the litigant a further source of appeal.

In Germany political fragmentation prevented the evolution of a law unified by administrative practice, so that the need for a common code was reflected in the formal Reception of Roman Civil Law in 1495, a measure without immediate or widespread practical consequences.

Throughout the medieval period, central courts tended to attract an ever larger proportion of litigation from the

local courts. As a court of appeal, the French *parlement* slowly eroded the importance of the local Grands Jours (the solemn sessions of the courts of the earlier great lordships), except in those cases in which these were themselves transformed into *parlements* at the end of the period. In England police jurisdiction passed into the hands of justices of the peace, unpaid crown officials appointed by the crown from among the local landowners; all important civil litigation was conducted in London, leaving the old shire courts, once omnicompetent, only a shadowy existence. The letters of the Paston family show a group of minor Norfolk gentry engaged in a long series of lawsuits that kept the head of the family almost constantly in London as well as employing a considerable number of the professional lawyers who were so characteristic a product of the 14th and 15th centuries, both in France and England. These lawyers were among the earliest literate laity, the first trained secular administrators, and the chief agents of the unification of postmedieval states (see further EUROPEAN LAW, MEDIEVAL).

Revenue and taxation. The whole fabric of the centralized state depended on the mobilization of great resources for the crown. Even the earliest rulers of the Heroic Age had needed a hoard from which to reward their followers and to draw the gifts upon which relations with neighbouring chieftains depended. Such hoards were accumulated by successful plundering and the tribute of subject peoples; they were themselves the great prize of war, as when Charlemagne's forces sacked the Avars' Ring (chief fortress) in 795/796, and the treasure was carried across Europe by the cartload. The rich furniture buried with the ship of a petty king in East Anglia at Sutton Hoo in the 7th century shows how magnificent and how diverse these hoards must have been. With the stabilization of settlement between the 9th and 12th centuries, the pursuit of plunder and tribute became a more marginal enterprise, still actively pursued by the Vikings and the Magyars (and later, in the 9th and 10th centuries, by the German settlers along the Eastern frontier) but rarely available to most princes of post-Carolingian Europe. By this time, the fragmentary system of Roman taxation had also broken down, so that the central or royal revenues that existed were mostly those open to any great landowner. For many rulers these resources were still largely in food rents, which the owners travelled about their estates to consume where they were produced. Throughout the medieval period these revenues from the king's own estates were an essential element of his power; the extent and distribution of the royal domain were touchstones of his authority.

The rise of a money economy

Beyond the revenue in kind produced on the king's estates or collected by his bailiffs, supplemented by a proportion of the trappings of his dead followers and any quasi-judicial fines that could be seized to his use, the king's revenue lay in services, such as his right to summon the army or to secure labour for his fortresses, cartage for his crops, or suitors for his courts. All these services and many of his food rents needed constant vigilance to enforce or secure; no prolonged expansion of central authority could rest on this anarchic foundation.

Such primitive sources of supply were the consequence of the almost complete collapse of a system of money exchanges across Europe, which was itself a consequence both of economic regression and perhaps a shortage of bullion; until the opening up of the central European silver mines on a large scale in the later Middle Ages, Europe had few supplies of ore for coinage. The monetary chaos of the 7th and 8th centuries was in part relieved by Charlemagne's introduction of the first medieval European silver coinage of commercial use, his pound of 240 pennies (deniers), coins of excellent quality and weight. Yet the multiplication of local mints and unstable values made even the abundant silver coinage of the 14th century awkward to use and offered innumerable opportunities for malpractice—from coining on a small scale by petty forgers to large but undisclosed devaluations by princes.

For all its inconveniences, however, the use of money to replace food rents or services rapidly became widespread as soon as the circulation of coinage and the evolution of a market made it possible. In the 10th century the great domains of Italy were already being leased out; and by the end of the 14th century rent rather than labour was the characteristic peasant due everywhere in Latin Europe except in parts of Germany and Central Europe. Similarly, it became increasingly common to commute the obligations of free tenants and feudal vassals to money payments; the very term aid (*auxilium*), originally meaning simply the help a man owed his lord, had by 1200 come to mean a tax. The king could in theory impose what dues he liked, as could any other lord; these arbitrary tallages were of considerable importance in areas where, as in Capetian France, there were large towns on the domain to contribute to it. The exploitation of the forests also contributed in fines and licenses; market tolls and even a primitive form of sales tax all went to swell the domain revenue. Though these profits could all be commuted and so more readily applied to the common needs of the kingdom, they still formed an extraordinarily complex and miscellaneous collection of rights of fluctuating value, which was far beyond the capacity of any central accounting system to assess or collect in detail. Hence, the importance of the local revenue collectors, stewards of the royal domain, sheriffs, provosts, vicars, and so forth; in general, the practice was to allow these officers to collect the normal profits of lordship for themselves in return for a fixed payment, accounting in detail only for such extraordinary revenue as heavy judicial fines or tallages. Fixed incomes of this kind suffered through a fairly marked inflation over much of 13th-century Europe.

Taxation

Rising costs of government in the 12th and 13th centuries required new and more comprehensive taxes little connected either with the ownership of land or the exploitation of vassals. The church, with its tithes, had long collected a form of income tax; under Innocent III the so-called charitable subsidy was imposed on the whole church by the papacy. Lay princes first employed a similar principle to finance the Crusades of the later 12th century by imposing a tax on the movable goods of their subjects. By the end of the 13th century such taxes were frequent. In many respects they were the first manifestation of an act of sovereign government in Europe for centuries, being applicable to all subjects upon a uniform principle—although in practice the collectors often struck bargains with whole communities rather than pursue the complexities of universal assessment. The fifteenth and thirtieth granted by the English Parliament at regular intervals from the reign of Edward I, were also employed occasionally in France, where a combination of a hearth tax and a variety of forms of sales tax (such as the notorious gabelle, a salt tax) became the most characteristic and permanent forms of national taxation. Customs dues, particularly on comparatively valuable and portable goods such as wool, wine, and cloth, also became widespread in the 13th century; it was at this time that elements of a genuine commercial policy first appeared in the conduct of princes, so that merchants began to secure legal and political rights as a class (as opposed to local privileges).

Loans from merchants and Jews

These merchants owed their political influence very largely to the possession of large sums of cash or increasingly trustworthy systems of credit. Kings, and lesser rulers too, constantly found that their need for money was urgent but that their means of realizing it were slow. It took several years for a grant of a subsidy by subjects to be assessed and collected, but no army or even household could go without wages so long. Until the late 12th century the sources of available credit were chiefly the Jews, exempt from the prohibitions of the church against usury, and goldsmiths. The Jews, generally prevented by law from holding land, were very widely found in the towns, living sometimes under the protection of the church and often under that of the king, who exacted heavy sums in return for his protection and sometimes assumed the rights of a Jewish creditor himself. The life of the Jewish community was extremely precarious, for the Jews were constantly subject to the civil disabilities

imposed by the church and were the most frequent victims of a wanton crusading zeal and sporadic violence; they found in kings harsh and inconsistent protectors who not infrequently expelled them from their lands as a measure of extortion or political appeasement.

In the 12th century the order of Knights Templar, whose widespread estates were intended to provide revenue for the Crusades to the Holy Land, acted as the first large-scale Christian bankers. By the 13th century, particularly, secular bankers began to appear in such Italian communes as Piacenza, Siena, and Lucca; ironically, many of them gained their experience by collecting papal revenues. The bankers often lent money to the Italian princes, who secured their loans by granting the bankers the right to collect customs revenue. The whole revenue system of the Angevin kingdom of Naples was controlled by the Florentine bankers who had financed the original conquest. Politically, this type of loan was widely suspect; and there were widespread failures of such enterprises in the years after 1350, as was also the case with the much humbler money changers north of the Alps, the hated Lombards or Cahorsins. In the 15th century a certain number of great banking houses north of the Alps appeared to rival the Medici of Florence; especially notable were the royal merchant Jacques Coeur in France and the later Fuggers of Augsburg. With the decline of the Italian bankers, princes were compelled to fall back upon their own merchants, and these became subject to frequent forced and interest-free loans, disguised under such a term as "benevolence." Only the grant of monopolies or freedom from other taxes could make such exactions tolerable, but with the aid of these the prince could maintain himself against the greater feudatories, whose particularism threatened the interests of the towns and king.

Military service. Changing patterns of taxation were closely related to changes in forms of warfare and of royal right to military service. The earliest war-band leaders were the leaders of an army that comprised, it seems, all free men of military age; their annual gathering could be regarded either as a military review or a public assembly. The special importance of a more heavily equipped and mounted warrior, already manifest in the 8th century and developing steadily thereafter, reduced the numbers of such warriors who could be found and altered their relations to the crown; it was as vassals, rather than as subjects, that they answered the king's call. The older and wider obligation on all men not of servile condition died out rapidly in the 11th and 12th centuries, except as a duty to provide labour, the most lightly equipped garrisons for the large stone castles now so widely present, or as a vague and largely unenforceable obligation to form a local militia in time of invasion. Only in the Italian towns was a public obligation to serve in the armies of the republic widespread and effective. Correspondingly, service with the army became attached to property, with the usual consequences. The vicissitudes of property made the gathering of an army difficult, and custom came to limit the length of service ever more narrowly until 40 days became the usual extent of obligation. Moreover, outside England the quotas of knights provided by the greater vassals for their king were usually far below his needs or their capacities. Under these conditions, wages early appeared as an extra inducement to service: the king would often contract with a magnate to secure a fixed number of knights, sergeants, and horses for a lump sum, which the noble or experienced contractor would then redistribute. By the 14th century this system was widespread; its effects in peacetime were often grave, but it did provide an alternative for the failures of the tenurial system.

Increasing use of mercenary soldiers

At the same time, the king's duty to provide for the safety of the realm allowed the use of a more universal obligation. In England all men were required to bear the arms appropriate to their rank and to practice their use; special commissioners of array then chose men to be paid in the king's service as archers or sergeants. In France the Valois kings required the community to provide contingents of troops but in practice normally used this as a means of raising money instead. A serious failing of such systems of raising forces was that at a war's end a large body of unemployed troops was unleashed on the countryside. France, Alsace, and Italy had as much to fear from the casual ravages of unemployed mercenaries as they had from the deliberate assault of an enemy. At the very end of the period the French kings, with their *gens d'armes*, created by an ordinance of 1445, laid the foundation of a permanent standing army that the emperor Maximilian sought to emulate. The European medieval aristocracy in its most characteristic form had been a military one, secure within its crowded fortresses and rendered formidable by its near monopoly of the essential arms of warfare, the mounted and fully equipped knights who were the great men's vassals. By 1500, however, the connection between noble birth, military prowess, and the basic composition of an army was fast breaking down, even before gunpowder hastened the last of the fully armoured knights from the battlefield.

IV. The world of the senses and the mind

The communities of the early Middle Ages lived on intimate terms with a largely hostile environment. Storms on sea and land, floods, pestilences, and famine were constant hazards. Even in the more densely settled parts of Europe, impenetrable forest and trackless fen covered large areas. These areas were often believed to be the haunts of demons and, in fact, provided refuge for brigands and outlaws as well as wolves and wild beasts. The uncultivated moorlands and mountain passes were safely crossed only in haste and with company.

Pagan religious survivals

The religious beliefs of the pagan tribes were appropriate to this environment of apparently random disaster. The gods of the Germanic tribes were wanton and crafty, and fertility cults were probably widespread though intensely local. Formal belief in Odin, Thor, or Freyja gave way rapidly before the missionary fervour of Arian and Roman missionaries, but many pagan elements passed into the superficial Christianity of the first converts. Devotion to local saints was often (even deliberately) based upon earlier pagan cults, while churches were built on the sites of temples or sacred groves. Early miracle stories often attribute markedly pagan qualities of willful jealousy or capricious good will to the saints. The mere possession of such wonder-working relics as the Holy Lance of the empire was supposed to confer a title to the crown and a formidable military authority to its holder. There was a brisk traffic in such objects among princes, while the sacred groves and wells of the earlier religion rapidly took on the name of some local saint for the peasant population. Augury, the sacrifice of cattle, and a host of other pagan rites continued throughout the period, though the ecclesiastical police system was more and more successful in sharpening the distinction between acceptable devotion and witchcraft.

Christian life at the parish level necessarily reflected such conditions. In a literate society it was the ceremonial performance of the sacraments that was of paramount importance; indeed, preaching was a late and occasional addition to the duties of the parish clergy. The blessing of harvests and houses, the averting of plague, fire, and invasion by regular formal intercession were the essential functions of the parish clergy. Correspondingly, communal disaster and outbreaks of anti-clericalism were closely connected. For the salvation of the individual the church required annual confession and very occasional taking of the sacrament. Private confession and private penance became the rule by 1100, but penance might be done by extensive travelling. By the 12th century, pilgrimages were made to the great international shrines of Jerusalem, Rome, Santiago de Compostela, or even St. Thomas of Canterbury, as well as to a host of lesser local shrines, insuring against the future or expiating past crimes. The Crusades, in part intended to restore the pilgrim road to Jerusalem, were themselves a means of securing full remission of sins for the participants. Their frequent failure or diversion to trivial or political ends made them a dying force after 1300, but the need for such concrete means of salvation continued (see CRUSADES). Papal jubilees and the organization of a system of indulgences

secured by generous alms met this need, though offering many opportunities for abuse.

Until the 12th century, when Manichaeism spread through Italy and southern France, the orthodox church in the West was notably free of the doctrinal divisions that rent the Eastern Church. Eradicated by prolonged and bloody war, this heretical view had little later influence, but the source of its success—the wealth and pastoral inertia of the secular church—was to produce a multitude of dissident movements that laid stress either upon the defects of the propertied clergy or upon the direct illumination of the individual, notably by the reading of the Scriptures; already in the 12th century the Waldenses had adopted such views. In the late 14th century, John Wycliffe in England preached individual salvation and criticized the whole fabric of the church, sacramental and hierarchic. Wycliffe's doctrines met with little response in England but provided the starting point for the Hussite movement in Bohemia, where economic, political, and doctrinal revolution united to threaten the whole social and ecclesiastical order of eastern Europe.

In part this multiplication of dissent drew on the anticlericalism of the more educated laity and on wider knowledge of the Scriptures in vernacular versions, but it also coincided with the end of the period of monastic reform and development; while the fervour of St. Francis produced a new order to perform a vital function within the church, the devotional tendencies of the 14th and 15th centuries associated with the names of Meister Eckehart or Thomas à Kempis were largely devoid of institutional consequences. At the opposite extreme of this interior devotion was the proliferation of extravagant sects such as the recurring appearance of the flagellants in the years after the Black Death and its successors ravaged Europe. The prophetic form of much of the Scriptures, the frequency of disasters that appeared, at least locally, to portend the collapse of human society, and the general belief that the centuries after Christ's coming represented the last age of mankind produced a steady trickle of sects believing in the imminence of judgment or the new Jerusalem, encouraged by an easy distortion of the view of such academic prophets as Joachim of Fiore.

Communal organization

Against the hostility of their world, seen or unseen, medieval societies also fell back upon a wide variety of communities vowed to mutual support. The minimal privacy of the characteristic medieval household gave all houses a communal quality. It was common for several generations to live, eat, and sleep under one roof, which often enough covered the livestock too. The earliest forms of Germanic settlement and organization rested upon the kin group, the early codes all supposing that the kin had absolute responsibility for its members. In the unstable society of the early Middle Ages, the bonds of commendation by which men bound themselves to a protector provided a substitute for the security and social cohesiveness of the earlier kin group. The medieval impulse toward the formation of communities bound together by oaths may be seen also in the trading guilds of the communes, in the councils of kings and princes, and even in the societies of rebels or robber bands.

The tightly knit character of most medieval communities was a result of the relative immobility of the population. While the church and warfare were international occupations, in which a man might serve from one end of Europe to the other, and though merchants travelled with their wares across a multitude of frontiers, the bulk of the agricultural population rarely journeyed more than a few miles from their village, and many townsmen were equally confined. News travelled slowly and inaccurately, borne by pilgrims, peddlers, bailiffs, and beggars; marriage outside the village was unusual; and local dialects sharply contrasted. On the edge of this society, however, some adventurous travellers were covering huge distances. The Vikings sailed south to the Mediterranean, east and south to the Black Sea, and west to Iceland, Greenland, and the coast of North America. In the 13th century, the Venetian traveller Marco Polo crossed the breadth of Asia to China, where several Franciscan missionaries were to follow. In the 15th century, Portuguese seamen groped their way south along the African coast in pursuit of gold and a sea route to the Indies, rounding the Cape of Good Hope in 1488. In 1492 a Genoese seaman, in the service of Ferdinand and Isabella of Spain, crossed the Atlantic and returned; remarkable though Columbus' voyage was, the rapidity with which its significance was assessed and exploited provided proof of how much the European view of the world was changing.

Technological advances

Primitive though the vessels and instruments of these navigators were, they showed a marked advance over the equipment of their predecessors, an advance that was widespread and accelerating in the 14th and 15th centuries. From the 9th century on, water power was extensively harnessed to mill flour, replacing the infinitely laborious querns of past millennia; it was also at work driving the hammers of the ironworkers and the first fulling mills, which revolutionized the social patterns of cloth working. In the late 12th century, windmills also began to be used widely, while at about the same time changes in the forms of harness allowed horses to take over some of the ploughing and carting formerly performed by oxen. In the 13th century, mining techniques for the first time allowed the driving of deep but drained shafts beneath the surface to the richer iron, copper, tin, and lead of Bohemia, Sweden, or Cornwall. By 1500 the demand for charcoal for smelting and timber for shipbuilding was already pressing hard upon the once inexhaustible forest of an older Europe.

As a consequence of these changes, the use of iron became common, even in the implements and houses of the poor. Metal cooking pots, glass bottles, and glazed wheel-turned pottery came into widespread use. Houses of stone or even brick often now had the chimneys and glass windows formerly found only in palaces. Domestic furniture of these centuries survives to show that it was no longer confined to crude benches, tables, and chests but was more carefully fitted and elaborately decorated. In the 15th century the former cloth-working town of Arras expanded its manufacture of the tapestry hangings for which it would be long famous, while silk, linen, and cotton came into wider use. Sumptuary laws showed how much wider a range of society was able to dress in the materials and fashions that had once been the hallmark of the highest rank.

Architecture

The earliest surviving medieval architecture is ecclesiastical; outside the area of continuing Byzantine occupation, these are small churches sometimes built out of the fragments of former Roman buildings and always derivative in style. Some of the greater Carolingian churches, however, were more ambitious in scale and conception, and the surviving 11th- and 12th-century Romanesque churches show the emergence of a wholly distinctive architectural convention in enormous and complex buildings. Gothic emerged in early-12th-century France and became the dominant style of the Latin West. In Italy the success of this style came late and was short-lived, for it was overtaken by the revived classical models before it had long taken root; but over much of northern Europe the Gothic survived into the 20th century.

The earliest surviving medieval domestic architecture is found in a few stone houses of the towns and in the modifications of fortresses to serve more domestic purposes. By the 13th century the stone palaces of princes were constructed with the skills first practiced in churches. Such planned towns as Aigues-Mortes foreshadowed the elaborate fortifications of many towns of the 15th century, within which numerous multistoried houses of wealthier merchants survive to the present. The single chamber of the lord's hall or the peasant's hut was increasingly giving way to the house divided into rooms, though in much of rural Europe there was little to distinguish the cabins of the poor of 1500 from those of a millennium earlier.

Recreation

The recreations of most ranks of society are poorly attested before the later Middle Ages. Among great men and warriors, the life of the hall and forest was pre-eminent. In the hall the long eating and heavy drinking of the lord's followers were favourite motifs of the heroic poetry once sung by household or itinerant poets, and feasts were an essential focus of social life. Coronations, wed-

dings, and funerals were all celebrated with banquets. Pagan customs of punctuating the year with such events were either contested or taken over by the church; the gathering of the harvest and the celebration of Christmas were regular occasions of the kind, and from time to time the enthronement of a bishop produced a banquet of extraordinary splendour. The earliest texts stress the quantity (rather than the quality) of food and drink, but the more detailed later medieval records show that scarce or exotic foods and an extreme elaboration in their preparation were becoming the marks of ceremonial extravagance, the highest achievement of the pastrycook being such "subtleties" as the Holy Trinity surrounded by choirs of angels. Eating, drinking, and singing were supplemented by various forms of gambling, to which the Germanic peoples were so addicted that men were known to stake their own liberty on the fall of the dice and so condemn their descendants to slavery. Backgammon, chess, and (late in the period) card games were played by those with leisure and means.

The principal recreations of the nobility in the countryside were hunting and fighting, both of which were originally conducted in an extremely dangerous way. The main beasts chased were the deer, boar, wolf, and bear, with rabbits, hares, and foxes serving as a lesser challenge; all were usually pursued on horses with packs of hounds, the variety and qualities of which were discussed in numerous later medieval treatises. The support of the chase was an important economic and legal institution. The rearing of the lord's dogs and other services for his hunting were conditions of some tenures, and the preservation of hunting rights was a source of revenue and a cause of oppression. Another honourable sport was falconry. There was an appropriate bird for each rank of society, and in this sport women might participate more often than in the hunt.

The holding of sham battles for enjoyment and exercise became general in the 12th century, when the ransom of prisoners was an accepted means of support for skilled but landless knights. By the end of the century, the practice was so widespread as to attract the condemnation of the church and to represent, it was supposed, a threat to the stability of the state, since gatherings of armed men could readily become rebellions. In later centuries, this generalized melee gave way increasingly to highly formalized jousting between individual knights wearing armour especially adapted for a variety of possible rules of combat. In the 14th century this was still a possible means of capture, ransom, or death; by the end of the 15th it was almost entirely a sport.

For the less wealthy, the available entertainments were rarer and less elaborate but not perhaps much safer. Football, wrestling, or fighting with staves, regulated by few conventions, produced a heavy casualty rate, as did brawling in innumerable unregulated alehouses. Brewing was a cottage industry frequently engaged in by women, who sold their product subject to a seignorial right to examine the quality of the drink. In a society in which sugar was expensive even for kings and in which honey was prized but scarce, the thick beer and mead of the north was not only a focus of recreation but also an essential element of diet, as was the wine of the south.

Medieval music

Among the earliest of recorded arts was that of song, and a substantial body of medieval music has survived. The best known is the ecclesiastical plainsong, whose origins are attributed to the liturgical reforms of Pope Gregory I, which was elaborated into the complex polyphony of the 14th and 15th centuries, when for the first time the names of individual composers are recorded. The songs of troubadours and minnesingers of the 11th and 12th century show the rise of a sophisticated secular music that was increasingly accompanied by a variety of musical instruments. Well before 1500 the playing of music as an autonomous and often purely secular art was already established.

The decorative arts of painting and sculpture had undergone a similar progress, but here the revived interest in the secular products of classical antiquity and extensive lay patronage may have hastened the emergence of such arts from a wholly ecclesiastical setting to produce the flourishing schools of Italian and Flemish painting and the extensive sculpture of 15th-century Italy. In both cases the remarkable triumphs of the earlier period, the mosaics of Ravenna or Norman Sicily, for instance, or the sequences of sculpture across the west face of Autun or Laon were subordinate to a larger ecclesiastical unity —the liturgical drama of the church; the sculpture of Donatello and the painting of Botticelli were not, whatever the piety of their creators.

The rise of the professional painter of pictures to be considered as a creator in his own right is largely matched by the decline of the illumination of manuscripts, which reached a peak of elaboration in the mid-15th century just as the earliest printed books were heralding the end of the role of the handcopied manuscript.

The rise of printing

Printing on wooden blocks spread rapidly in the second half of the 15th century, and by 1500 presses were at work in every major country in Europe. Though the earliest printed books were extremely expensive and their purpose frequently liturgical, their potential importance was great because they could cater to the new market of the educated laity. These now included far more than the circle of great men who commissioned the splendid psalters and books of hours of the 14th and 15th centuries, which had overthrown a monastic monopoly of lavish manuscripts that had prevailed since the time of Cassiodorus and his copyists at Vivarium in the 6th century.

Education

The institutions and the subject matter of education had already proceeded according to a similar rhythm. The only schools of Europe in the 8th century, except perhaps in parts of northern Italy, were those attached often to monasteries and more rarely to bishoprics. Independent thought, even in theology, was extremely rare; theology remained under the long shadow of St. Augustine of Hippo, and the Greek learning and originality of John Scotus Erigena in the 9th century was little pursued. The rise of the new skills in dialectic in the 11th and 12th centuries produced two phenomena: first, a confidence in rational thought as a means of solving problems, especially those raised by the conflict of authorities; and, second, a number of teachers whose exceptional talents attracted scholars from the farthest ends of Europe. The self-confidence and European reputation of Abelard reveal this movement at its most distinctive. Around such teachers grew up either religious communities such as that of Saint-Victor of Paris or the earliest universities. In the 12th century the lawyers of Bologna, the doctors of Salerno, and, above all, the theologians of Paris were becoming organized bodies governed by a chancellor; by the 13th century the universities possessed their own statues regulating the arduous courses of study toward recognized degrees. The crown of studies was the pursuit of the highest knowledge, theology. The forms of 13th-century university study gave rise to the characteristic theological achievements of the period, the *summae* of the Dominican St. Thomas Aquinas and the Franciscan St. Bonaventure. The founding of universities received a new impetus at the end of the 14th and 15th centuries, when they spread into Scandinavia, Scotland, and eastern Europe. It was in these years that most secured a large independence from external ecclesiastical government, and in the councils of Constance and Basel the universities claimed a position of the highest authority.

Much of the earlier confidence in the capacity of intellectual endeavour according to the established forms of enquiry drained away in the 15th century. Logicians in the tradition of Duns Scotus and William of Ockham asserted the essential disparity of faith and reason; the canon law proved incapable of resolving the most pressing problems of ecclesiastical authority or of securing effective reforms; and the only literary forms that offered novelty and room for growth were the vernacular literature of the court and the classical poetry of the Humanists. Whatever changes occurred in education were generated not in the universities but in the schools of such Italians as Vittorino da Feltre near Mantua or of the Brethren of the Common Life in the Netherlands. These first insisted upon the effects of education, on the whole

personality, where the numerous grammar, guild, and charitable schools had provided only a grounding in the mechanics of literacy. The retreat of the monopoly of the church in education stands beside the work of the explorers and the rise of the absolute monarchies as an important mark of the ending of medieval society.

BIBLIOGRAPHY. The fullest account of medieval Europe in English is that of the *Cambridge Medieval History*, 8 vol. (1924–36; 2nd ed., 1966–), of which only the Byzantine volumes have been re-edited, as has the first volume of the companion *Cambridge Economic History of Europe* (1966). GEORGES DUBY, *L'Économie rurale et la vie des campagnes dans l'occident médiéval*, 2 vol. (1962; Eng. trans., *Rural Economy and Country Life in the Medieval West*, 1968), summarizes much recent work and shows well the directions of current research; FRANCOIS L. GANSHOF, *Qu'est-ce que la féodalité?*, 2nd ed. (1947; Eng. trans., *Feudalism*, 1952), is the classic analysis of the structures of feudal tenure. FRITZ KERN, *Gottesgnadentum und Widerstandsrecht im früheren mittelalter* (1914; Eng. trans., *Kingship and Law in the Middle Ages*, 1939, reprinted 1968), is the most convenient introduction to its subject. *The Handbook of Church History*, ed. by HUBERT JEDIN and JOHN DOLAN (1965–), gives a full and recent account of the period from the 8th to the 12th centuries; for the later period, A.C. FLICK, *The Decline of the Medieval Church* (1930), is still useful; R.W. SOUTHERN, *Western Society and the Church in the Middle Ages* (1970), gives a more general introduction. On the towns HENRI PIRENNE, *Medieval Cities* (1925, reprinted 1956), has lost little of its value; DANIEL P. WALEY, *The Italian City-Republics* (1969), covers the period between the 11th and the 14th centuries in greater detail. GEOFFREY BARRACLOUGH (trans.), *Medieval Germany, 911–1250: Essays by German Historians* (1961); JOHANNES HALLER, *Die Epochen der deutschen Geschichte* (1928; Eng. trans., *The Epochs of German History*, 1930); ROBERT FAWTIER, *Les Capétiens et la France* (1942; Eng. trans., *The Capetian Kings of France*, 1960); and EDOUARD PERROY, *La Guerre de cent ans* (1945; Eng. trans., *The Hundred Years War*, 1951), give useful examples of national history. On England the first 5 volumes of the *Oxford History of England* (1937–59), are authoritative. JOHAN HUIZINGA, *Herfsttij der Middeleeuwen* (1950; Eng. trans., *The Waning of the Middle Ages*, 1954), is a controversial classic. The largest and most representative collection of medieval texts in translation is that in the Columbia "Records of Civilization" series, supplemented by that of the *Oxford* (formerly *Nelson*) *Medieval Texts*. The standard English bibliography of medieval history is L.J. PAETOW, *A Guide to the Study of Medieval History*, rev. ed. (1931, reprinted 1964).

(Ma.Br.)

Middle American Peoples and Cultures

The physical spine of Middle America is the broad mountain chain extending from the southern end of the Rockies to the northern tip of the Andes, with Middle America in the area from northern Mexico to Nicaragua. The mountain chain marks off the area into four major regions. The heartland of Middle America is the central valley of Mexico. A second region is the highlands along the southern Pacific slope of Mexico. Beyond the Isthmus of Tehuantepec are the southeastern highlands in the Mexican state of Chiapas and in Guatemala. The arid region in the northwest of Mexico is a fourth region.

Within these four major geophysical regions there is tremendous variety in ecology, climate, soil, and the possibilities of human life. The mountains crumple the face of the land into a multitude of valleys and micro-environments; the result is a mosaic of crops, peoples, and settlements about which it is difficult to generalize. The high valleys of central Mexico, Oaxaca, Jalisco, and Guatemala have been the most densely settled parts of Middle America. But the lower slopes of mountains near the seacoasts have also carried substantial populations. The steamy tropics of the Isthmus of Tehuantepec and the hot limestone thumb of Yucatán have also been heavily populated.

The Indians of Middle America live almost everywhere in the region. The basic requirement for human settlement is water. The major river systems and the high valley lakes have been the primary settlement sites since prehistoric times.

CULTURAL PATTERNS

Cultural areas. While the social heritage of Middle America is highly complex, within the broad historical flow five separate cultural areas can be distinguished. They are regional configurations of the basic Middle American cultural patterns. One cultural area is that of the Maya. The southern, highland Maya were and are concentrated in western Guatemala and the state of Chiapas in Mexico. The northern Maya inhabited the Yucatán Peninsula of Mexico and the jungle of Petén in Guatemala. The Maya of these two regions form a continuous territorial and historical entity. (There are also contemporary Maya people in Veracruz and San Luis Potosí in Mexico, known as the Huastec.) The monumental ruins left by the pre-Columbian Maya are one of the puzzles of anthropology; theirs is the only civilization known to have flourished in a tropical rain forest.

The Mayan cultural area

The southern Mexican highlands and the adjacent coastal regions form a second cultural area within the basic Middle American pattern. This region covers most of the present Mexican states of Guerrero and Oaxaca, the southeastern part of Veracruz, and parts of Puebla and Morelos. Its highland people developed the traditions of the Mixtec and Zapotec, whose ruins survive at Mitla and Monte Albán, whereas the coastal people seem to have been somewhat isolated from them.

A third cultural area is the central Mexican highlands, including the valleys of Puebla, Toluca, and Morelos, along with the eastern slopes of the Mesa Central and parts of the Balsas River Basin. This area was the centre of the Aztec Empire. Mexico City is built on the ruins of the Aztec capital of Tenochtitlán, and descendants of the Aztecs still live in the area.

The mountain chain around the high lake of Pátzcuaro, in present-day Michoacán, forms another cultural area. The relative isolation created by the mountains permitted the Tarascans to work out their own cultural variant. They reached a level of social and political organization comparable to that of the Aztec and the Maya.

A fifth cultural area is northwest Mexico. This region is not historically or culturally a single unit; it exhibits three major types of ecology and three major types of human adaptation. The high mountains provided possibilities for simple agriculture without irrigation, whereas the desert required settlements around valley bottoms dependent on floodwaters in the rainy season. In the west the abundant shellfish were the basis for a coastal culture. The historical forces at work in the northwest area were different from those in the four other cultural areas. The population was taken over by the Jesuits, who built mission communities. The relative isolation of the region following the collapse of Spanish power, together with the weakness of the succeeding Mexican governments, permitted the survival of a blend of Spanish-Indian culture.

Language groups. Hundreds of languages were spoken in Middle America. Some linguists have grouped them in a number of phyla, or superfamilies, each phylum being at the same classificatory level as, say, Indo-European. The Hokaltecan superfamily includes the Yuman family (four surviving languages, two extinct); the Serian family (one surviving language, four extinct); the Coahuiltecan family (four extinct languages); the Tequistlatecan family, with one living language; the Supanecan family (one surviving language, two extinct); and the Jicaquean family, with one living language. A second phylum, Uto-Aztecan, comprises the Piman family (four surviving languages, eight extinct); the Taracahitian family (two surviving languages, 39 extinct); and the Aztecoidan family (three surviving languages, 18 extinct).

Attempts have been made in the past to associate the large family of Mayan languages (about 27 living languages) with several other Meso-American Indian languages (the Mixe-Zoque and Totonacan families, primarily) as the Macro-Mayan or Macro-Penutian languages; the latter term supposes an association also between the Mayan languages and the Penutian phylum of North American Indian languages. These groupings are not generally accepted by modern linguists.

The Oto-Manguean phylum includes the Oto-Pamean

family (six surviving languages, one extinct); the Chinantecan family (one living language); the Zapotecan family (two surviving languages, one of which, Zapotec, is so diversified that its many dialects constitute mutually unintelligible languages); the Mixtecan family (three living languages); the Popolocan family (four surviving languages, one extinct); the Chorotegan family (eight extinct languages); and the Amuzgo family (one living language).

In addition to the four phyla there are the Tarascan family of one living language; the Huavean family of one living language; and the Xinca-Lenca languages spoken in small enclaves in the southern part of the region. Finally, there are also 39 extinct languages that linguists have not been able to relate to each other or to any other (see MESO-AMERICAN INDIAN LANGUAGES).

Characteristics of Indian cultures. This linguistic variety shows how diverse were the cultures of the aboriginal population. Even though it is possible to generalize broadly about cultural resemblances or patterns, the actual cultural unit was local or regional. This is true despite the existence of larger political units, such as the Quiché of classical Guatemala, or the Aztec Empire. The broad characteristics of the Indians of Middle America may nevertheless be sketched in profile, if the caveat that the profile is everywhere varied by local circumstance is heeded. When the Spaniards conquered the Indians, they removed the indigenous ruling class and placed themselves at the apex of society. They also brought Catholicism, horses, cattle, wheels, iron, and new forms of political and economic organization. The Indian culture of today is a blend of indigenous elements, the culture of the Spanish, and the historical precipitate of the 500 years since the conquest.

Middle American cultures exist in located, named communities, each of which has a physical centre housing the patron saint of the community. The basis of subsistence is maize (corn), cultivated in small plots; but there is a myriad of crafts and artisans, and communities tend to be economically specialized. The family is the basic social unit, each living in a separate structure. Males and females wear distinctive costumes; where they have adopted modern dress, it is chiefly among the men rather than the women. Men do the heavy agricultural tasks and women the domestic chores. Men are in charge of the indigenous cults; women are more prominent in the Catholic aspects of religion. There is little restriction on any economic activity, and no taboos hedge occupational choice; but modern economic organization is little developed. The marketplace is the focus of economic life.

Families are internally hierarchical, with males and elders dominating. Both the father's and the mother's kin are recognized in tracing relatives, but there is a patrilineal emphasis in the transmitting of names and of some real property. Marriages are arranged for sons and daughters by their elders, who negotiate among themselves through a series of fixed visits and gift exchanges. Often the groom must perform service for his in-laws; during this period he may reside with them, later to set up his own abode. Marriages are easily dissolved in the absence of children, and a man may have a succession of wives; when children arrive, however, they stabilize a marriage.

Community political organization

The community's political organization is housed in a central building, usually opening upon a plaza. The personnel of the political organization form a hierarchy, with its top members recognized as spokesmen or representatives of the Indian community by the national government. The political offices are closely interwoven with a similar hierarchy, whose personnel serve the local church and the religious brotherhoods (*cofradías*), and plan the annual festal cycle. The personnel of the two hierarchies tend to alternate their periods of service between the civil and the religious wings. All adult men serve in this civil-religious hierarchy; in small communities they eventually reach the top posts and retire to become respected elders (*principales*). The annual change of personnel is accompanied by ritual.

There are no social classes in the Indian communities, but there are considerable differences in prestige, wealth, and individual achievement. There is an age hierarchy, especially among males, largely based on previous public service in the civil-religious hierarchy.

In the life cycle of these Indians, the important events are Baptism and marriage. There are no puberty rites; and death is accepted matter of factly, followed by a Catholic wake. The concepts of sickness and disease are various. Illness is often thought to be caused by invasions of wind (*aires*) into the body or by disturbed emotional states, or it is thought to be most likely to strike during certain periods. Foods and natural substances are thought to be hot and cold, strong and weak; these conceptions are used both in diagnosis and in curing. Those who cure are specialists or semispecialists, frequently believed to perform witchcraft or other noxious magic.

In the realm of the sacred there are a large number of supernatural beings and places. The deities are arranged in a vague hierarchy, sometimes with the Christian God at the apex. Christian saints are the chief focus of worship, but with them are associated various pagan attributes and forces, including the natural forces of wind, rain, and lightning. Some saints have strong cults, their effigies being housed in a cult house or special temple and cared for on a daily basis. Liquor is consumed during sacred ceremonies and in that context is itself sacred.

The annual cycle is regulated by the European calendar. There is a series of religiously derived festivals, the chief of which are All Saints' Day, All Souls' Day, Easter week, and the Day of the Cross (in September). At least 82 different Indian communities still use the old pre-Conquest calendars in some agricultural rituals and for purposes of divination.

The prevalence of animism

The world view is animistic in the sense that the Indians see the world as peopled by spirits, souls, ghosts, and witches capable of inflicting harm if the proper ritual precautions are not taken. Omens, dreams, and talismen are of great importance. People are also believed to transform themselves into animals and mystically eat the life from a victim. Some communities execute witches. In religious practice, ritual conformity is more important than inner piety; if a person does not hold a ritual office, he engages in very little daily religious activity.

Slander, gossip, and envy are strongly condemned, although they are an important means of influencing social behaviour. In general, however, the rule of law, as expressed in the formal organization of the civil-religious hierarchy, is much more important than personal leadership. Great ambition is discouraged, but industry and diligence are much lauded.

This generalized profile of Middle American cultures fits most closely the areas of the former high cultures—the Aztec, Maya, Zapotec, Tarascan, and Mixtec areas. It needs modification for the northwest culture area, where the communities have less economic specialization and interdependence. It also does not fit some coastal Indians who have become part of the export economy, such as the vanilla-growing Totonac of Veracruz, or the Indians who work on the sisal plantations of Yucatán, where many approach the status of a rural proletariat.

DEMOGRAPHY AND HISTORY

The Indians of Middle America are all descended from Asiatic forebears who crossed the Bering Strait and moved southward. They tend, except in the Isthmus of Tehuantepec, to be small in stature (155–160 centimetres or a little over five feet on the average), with brown to coppery skin, straight black hair, and dark-brown eyes often set above high cheek bones, sometimes with epicanthic folds. The Maya facial features are particularly distinctive, being flatter than those of the other groups; the Maya also have more prominent noses, and a tendency to rounder heads. Mexico is basically a mixed (*mestizo*) nation; there has long been extensive interbreeding between Indians and non-Indians. In Guatemala there has been much less interbreeding. But the term "Indian" is not a biological designation so much as a social, cultural, economic, and linguistic summary of the differences between some rural ways of life and the dominant national

culture. Race in and of itself is not socially as important as it is in other parts of the world. The usual census definition of "Indian" is based on linguistic criteria, and the population figures for Indians must therefore be read as figures for speakers of Indian languages.

In Mexico there are more than 2,500,000 speakers of Indian languages, chiefly in the states of Oaxaca, Chiapas, and Michoacán, and in the Yucatán Peninsula, although some Indian language speakers are found in all of the states except Aguascalientes, Nuevo León, and Tamaulipas. More than half the population of Guatemala, or nearly 2,000,000 persons, are speakers of Indian languages. More than 1,000,000 of them are concentrated in the western highlands; other Indian areas in Guatemala are the Verapaz and the eastern highlands.

The earliest inhabitants

Prehistoric periods. More than 10,000 years ago, the Middle American Indians were hunters who roamed the country in bands of four to ten persons. Their quarry were mammoths, llamas, bison, and wild horses, as can be seen from the remains found at Tequixquiac, north of the central valley of Mexico. Some also hunted small game and gathered the seeds of wild plants. The seed gatherers and the big-game hunters coexisted for thousands of years, until a climatic shift around 7500 BC favoured the seed gatherers.

About 4500 BC, cultivated squash and gourds became part of the subsistence of the Indians, and they adopted a seasonal nomadic pattern. Around 3500 BC the basic crops were domesticated: maize, beans, and squash have been found in remains of Tehuacán. The Neolithic era was beginning. By 2500 BC farming supplied about 10 percent of the food intake of the people, and by 2000 BC the first permanent settlements based on farming had been established, apparently villages of fewer than 100 persons. After 2000 BC farming became of major importance, and in the main areas the people were fully sedentary. They now produced forms of pottery.

From 1500 BC to the beginning of the Christian Era (often called the Formative period), the basic techniques of intensive agriculture were worked out, the full range of cultivated plants was developed, the pottery and art styles were formed, and the transition from small villages to ceremonial towns of 5,000 inhabitants was completed. The archaeological evidence of this may be seen in the central valley of Mexico at El Arbolillo, Zacatenco, Tlatilco, and, finally, Ticoman. The same developmental sequence occurred in the Formative period of highland Guatemala, as shown in the excavations at Kaminaljuyú near Guatemala City.

From the beginning of the Christian Era to about 1000 AD, the cultures of Middle America passed through their Classic or Florescent period. In the Pyramids at Teotihuacán and at Cholula, in the Oaxaca valley at Monte Albán, in the ruins at Tikal and Uaxactún in Petén, and at Copán and Piedras Negras, there is monumental testimony to the rise of theocratic civilizations. The differentiation between the earthbound peasants and the class of priests is marked by the spread of the hieratic art called by archaeologists the Olmec style, the greatest example being at La Venta in Tabasco, Mexico. Olmec art made much use of carved jade and the symbol of the jaguar. In the symbolic vocabulary of Middle America, the jaguar is associated with Tlaloc, the god of rain and fertility and therefore represents the forces of nature that make the crops grow. Tlaloc also dwells in caves from which lightning is thought to come; caves mean towns and settlements, so the jaguar probably represents the control of society as well as of nature. Jaguars are carved on axes, pots, and human likenesses; men wear jaguar skins; and human faces are depicted with jaguar-like mouths.

Mayan theocratic societies

This symbolism fitted well the new theocratic societies of the Maya—the Nahuatl, the Zapotec, and the Mixtec. Large raised pyramids were built for the public enactment of religious rites by the ruling priestly class. These structures were ceremonial centres where priests and specialists lived, and where the populace assembled for periodic rituals. The priests were important in the agricultural cycle, for which the ancient calendars were first devised. These calendars evolved into precise instruments for the measurement of time and came to express the philosophical ideas of the priestly intellectuals. The full poetry of Middle American calendrics is expressed in the Maya calendar "long counts," for which stone pillars were erected.

The ceremonial centres of the Classic period produced many luxury items made from imported goods such as seashells, feathers, obsidian, fine flint, jade, cotton, and cocoa beans. The centres paid for these imports by exporting finished ceremonial objects to the peripheries. The economy of the city-states of this period depended on the intensive cultivation of maize and on huge drafts of peasant labour for the building of the temple complexes. The building was done with a rather simple technology for beasts of burden, a functional wheel, and metal tools were lacking. The major achievements of the Classic period were its calendrics and its religious thought, which represented the efforts of the priestly class to systematize the universe and make it predictable.

Around the end of the Classic period a warrior group came into power. With the coming of the warriors a period of militarism began. War gods replaced the rain gods, and the jaguar symbolism gave way to the eagle complex in iconography. Human sacrifice, an occasional thing in the Classic period, became a mass phenomenon under the warlords of the Toltec and Aztec empires. The Toltecs, and later the Aztecs, were expansionist. When the Spaniards arrived, the Aztec Empire extended over most of Middle America, from the central valley to Yucatán, with garrisons, frontier fortifications, and foreign colonies. With the fall of the Aztec capital in 1519, the Guatemala conquest of 1524, and the long-drawn-out battles with the Indians of the northwest region, the Spaniards ended the isolation of Middle America.

The Conquest and its effects. The Spaniards overthrew the urbanized, class-structured high civilization of the Aztecs and established a system of alliances with the tributary states. The conquerors decapitated native society, substituting the Spanish for the Indian nobility. They introduced a host of new agricultural techniques and crops, along with steel, horses and cattle, mines, European crafts, and new forms of social organization. They also created dichotomies between ethnic and racial groups, giving them different rights and obligations.

From the colonial period to independence

In the early colonial period following the conquest, a small minority of Spaniards administered and controlled vast Indian populations. The religion of the conquerors spread rapidly, as did many of their domesticated plants and animals. In this period the Indians were grouped into villages modelled on the grid plan, with a central plaza on which stood the church and town hall. The basic economic institutions were the encomienda and the religious reservation: the encomienda was an allotment of land and labour to a Spanish overlord, used in the densely peopled areas of the former high civilizations; the religious reservation was established by the Jesuits in the northwest. The encomiendas integrated the Indians into the whole economy of New Spain, while the religious reservations tended, on the whole, to isolate them.

The later colonial period saw the abolition of the encomienda and the secularization of the religious reservation. Where previously the upper strata of the Indian population had intermarried with the Spanish, the later colonial period brought a more rigid separation of the Spanish and the Indians. Although the encomienda was legally dead, its economic consequences persisted: the colonial-plantation economy developed, along with the cattle ranch and the mining complex. Indian lands were taken over at an increasing rate, and the Indians gave up whatever illusion they may have had as to the benefits to be gained from the white man's culture.

Mexico and Guatemala became independent of Spain in 1821. The early republican period involved the Indians only marginally. As the native Spanish and mestizos threw off European domination with the slogans and banners of the French Revolution, they neither considered the Indians nor involved them in their struggles. For the Indians it was a time of cultural and social consolidation, spent in the building of defenses against land encroach-

ment and in the erection of communication barriers to protect their cultural heritage.

The later republican period was for Mexico and Guatemala a time of nation building, of moderate industrialization, and of commercial agriculture. The Indian was seen as anachronistic, an obstacle to the march of triumphant nationalism. For the Indian, the consequences were drastic. The national and international economy led him to develop a social structure and cultural pattern that would keep the Indian community intact, even at a low economic level. The strong corporate Indian community reached its most definitive form under the pressures of land-grabbing and anti-Indian social policies.

In the Mexican revolution that began in 1910, the Indians won a place in the political structure. The official policy became pro-Indian. This policy exalts the Indian heritage and makes it the root of the national heritage. In its economic and political aspects it aims at "incorporation" of the Indian into national society on terms set by the Indians themselves. It also provides special services to Indians, as demonstrated in the many programs of the Instituto Nacional Indigenista de Mexico in health, road building, agriculture, and literacy. It employs themes from Indian culture, as in the murals of the Mexican neo-Realists and the music of modern Mexican composers. In Guatemala a similar policy had a brief vogue, but it was reversed following the counterrevolution of 1954.

But the forces shaping contemporary Indian life lie largely outside these official policies. Population increase, the expansion of physical and cultural communication, industrialization, urbanization, and political participation are the basic forces that will determine the future of the Indians of Middle America.

The Indian population in Mexico in the 1970s was stable or declining, while that in Guatemala was increasing. The differences lay not so much in vital rates as in social and cultural dynamics. In Mexico there has been considerably more urbanization—with a greater flow of Indians to the cities—than in Guatemala; and when Indians are urbanized they become mestizos and ladinos. In the rural parts of Mexico there has also been a movement of Indians into Mexican society at a faster rate than in Guatemala. Thus, in the 1970s, only about 12 percent of the Mexican people were still counted as Indians, while the category comprised more than 50 percent of the Guatemalans.

BIBLIOGRAPHY. The most authoritative work on Middle America is the monumental *Handbook of Middle American Indians*, ed. by ROBERT WAUCHOPE, 11 vol. (1964–71). These volumes cover all aspects of the area from physical geography, linguistics, archaeology, physical anthropology, ethnology, to social anthropology. ERIC WOLF, *Sons of the Shaking Earth* (1959), is the best single book on Middle America. SOL TAX (ed.), *Heritage of Conquest* (1952), is a professional anthropological estimate of the content of Middle American Indian culture. *The Maya and Their Neighbors* (1940), is an introduction to the archaeological and historical problems of the region and contains the basic essay on Indian history by OLIVER LA FARGE. S.G. MORLEY, *The Ancient Maya*, 3rd rev. ed. (1956); J. ERIC THOMPSON, *The Rise and Fall of Maya Civilization*, 2nd ed. (1970); and MICHAEL D. COE, *The Maya* (1966), give the different views on that civilization. G.C. VAILLANT, *The Aztecs of Mexico*, rev. ed. (1956), is a basic book; and H.J. SPINDEN, *Ancient Civilizations of Mexico and Central America*, 3rd rev. ed. (1968), is important for its calendrical correlations. For an understanding of the colonial period, the works of BERNARDINO DE SAHAGUN, trans. by A.J.O. ANDERSON and C.E. DIBBLE, *The Florentine Codex*, pt. 9–10 (1959); and A.M. TOZZER'S annotated translation of Landa's *Relación de las cosas de Yucatán* (1941), is essential. For historical accounts, R.N. ADAMS, *Crucifixion by Power: Essays on Guatemalan National Social Structures, 1944–1966* (1970), is an up-to-date survey. GEORGE M. MCBRIDE, *The Land Systems of Mexico* (1923, reprinted 1971); FRANCOIS CHEVALIER, *La Formation des grands domaines au Mexique* (1952; Eng. trans., *Land and Society in Colonial Mexico: The Great Hacienda*, 1963); and NATHAN L. WHETTEN, *Rural Mexico* (1948), give an adequate picture. The modern studies in Middle American communities are listed in HOWARD CLINE, "Mexican Community Studies," *Hispanic American Historical Review*, 32:212–242 (1952); and in R.H. EWALD, *Bibliografía comentada sobre antropología social guatemalteca, 1900–1955* (1956). Modern developments are described in J.L. ARRIOLA (ed.), *Integración social en Guatemala* (1956); W.E. MOORE, *Industrialization and Labor* (1951); and by R.N. ADAMS in *Political Changes in Guatemalan Indian Communities* (1957).

(Ma.N.)

Middle Eastern and North African Peoples and Cultures

The name Middle East has become a conventional term designating the whole area extending from the Atlantic coast of Morocco and Mauritania eastward across North Africa to Afghanistan. So defined, it includes the Sahara and the Mediterranean coast of Africa (the countries of Morocco, Algeria, Tunisia, Libya, Egypt, and The Sudan), the Arabian Peninsula, Iraq, Turkey, and Iran. This area is now regarded as a broad cultural unit, primarily in terms of its overarching Islāmic civilization.

A number of distinctive cultural features give the area unity and justify the use of the term Middle East. These take the form of similar constellations of subcultures across the area consisting of urban networks with their agrarian and pastoral hinterlands. Parts of the area are often treated separately: the Maghrib, a term that designates the peoples of present-day Morocco, Algeria, Tunisia, and Libya; Mashriq, the eastern area including the present countries of The Sudan, Egypt, Syria, Lebanon, Israel, Jordan, Iraq, Saudi Arabia, Yemen (Aden), Yemen (Ṣan'ā'), Oman, Kuwait, and other Gulf emirates; Anatolia (Turkey); and Iran.

CHARACTERISTICS OF MIDDLE EASTERN CULTURE

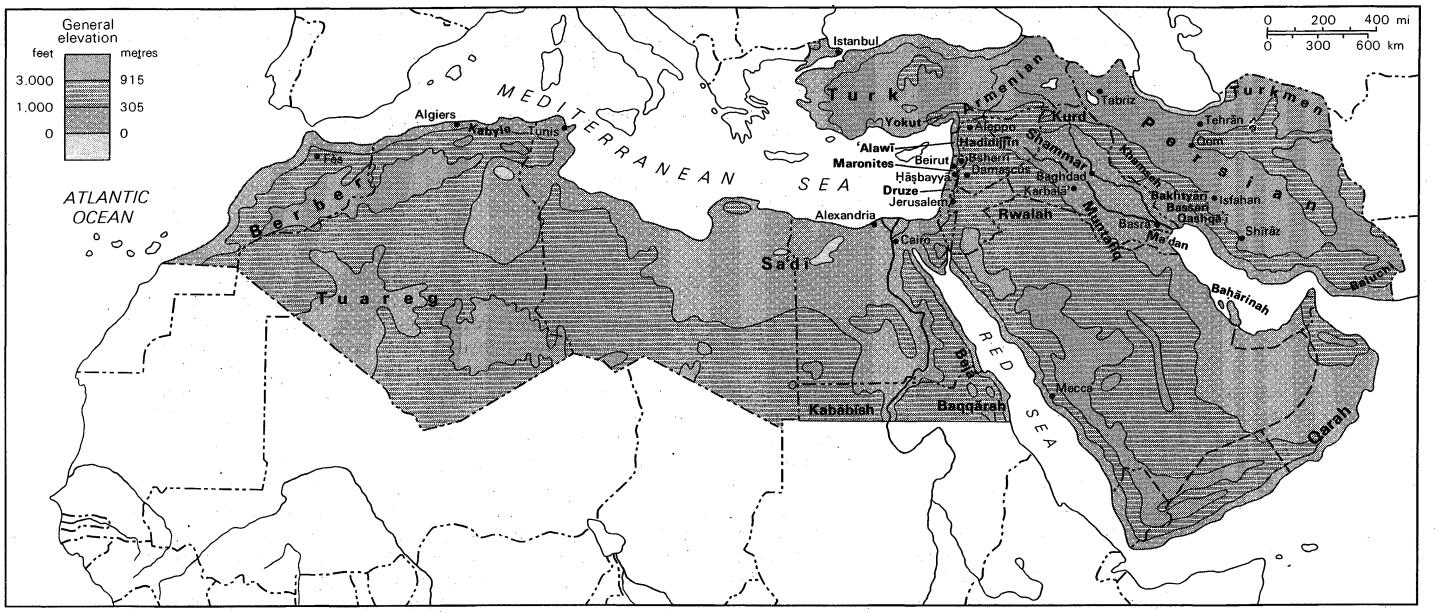
Historical overview. Scattered Paleolithic evidences of man have been found throughout the Middle East. With the Neolithic Period, however—that period commencing about 10,000 years ago, during which plants and animals began to be domesticated—a geographical concentration appears, and an evolution of Middle Eastern cultures took place in the sickle-shaped zone extending from the Nile Valley to the valleys of the Tigris and Euphrates rivers and down the eastern slopes of the Zagros Mountains in Iran. Exploration of eastern Turkey and northern Iran has indicated that the earliest domestication of plants and animals occurred in that central area. About 2,000 years later, civilizations based on irrigation began to develop, first in the lower reaches of the Tigris–Euphrates Valley and then in the Nile Valley. This irrigation technology led to the development of states and empires. Scholars traditionally have looked upon the areas along the Nile and Tigris–Euphrates rivers and between as the region of the earliest development of the ancient high civilizations of the Old World. In recent times, however, evidence from southern Arabia, especially the Yemen, has led to an extension of this area.

By Greco-Hellenic times, trade connections had been established with the Far East. As civilization developed in Europe, the Middle East from Morocco to Iran became an area of passage. There was increasing movement into and across the area, not only in the form of massive invasions but in commercial, diplomatic, and religious contexts. This began most intensively at the time of the rise of the Arab kingdom and of Islām (AD 640), and it accounts for the concept of the Middle East as that central area of Islāmic control that reached from Morocco to Iran, incorporating or incapsulating a multitude of peoples, languages, and cultural adaptations.

Unity and diversity. In very broad terms, the culture of the Middle East consists of great urban centres separated by dispersed rural peasantries in the mountains and plains, along with a variety of patterns of nomadic pastoralism in the steppes and deserts. But this tripartite division tends to oversimplify the ecological diversity and ignore the cultural continuity of the whole Middle Eastern area, which should be thought of as laced together by a number of connecting institutions. Indeed, the Middle East has always been linked by its merchants, scholars, religious brotherhoods, and other communicating agents.

The cities

A complex urban life is one of the most distinctive characteristics of the Middle East, and this has been true throughout its recorded history. More than a dozen of the world's greatest cities are found there: Alexandria, Al-



Distribution of Middle Eastern and North African peoples and cultures.

giers, Aleppo, Baghdad, Basra, Beirut, Cairo, Damascus, Fès (Fez), Isfahan, Istanbul, Jerusalem, Mecca, Shīrāz, Tehrān, Tabriz, and Tunis. Jerusalem is a world centre for three major religions: Judaism, Christianity, and Islām. Mecca is another major centre for Islām, and many secondary and tertiary urban centres are also religious centres, such as the Druze community's holy towns of Ḥāṣbayyā, Rāshayyā, and B'qlīn, in the Lebanese mountains. There are also the sacred Muslim towns of Tarīm in southern Arabia, Karbalā' in Iraq, and Qom in Iran. The Maronite Christians of northern Lebanon have their sacred town in Bsharrī. The city as sacred ceremonial centre has from antiquity been a major cultural characteristic of the Middle East.

The extensive urban civilization of the area is closely associated with nonurban patterns of nomadic pastoralism and various types of peasant agriculture, as well as with the connecting institutions mentioned above. It is important to bear in mind the antiquity and continuity of these stable, complex centres of civilization situated in optimum locations, with less complex cultural systems effectively tied to them by the integrator networks. As a whole, this culture has survived 10,000 years with few breakdowns until the most recent times.

Peoples and languages

Another characteristic of the Middle East is its cultural diversity. It includes the Bedouin tribes of northern Arabia, the Oriental Jewish and Assyrian Christian communities, the Muslim Ibāḍīs of Algeria and Oman, the Turkmen of northeastern Iran, the Kabyle Berbers of Algeria, and the Kurds, Armenians, and others. But most of these communities are widely dispersed. The Middle East is at once a vast patchwork of dispersed peoples and languages and a fragmented area of enclaved ethnic groups locally dominant over others scattered among them. Probably there is no complete enumeration of such groups, much less an agreement upon their status or their demographic features. Even as some are clearly disappearing, new ones are forming. Most conspicuous among the new ethnic groups are the Algerian and Palestinian nationals and the Afro-Lebanese caste appearing in Lebanon.

The linguistic structure of the Middle East is also very complex. Four major language stocks are represented: Semitic, Hamitic, Indo-European, and Ural-Altaic. In all but the most remote communities, people are multilingual, commonly with some knowledge of at least one European language. But there are marked dialectal distinctions carrying caste, class, and political significance, especially in urban centres. These include such distinctive hybrids as the argots of Persian Gulf sailors or of the servant classes in Mediterranean cities. Finally, every urban quarter, rural village, religio-ethnic group, or pastoral tribe has its own linguistic forms, its own vocabulary, its secret names for people and places.

CITIES AND LINKAGES

Urban life. Whether it is a port of trade or a ceremonial centre or a political centre, the Middle Eastern city possesses certain marked characteristics. It is roughly circular in shape and divided into a number of distinct sections or "quarters." Some cities still have walled quarters, with gates shutting them off from each other guarded by watchmen. The government or political centre of the city forms one quarter, with its garrison, hospitals, and palace or fortress; it also contains the major mosques associated with the particular Islāmic sect of the ruling elite, together with Muslim schools and universities.

Structure of the Middle Eastern city

A second important quarter of every city is the permanent covered marketplace, the bazaar or suq, situated near the government quarter. The bazaar is a labyrinth of alleyways, divided into sectors according to crafts and types of merchandise and with small mosques and chapels scattered through it. The alleyways are covered; often they are arched with stone, as in the old bazaar at Aleppo or in those of the great cities of Iran; sometimes, in the smaller towns, they are simply shaded by mats. The caravansaries of inter-city and inter-regional trade are located in the bazaars.

The residential quarters were traditionally reserved to Muslims. Depending upon the size of the city, these might be divided according to sect, ethnic group, or city of origin. Non-Muslims generally live outside the city walls in their own quarters: Christians of many sects, Jews, and villagers from the hinterland. Each of the residential quarters of these other ethnic groups or non-Muslim religious groups contained its own small bazaar and its religious centres and schools.

Outside the city lies an agricultural zone from which the city draws food and other necessities. In this area are located the slaughterhouse, tanneries, and the spinners and dyers that supply the textile workshops in the city bazaar.

Integrating networks. Middle Eastern cities have historically been connected with each other by caravans and coastwise or riverine sailing craft. The organization of commercial ventures between cities was undertaken by the merchants and carried out as joint endeavours. Merchant families maintained houses or branches in a number of cities—kinship networks that continue today at all levels of society. The caravan routes have existed since antiquity, adapted to the major carrying vehicle, the camel. Today trucking routes follow virtually the same tracks. Along the routes are stopping places, often forti-

Merchant organizations and religious brotherhoods

fied *khāns* (caravansaries) associated with villages or towns where food and drink and other amenities were available to the caravan people. The locations of many cities and towns all across the Middle East, from Morocco to Iran and well into Anatolia and Armenia, were determined by the caravan routes.

Another institution linking the Middle Eastern cities is that of the religious brotherhoods, both Muslim and Christian, whose networks of mission stations or religious centres have often been parallel to and part of the commercial network. Sometimes a particular brotherhood was itself also a commercial network covering vast areas in which people were sparsely dispersed in oases or nomadic groups but across which it was important to transport goods. One recent example was the Sanūsīyah Brotherhood, whose founder, Sīdī Muḥammad as-Sanūsī, moved from Iraq in the 19th century to the oases of present-day Libya. Ultimately the expansion of Sanūsīyah followers gave rise to a network linking the Mediterranean ports with the Saharan and trans-Saharan trade routes. The order promulgated learning, stimulated and promoted commerce, and maintained a communications network, emerging as the military leadership of Libya in the wars of the 20th century. The leader of the brotherhood, a descendant of the founder, became the ruler of the Kingdom of Libya for its short life.

Many other institutions link the cities and settlements of the Middle East with the rural agrarian hinterlands and the pastoral tribes. Some are extensions of the power centres of the kingdoms, provinces, or empires that have appeared and disappeared before and since the rise of Islām, such as the armies of the ruling elites and the courier services of rulers and governors. At lower levels there are the mass migrations of ethnic groups, the annual seasonal movements of nomads, the networks of peasant markets, and the itinerant gypsy groups providing tinkering, tattooing, and stud animals for breeding. Among the wandering groups and individuals are Ṣūfī mendicants, essentially beggars but often reputed to be agents of the state; the jongleur, or itinerant singer, who sings his way across the country and may also be an agent of a particular ruler; itinerant medical practitioners, barbers, circumcisers, tooth pullers, and census takers; country peddlers; and animal traders. Thus no district or community of the Middle East has ever been isolated from news of the world or from the centres of power. In the most remote areas one is always surprised at how much the peoples of the villages know about the external world. No tribesman or goat herder is under the illusion that his people are the only living people on Earth. A continuity and ramification of contact across languages and religions are quite clearly built into the ideologies, the social relationships, and the economy and technology of the traditional Middle East.

Migrant groups

The cities and the integrating mechanisms described above are the dominant Middle Eastern social structures, the major centres of political power, knowledge, formal culture, and sacred ceremony for very large numbers of people. It is incorrect to think of the Middle East as an agglomeration of ethnically distinct people and cultures. With the exception of dense concentrations of cultivators along the two great river systems, the Nile and the Tigris–Euphrates, the cultural and population centres lie in the network of ancient and contemporary cities. The cities themselves are multi-ethnic in character. Even though a religious centre such as Mecca may exclude all except Muslims, it is still a great gathering place of peoples, illustrating the ethnic diversity of the followers of Islām. The very structure of the Middle Eastern city is exemplary of the area in general, of the coexistence of many different ethnic, religious, and linguistic groups. The governing apparatus of the state is everywhere the mediator between these ethnic groups and has in fact served to maintain an equilibrium among them for thousands of years. Only since the 19th century has this ancient network, upon which the Ottoman Empire was built, begun to dissolve and break down. This fragmentation has taken place under the impact of the expansion of Western and Eastern nations and of industrialization.

PEASANT CULTIVATORS

Perhaps 70 percent of the peoples of the Middle East are peasant cultivators. With a few notable exceptions, such as those in the Berber highlands of Morocco and Algeria, peasants live in nucleated village settlements on marginal lands near their fields. Berber highland settlements tend to be dispersed hamlets within areas belonging collectively to kinsmen.

Property and tenure

Within the sphere of state or provincial control, the peasant communities are often dominated by landlords and pay taxes to the state. Elsewhere the cultivating population is often found in ethnically distinct enclaves that challenge the power of the state to collect revenues. The "land of government" generally includes areas in the vicinity of urban centres, open plains easy of access, and settlements on established trade routes. Examples are the Nile Valley; the lower reaches of the Tigris and Euphrates, from Baghdad to the marsh areas of the Shatt al-Arab; the plains of Esdraelon and the lower foothills of old Palestine; the vicinities of Tunis and Fès in the Maghrib; the plains of western Syria from Damascus north to Aleppo; and the coastal plains of Turkey and Iran along the Black and Caspian seas.

Forms of land tenure vary exceedingly. Individual tenure is rare in comparison to the great estates of urban landlords and rural aristocratic families. The latter might in the past encompass as many as 90 villages, as in the Syrian plains or in Egypt, or be interspersed with freehold peasant plots, as were the estates in the Lebanese mountains controlled by the great feudal families. Communal management of the lands associated with a village was a customary means of distributing the land among the villagers, according to the supply of men and oxen. Such communal tenures were known in the Levant as *mushāʿ* but appear in many parts of the Middle East in various forms depending on local resources. Everywhere, however, a major category of property has been *waqf* (Arabic) or *habous* (Berber) land—*i.e.*, land or other property controlled as a religious or charitable endowment by ecclesiastics or trustees. Control of such property was the basis of the power of the Muslim religious brotherhoods, the Christian monastic orders, and the saintly lineages. Even at the lowest levels of society, the classes of priests and *imāms* or mullahs of the village churches and mosques based their power upon the control of property.

Village communities

Modern Eastern villages show common organizational traits. Everywhere the patrilineal family dominates, with alliances between families made by marriage and economic cooperation. A village council composed of the heads of larger families maintains local customs, decides on the filling of such local offices as village herdsman and field guards, and acts in times of threat to the village as a whole from predatory landlords, armies of the state, or pastoral tribesmen. Internal economic specialization is not very great, especially in the villages of the state-controlled areas. Most families are relatively self-sufficient, and productive and maintenance tasks are divided according to sex. The status of women has long been misrepresented as "low"; it is, rather, protected and varies with the status of the whole family.

Communities of the relatively autonomous areas (highland Berbers, Druzes, Maronites, ʿAlawīs, Kurds) are often more complex. Here are found patrilineal alliances of families in the form of clans and a ranking of clans and families as political and religious leaders. Such communities may ally with each other territorially, forming "tribal" entities.

A universal feature of the Middle Eastern social landscape is its lack of homogeneity at all levels, even familial. Most villages in which one ethnic or religious category predominates will also have at least a few families of a different group who provide a local specialization, such as making pottery or herding the village flocks. This is also true for many rural areas. For example, the Kurdish mountain area at the juncture of the countries of Iraq, Turkey, and Iran is also inhabited by communities of Jews, Arabs, Assyrians, Turks, and Armenians. The most ethnically homogeneous areas are probably the Berber

highlands of Morocco and Algeria, but even there the "tribes" are usually heterogeneous in origin.

Crops and crafts

Methods of cultivating the land are various. Cash cropping in cotton, cereals, and sesame is widespread in the plains areas; truck gardening in vegetables is also important, especially near urban centres. Wheat and barley and such legumes as chick-peas and lentils are ancient and ubiquitous throughout the Middle East. Arboriculture is equally important. A variety of nut and fruit trees is cultivated in the Zagros and Taurus mountains. The olive and the fig are apparently almost as ancient as cereal cultivation around the Mediterranean. Southern Iraq is the major date-producing area of the world. A rich variety of subtropical fruits, including apricots, oranges, and tangerines, is also long known in the area, deriving from the east. Tea, rice, and coffee are found in Iran and the Yemen.

Rural crafts are also widespread. To a great extent they belong to the relatively self-sufficient peasant householding economy, but there is some production for the market, especially in the vicinity of major cities. Manufactured pottery, textiles, mats, trays, and baskets form staple rural crafts. The "industrial village" of Bayt Shabāb, in central Lebanon, has for many centuries produced cotton textiles, church bells, and large pottery jars. While various ethnic groups tend to specialize in particular crafts, this specialization is often determined by the locality. Thus, the Jewish communities of the Yemen specialized in metalcrafts, vineyard cultivation, and wine making, but in the Levant and North Africa they found other occupational niches related to literacy and money handling. Armenians in their traditional homeland area engaged in all occupations; elsewhere, they held special status as merchants to the *shāh* in Iran or as craftsmen in southern Turkey and in the cities of the Levant.

NOMADIC PASTORALISTS

Kinds of nomadic tribes. The various nomadic pastoralist societies of the Middle East have been much written about but constitute only 10 to 20 percent of the population. Control of the long-distance caravan routes and of particular resources that can only be reached by caravan, such as salt in the Sahara, lie with the nomadic tribes and have given rise to concentrations of power, such as the confederations of the Tuareg of the Sahara, of the Shammar of northeast Arabia, and the Khamseh of the central Zagros Mountains of Iran. Many nomadic pastoralist societies, however, are primarily specialized animal breeders occupying marginal lands and usually combining small-scale cereal cultivation with their pastoralism. The Kabābīsh of the northern Sudan, the shepherd peoples of the Levant and of parts of Libya, and others depend not only upon their flocks and herds but also upon agricultural or other resources. Sheep and goat breeders such as the Sa'dī of Libya and Egypt, the Mawālī and Ḥadidiyyīn of the Levant, the Muntapīq of southern Iraq, and the Bāṣerī (Basseri), the Bakhtyārī, and the Qashqā'ī of Iran are marginally associated with agriculturalist. The Bedouin of the Arabian Peninsula and North Africa, who specialize in camel breeding, occupy steppe and desert areas. Cattle herders, much less common, are represented by the Baqqārah (Baggara) of The Sudan and the Qarah of southern Arabia. Unique in the area, if not in the world, are the marsh-dwelling Ma'dan of southern Iraq, whose water buffalo are their major economic resource. The camel, by virtue of its speed and mobility, its weight-bearing capacity, and its ability to live upon desert vegetation, enables specialized camel pastoralists to dominate other societies.

Control of caravan routes

Patterns of movement. The nomadic pastoralists follow two patterns of movement: horizontal nomadizing and vertical transhumance. Both are seasonal adaptations to grazing conditions. In horizontal nomadizing the flocks move from permanent sources of water during hot, dry summers into steppe and desert pasturages during the cooler, moister winters, returning to the permanent wells in late spring. The Rwalah Bedouin of northern Arabia have traditionally summered in the vicinity of the Azraq oasis, patronizing the markets of 'Ammān and Damascus, and moved deep into the north Arabian steppes during the winter. Vertical transhumance is characteristic of sheep and goat herders in mountainous areas. They winter their flocks on warm lowland pasturages and trek to highland summer pasturages through areas often rather densely occupied by peasant cultivators. Numerous Berber tribes of Morocco and Algeria have subsisted in this way, as have the Yokut peoples of southern Turkey. Best known, however, are the great shepherd tribes of Iran's Zagros Range, the Bakhtyārī and the Bāṣerī.

Nomadic social organization

The nomadic pastoral peoples of the Middle East all have a clan or lineage organization, and all appear to be patrilineal in descent and inheritance, except for the matrilineal Berber Tuareg of the Sahara. Most large and wealthy tribes also have socially ranked clans, with aristocratic lineages holding leadership. The camping units of nomadic pastoralists generally comprise a core of kinsmen; the camps of tribal leaders contain more distant kin, along with such members of specialist castes as blacksmiths, camel traders, or hunters and water scouts.

The ethnic diversity general in the Middle East is found also among the nomadic pastoral peoples, who include Berbers, Arabs of North Africa, Beja (Bujah) and Baqqārah peoples of the Red Sea littoral and northeastern Sudan, Turks, Turkmen, Kurds, and various Persian and Baluch peoples. All are nominally Muslim.

CULTURAL CHANGES IN THE MIDDLE EAST

Colonialism and revolution

Marked economic and political changes in the Middle East began in the early 19th century with the European colonial penetration of North Africa, southern Arabia, and the Persian Gulf. At the end of World War I, the Ottoman Empire was dismembered, and Iran was united under a new dynasty. During the period between the wars, national boundaries and parliamentary institutions were arbitrarily established throughout the Middle East, except in the Arabian Peninsula and Palestine. Nationalist movements appeared, and the years 1925–30 saw rebellions and disturbances throughout the area. There were large population movements, such as the exchange of Greeks and Turks. The repressive pacification and suppression of the Berber peoples of Morocco and Algeria was carried out. The colonization of Palestine by European Jewish communities brought clashes with the British and with the indigenous peoples, mostly Arab peasants. Within a decade after World War II, national independence had been achieved by Syria, Lebanon, Iraq, and Jordan. Israel had been established, the long Algerian revolution was underway, and Morocco, Tunisia, Libya, and Egypt had been declared sovereign states. Everywhere the traditional economic patterns and institutional networks had been severely affected. Foreign economic enterprises controlled the exploitation of Middle Eastern resources, using the area as a market for manufactured goods and as sources of petroleum, fruits, and a few minerals. Industrialization of the Middle Eastern economies was minimal, except in the state of Israel, dominated by a sophisticated and highly trained elite largely of European origin.

By the mid-1960s, however, a number of countries in the Middle East were moving toward secularized nationalist regimes and acquiring greater control over their own resources and populations. The ancient cities had become centres of industrialization; the old urban–rural networks fed a steady flow of peasants to the cities, and, as in the rest of the industrializing world, shantytowns sprang up around the major cities. Both Algeria and the Levant supported internment camps of hundreds of thousands of refugees, and it is these concentrations of displaced peoples, chiefly peasants, that gave rise to revolutionary movements and the beginnings of organizational change in the Middle Eastern societies.

By the 1970s the Middle Eastern societies had not yet made the transition to modern organization. Throughout much of the area there was a greater division than ever between tiny, rich elites and masses of displaced peasants and impoverished pastoralists. Land reform and redistribution had only partially succeeded in Iran, Algeria, Egypt, and Syria. Secularized legal and judicial systems,

expanded national education systems, and basic industrialization were developing slowly and sporadically. In farming, modern technology and marketing practices were gradually being adapted to Middle Eastern climatic and soil conditions.

BIBLIOGRAPHY

General surveys, histories, essays, and anthologies: A.J. ARBERRY (ed.), *The Legacy of Persia* (1953, reprinted 1963); B.C. ASWAD, *Property Control and Social Strategies: Settlers in a Middle Eastern Plain* (1971); PETER AVERY, *Modern Iran* (1965); JACQUES BERQUE, *Le Maghreb entre deux guerres* (1962; Eng. trans., *French North Africa: The Maghrib Between Two World Wars*, 1967); PHILIP K. HITTI, *History of the Arabs*, 10th ed. (1970); BERNARD LEWIS, *The Emergence of Modern Turkey*, 2nd ed. (1968); RAPHAEL PATAI, *Golden River to Golden Road: Society, Culture, and Change in the Middle East*, 2nd ed. (1962); K.S. SALIBI, *The Modern History of Lebanon* (1965); PATRICK SEALE, *The Struggle for Syria: A Study of Post-War Politics, 1945–1958* (1965); LOUISE E. SWEET (ed.), *The Central Middle East: A Handbook of Anthropology and Published Research on the Nile Valley, the Arab Levant, Southern Mesopotamia, the Arabian Peninsula, and Israel*, 2 vol. (1968), and *Peoples and Cultures of the Middle East: An Anthropological Reader*, 2 vol. (1970); W. MONTGOMERY WATT, *Islam and the Integration of Society* (1961).

Geographical studies: W.C. BRICE, *South-West Asia* (1966); K.W. BUTZER, "Physical Conditions in Eastern Europe, Western Asia and Egypt Before the Period of Agricultural and Urban Settlement," *Cambridge Ancient History*, 3rd ed., vol. 1, pp. 35–69 (1970); W.B. FISHER, *The Middle East*, 6th ed. (1971). Outstanding recent works on Middle Eastern cities include: J.L. ABU-LUGHOD, *Cairo: 1001 Years of the City Victorious* (1971); ABDALLA S. BUJRA, *The Politics of Stratification: A Study of Political Change in a South Arabian Town* (1971); JOHN GULICK, *Tripoli: A Modern Arab City* (1967); A.H. HOURANI and S.M. STERN (eds.), *The Islamic City* (1970); I.M. LAPIDUS, *Muslim Cities in the Later Middle Ages* (1967), and (ed.), *Middle Eastern Cities* (1969); ROGER LE TOURNEAU, *Fez in the Age of the Marinides* (1961); BERNARD LEWIS, *Istanbul and the Civilization of the Ottoman Empire* (1963).

Studies of religious brotherhoods: J.M. ABUN-NASR, *The Tijaniyya: A Sufi Order in the Modern World* (1965); A.J. ARBERRY, *Sufism: An Account of the Mystics of Islam* (1950); E.E. EVANS-PRITCHARD, *The Sanusi of Cyrenaica* (1949); ERNEST GELLNER, *Saints of the Atlas* (1969); RICHARD P. MITCHELL, *The Society of the Muslim Brothers* (1969).

Rural studies and studies of particular communities: HAMED AMMAR, *Growing Up in an Egyptian Village: Silwa, Province of Aswan* (1954, reprinted 1967); HAROLD B. BARCLAY, *Buurri al Lamaab: a Suburban Village in the Sudan* (1964); FREDRIK BARTH, *Principles of Social Organization in Southern Kurdistan* (1953); PIERRE BOURDIEU, *Sociologie de l'Algérie* (1958; Eng. trans., *The Algerians*, 1962); JEAN DUVIGNAUD, *Chebika, mutations dans un village du Maghreb* (1968; Eng. trans., *Change at Shebika: Report from a North African Village*, 1970); PAUL W. ENGLISH, *City and Village in Iran: Settlement and Economy in the Kirman Basin* (1966); ROBERT A. FERNEA, *Shaykh and Effendi* (1970); S.M. SALIM, *Marsh Dwellers of the Euphrates Delta* (1962); LOUISE E. SWEET, *Tell Toqaan: A Syrian Village* (1960); J.S. SZYLIOWICZ, *Political Change in Rural Turkey: Erdemli* (1966); ALEX WEINGROD, *Reluctant Pioneers: Village Development in Israel* (1966).

Studies of nomadic pastoralism: TALAL ADAD, *The Kababish Arabs* (1970); FREDRIK BARTH, *Nomads of South-Persia: The Basseri Tribe of the Khamseh Confederacy* (1961); IAN G. CUNNISON, *Baggara Arabs: Power and the Lineage in a Sudanese Nomad Tribe* (1966); EMANUEL MARX, *Bedouin of the Negev* (1967); P. SALZMAN (ed.), "Comparative Studies of Nomadism and Pastoralism," *Anthrop. Q.*, 44:104–210 (1971).

Translations from contemporary literature: An important source of appreciation of Middle Eastern life may be found in these rare translations. A few titles are: F.M. ESFANDIARY, *Identity Card* (1966); YASHAR KEMAL, *Ortadirek* (1960; Eng. trans., *The Wind from the Plain*, 1969) and *Ince Memed* (1958; Eng. trans., *Memed, My Hawk*, 1961); NAGUIB MAHFOUZ, *Zuqāq al-madaqq* (1961; Eng. trans., *Midaq Alley, Cairo*, 1966); MOULOUD MAMMERI, *Le Sommeil du juste* (1955; Eng. trans., *The Sleep of the Just*, 1958); A.R. SHARKAWI, *al-Ardh* (1954; Eng. trans., *Egyptian Earth*, 1962); LEON SURMELIAN, *Daredevils of Sassoun: The Armenian National Epic* (1966).

(L.E.S.)

Mies van der Rohe, Ludwig

Ludwig Mies van der Rohe was one of the leading architects of the 20th century. Regarding architecture as "logical structure," he evolved a classically balanced, utterly simple architectural style. His buildings were marked by an uncluttered, almost skeletal exterior framework, little division of interior space, and masterful craftsmanship and simplicity of detail. World architecture continues to oscillate between his precepts and those of his contemporary, Le Corbusier. But while Le Corbusier, inspired by the white plasticity of Mediterranean architecture felt that "passion can create drama out of inert stone," Mies, brought up in the chill, rationalist climate of northern Europe, was a classicist. He was the supreme craftsman, who kept applying and refining the same basic structural and aesthetic principles he had discovered early in his career to every new architectural problem he was asked to solve. "I would rather be good than original," he has said. But Mies's mastery of proportion and detail, which has been compared to that of great Renaissance artists, such as Lorenzo Ghiberti, is not easy to match. In an age of built-in obsolescence, his was most of all an architecture of quality.

Artemis Verlag und Verlag fur Architektur, Zurich

Mies van der Rohe.

Early training. Ludwig Mies (he added his mother's surname, van der Rohe, when he had established himself as an architect) was born on March 27, 1886, in Aachen, or Aix-la-Chapelle, Germany, one of the great historic cities of Europe and once the centre of Carolingian culture. His father was a master mason who owned a small stonecutter's shop. Helping his father and working on various construction sites as a boy gave Mies a feeling for the nature of materials and craftsmanship that is rarely acquired in architecture schools. Mies never received any formal architectural training, a lack of which he seemed almost proud. At the age of 15 he was apprenticed to several Aachen architects for whom he sketched charcoal outlines of "Renaissance" ornaments on large sheets of brown paper pinned to the wall, which the plasterers would then form into stucco building decorations. Although Mies thought it a grueling task, it developed his skill for linear drawings that was to produce some of the finest architectural renderings of his time.

Influences on Mies's architectural development. In 1905, at the age of 19, Mies went to work for an architect in Berlin, who, it turned out, specialized in the design of wooden buildings. Mies felt that he knew too little about this material and soon left his job to become an apprentice with Bruno Paul, a leading furniture designer who worked in the Art Nouveau style of the period. Two years later he received his first commission, a traditional suburban

house. Its perfect execution so impressed Peter Behrens, then Germany's most progressive architect, that he offered the 21-year-old Mies a job in his office, where, at about the same time, Walter Gropius and Le Corbusier were also just starting out.

Association with the Werkbund

Behrens was a leading member of the Deutscher Werkbund, and through him Mies established ties with this association of artists and craftsmen, which advocated "a marriage between art and technology." The Werkbund was inspired by the English arts and crafts movement of William Morris and John Ruskin. But while Morris set out to revive medieval craftsmanship in romantic protest against the industrial revolution, the Werkbund, with equally romantic zeal, welcomed technology almost to the point of idolatry. What was needed, its spokesmen declared, was to give form and meaning to machine-made things, including machine-made buildings. This new and "functional" design for the industrial age would then give birth to a *Gesamtkultur*, that is a new universal culture in a totally reformed man-made environment. Endlessly and often poetically discussed, these ideas were not so much a program as a vision, an article of faith. It was a faith that, throughout the first half of the 20th century, motivated the "modern" movement in architecture and was designated the "International Style," by the architectural historian Henry-Russell Hitchcock and the architect Philip C. Johnson, in their book of the same name published in 1932.

It was soon obvious to Behrens that Art Nouveau, with its sensuous undulations, was hardly what the Werkbund envisioned. Instead, he began to emulate the pure, bold and simple Neoclassic forms of the German early 19th-century architect Karl Friedrich Schinkel. Although Mies was impressed by many of Behrens's works, such as the Berlin Turbine Factory of 1909—an original Neoclassic statement in steel and glass—and the German Embassy at St. Petersburg of 1911, it was Schinkel, not Behrens, who became the decisive influence on Mies's search for an architecture of *Gesamtkultur*. Throughout his life, the elegant clarity of Schinkel's buildings seemed to Mies to embody most perfectly the form of the new environment.

Influence of Karl Schinkel and Hendrik Berlage

Another decisive influence was Hendrik Petrus Berlage, a pioneer of modern Dutch architecture, and the first European actually to view Frank Lloyd Wright's American houses that so excited the European avant-garde. Mies encountered Berlage in The Hague, where Behrens had sent him to supervise the construction of a full-scale wood and canvas mock-up of a large house for Mme H.E.L.J. Kröller, a prominent art collector. Mies stayed a year because Behrens's design did not satisfy Mme Kröller, who then asked Mies to submit his own. Although Mies's more elegant version was never built either, his year in Holland gave him time to acquaint himself with Berlage's work and philosophy. Berlage's work inspired his own love for brick and the Dutch master's philosophy inspired Mies's credo of "architectural integrity" and "structural honesty." Later, teaching in Chicago, Mies defined "architectural integrity" to his students: "I hope you will understand that architecture has nothing to do with the invention of forms." As to structural honesty, Mies went further than anyone else to make the actual rather than apparent or dramatized supports of his buildings their dominant architectural features.

Work after World War I. In Imperial Germany only university graduates could become officers, and thus throughout World War I Mies served as an enlisted man, building bridges and roads in the Balkans. When he returned to Berlin in 1918, the Kaiser's Germany, with its arrogant hostility to the modern movement, had gone down in defeat, and despite the hunger and violence there seemed to be in the turmoil of the revolution a new hope that gripped every artist and intellectual and inspired a prodigious burst of new creativity. Architecture, painting, and sculpture, according to the manifesto of the Bauhaus —the avant-garde school of the arts just established in Weimar—were not only moving toward new forms of expression, but were becoming many-facetted and international. Never before had painting and sculpture—among them Dadaism from Zürich, a nihilistic, iconoclastic, antirational movement in the visual arts and literature; de Stijl from Holland, an abstract geometric movement characterized by the use of rectangles and primary colours; and the Russian abstract movements of nonobjectivism in painting and sculpture known as Constructivism and Suprematism—so decisively determined architectural thought. Mies was in the thick of it. Though hardly gregarious, he became a leader of the *Novembergruppe*, a group named after the German revolution of November 1918, which had also advocated an artistic one. He joined the *Zehnergruppe*, a group dedicated to popularizing the concepts of modern architecture, helped found and finance the magazine *G* (for *Gestaltung*, or "form giving") in which many of his theoretical projects were published, and organized many exhibitions. But there was virtually nothing for him to build, and his foremost building of this period—an Expressionist memorial to the murdered Communist leaders Karl Liebknecht and Rosa Luxemburg, dedicated in 1926—was demolished by the Nazis.

Mies's most important work of these years remained on paper. In fact, these theoretical projects, rendered in a series of drawings and sketches that are now in the New York Museum of Modern Art, foreshadowed the entire range of his later work. The Friedrichstrasse Office Building (1919) is one of the first proposals for an all-glass building and establishes the Miesian principle of "skin and bones construction." The "Glass Skyscraper" (1921) applies this idea to curved forms so that the facade reflects itself. Both of these building designs were uncompromising in their utter simplicity. Other theoretical studies explore the potentials of concrete and brick construction, and of de Stijl form and Frank Lloyd Wright concepts. Few unbuilt buildings surpassed them in the variety of ideas and in their influence on the development of architecture.

Werkbund exposition of 1927

This influence was apparent at the first postwar Werkbund exposition at Weissenhof near Stuttgart in 1927. The exhibition consisted of a housing demonstration project planned by Mies, who had by then become the Werkbund's vice president. Europe's 16 leading modernists, including Le Corbusier and Mies himself, designed various houses and apartment buildings, 33 units in all. Weissenhof demonstrated, above all, that the various architectural factions of the early postwar years had now merged into a single movement—the "International Style" was born. Though not a popular success, it was a critical one, and Europe's elite suddenly began to commission modern villas, such as Mies's Tugendhat house (1930) at Brno, Czechoslovakia. Intellectuals who could not afford villas took to the new steel furniture and interior designs Mies, architect and designer Marcel Breuer, and designer Mart Stam had displayed. But the exhibit did little, if anything, to advance the Werkbund's social purpose. None of the buildings designed for low-cost prefabrication and industrial production was actually prefabricated or mass-produced. The problem, it seemed, was not design alone, but lack of popular demand for the new style. Yet Mies held fast to his belief that the path of industry "must lead from quantity towards quality—from the extensive to the intensive" so that "industry and technology will join the forces of thought and culture."

Technology and culture did merge in the German Pavilion (also known as the Barcelona Pavilion), which is indeed best described as an architectural statement of quality, thought, and culture. The pavilion was commissioned by the German government for the 1929 International Exposition at Barcelona. It exhibited nothing but itself—a sequence of marvelous spaces on a 175 by 56 foot (53.6 by 27 metre) travertine platform, partly under a thin roof, and partly outdoors, supported by chromed steel columns. The spaces were defined by walls of honey-coloured onyx, green Tinian marble, and frosted glass and contained nothing but a pool, in which stood the sculpture of a nude by Georg Kolbe, and a few of the famous chairs Mies had designed for the pavilion. It was art for art's sake and luxurious art at that.

Appointed director of the Bauhaus

In 1930, Mies was appointed director of the Bauhaus, which Gropius had brought from Weimar to Dessau in 1925. Between Nazi attacks from outside and left-wing student revolts from within, the school was in a

state of perpetual turmoil. Though not cut out to be an administrator, Mies soon won respect as a stern but superb teacher. "If Mies doesn't make good architects of us, he'll at least teach us to judge what good architecture is," one of his American students wrote home at the time. When the Nazis closed the state-supported school, Mies tried for a few months to continue it on a private basis in an empty telephone factory in Berlin. But modern design was as hopeless a cause in Hitler's totalitarian state as was political freedom. Mies announced the end of the Bauhaus before the Nazis could close it.

Mies in America. Four years later, in 1937—again after working mainly on projects that were never built—Mies moved to the United States. Philip Johnson, who had met him at the Bauhaus and was an ardent admirer, had arranged for a commission, a country house in Wyoming. It was never built either. Soon after he emigrated Mies met the architect John Holabird who was looking for a man to head the School of Architecture at Chicago's Armour Institute (later the Illinois Institute of Technology). Mies demanded "a completely free hand and $10,000 a year." He got a free hand but a lower salary. By the time the age limitation forced him to retire in 1958 the school was world-renowned for its disciplined teaching methods as well as for its campus, which had been designed by Mies. A cubic simplicity marked the buildings that could easily adapt to the diversified demands of the school. Exposed structural steel, large areas of glass reflecting the grounds of the campus, and a yellow-brown brick were the basic materials used.

Design of Illinois Institute of Technology

Also in Mies's retirement year the Seagram Building on New York's Park Avenue, a skyscraper office building with a glass, bronze, and marble exterior, which Mies designed with Philip Johnson, was dedicated. The International Style, with Mies its acknowledged leading master, seemed at its zenith. The United States in the '50s had a faith in material and technical progress that seemed similar to the earlier German notion of *Gesamtkultur*. It was forgotten that only eight years earlier the magazine *House Beautiful*, voicing widespread unease, had attacked Mies's first buildings in America, specifically the house for Dr. Edith Farnsworth in Plano, Illinois, as a sinister threat to the American way of life, and Frank Lloyd Wright had even hinted that the men of the "International Style" might not be altogether "wholesome people." But now Henry Luce, the publisher of *Time*, *Life*, and *Fortune*, proclaimed that "the twentieth century revolution in architecture has been accomplished and it has been accomplished mainly in America." Miesian glass boxes, largely realizations of those paper projects of the post-World War I days, appeared all over the land, indeed all over the world.

Soon, however, the pendulum of taste and conviction swung back. Steel bones and glass skins seemed no longer a new beginning but the end of architectural creativity. "We must explore and expand the horizons of our architecture," said Eero Saarinen, perhaps the most promising of the younger generation of modernists. And he added: "In this sense, I align myself humbly with Le Corbusier and against Mies van der Rohe." Even Philip Johnson, the faithful disciple, turned away from the master's single-minded devotion to simplicity to a more romantic historicism.

Late work. Mies himself remained unperturbed. Busier than ever in his life, he remained faithful to the tenets he had developed in the 1920s and created beautiful buildings with them—among them the Chicago apartment houses (1949 and 1951); Lafayette Park, a housing development in Detroit (1956–63); the One Charles Center office building in Baltimore (1963); a museum in Houston (1958); the Federal Center in Chicago (1964); the Public Library in Washington, D.C. (1967); and, most Miesian of all, the Gallery of the Twentieth Century in Berlin, dedicated in 1968.

A heavy man, badly plagued by arthritis, he had come to look, as Peter Blake, one of his biographers, put it, like a granite monument of himself. Divorced in 1930, he lived alone in a spacious apartment in an old building near Lake Michigan in Chicago until his death on August 17, 1969. His office was in a nearby old loft building. He loved his splendid collection of paintings by Paul Klee, smoked expensive cigars, wore exquisite clothes, and read German philosophers. Though he seemed hard to approach, those who did found him friendly, utterly unpretentious, and cheerfully apt to reminisce.

Last years

To the charge that his buildings are cold, he replied that though milk can be cold or warm, there could hardly be a cold or warm architecture. He saw no necessity, he said, "to invent a new architecture every Monday morning." He disdained any style that was nothing more than a personal statement of the architect, a monument to his own ability. Mies's architecture was meant to be anonymous. Architecture, he said, "is the will of an epoch translated into space—living, changing, new." Mies never considered himself more than an instrument of that will.

MAJOR WORKS

Perls house, Berlin (1911); Kröller house project, The Hague (1912, not executed); Urbig house, Neubabelsberg (1914); Liebknecht-Luxemburg Monument, Berlin (1926); Wolf house, Guben (1926); cantilevered tubular-steel chair (1926); Weissenhof flats, Stuttgart (1927); German Pavilion, Barcelona Exposition (1929); Barcelona chair (1929); Tugendhat house and chair, Brno, Czechoslovakia (1930); Illinois Institute of Technology, Chicago (1939–41); Farnsworth house, Plano, Illinois (1946–50); Promontory Apartments, Chicago (1949); Apartment houses, Lake Shore Drive, Chicago (1949–51); Seagram Building, New York (1956–58); Bacardi Building, Mexico City (1961); Federal Center, Chicago (1964); Gallery of the Twentieth Century, Berlin (1968); New National Gallery, Berlin (1968); I.B.M. Building, Chicago (1971).

BIBLIOGRAPHY. The most important collection of drawings is in the Ludwig Mies van der Rohe Archive of the Museum of Modern Art in New York City. A selection of these has been published as *Mies van der Rohe: Drawings in the Collection of the Museum of Modern Art,* introduction and notes by LUDWIG GLAESER (1969). Bibliographies are contained in PHILIP C. JOHNSON, *Mies van der Rohe,* 2nd ed. (1953), a brief, scholarly, and most readable book that first introduced Mies to America; ARTHUR DREXLER, *Ludwig Mies van der Rohe* (1960), a comprehensive and critical analysis of the architect's work; PETER BLAKE, *The Master Builders: Le Corbusier, Mies van der Rohe, Frank Lloyd Wright* (1960, each of these essays published separately in 1964), the most detailed, personal, and authoritative biography to date; and WERNER BLASER, *Mies van der Rohe: The Art of Structure* (1965), a fairly complete collection of photographs and drawings of Mies's work with his own brief introduction. Information may also be found in publications on 20th-century architecture, notably S. GIEDION, *Space, Time and Architecture: The Growth of a New Tradition,* 5th ed. (1967); and on the Bauhaus, notably *Bauhaus and Bauhaus People* (1970), a collection of personal reminiscences by former Bauhaus members in which Mies is recalled by HOWARD DEARSTYNE.

(W.V.E.)

Mi Fei

One of the dominating figures of Chinese art, Mi Fei's (in Pin-yin romanization also Mi Fei) erudite yet natural calligraphy and his landscape style were complemented by an eccentric personality and caustic tongue. His far-reaching influence owes as much to his scathing critical writings and lively connoisseurship as to his artistic activities. For later generations he came to represent the ideal type of scholar-gentleman who combined artistic excellence with a leisured independence of the material world.

Mi Fei's original name was Mi Fu. The second character of his name, Fu, was later changed to another, originally homophonous character, the reading of which was subsequently confused with a third character read as Fei. This is the name by which he is known today. In accordance with Chinese practice, Mi Fei also used many other names and nicknames to sign paintings, pieces of calligraphy, or articles; among the best known are Yüan-chang, Hai-yüeh Wai-shih, and Hsiang-yang Man-shih.

Mi Fei was born in 1051 of a family that had held high office in the early years of the Sung dynasty (960–1279). Because his mother had been the wet nurse of the emperor Ying Tsung (reigned 1064-67), he was brought up within the Imperial precincts, mixing freely with the Im-

"Tower of the Rising Clouds," attributed to Mi Fei, colour on silk. In the Freer Gallery of Art, Washington, D.C.
By courtesy of the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.

perial family. As a young boy, Mi Fei showed precocious talent, particularly in calligraphy. Although he expressed a distaste for the formal lessons prescribed for a future official, he displayed a lively intelligence by his quickness in understanding learned argument, an aptitude for excitingly original poetry, and an ability in painting and calligraphy.

His professional career

In later life Mi Fei followed a checkered official career with frequent changes of post. He began as a reviser of books in the Imperial library and subsequently served in three posts outside the capital of K'ai-feng, in the Honan Province. In 1103 he was appointed a doctor of philosophy and was briefly military governor of Wu-wei in the province of Anhwei. He returned to the capital in 1104 to serve as a professor of painting and calligraphy, taking this opportunity to present to the Emperor a painting by his son, Mi Yu-jen. He then undertook the position of a secretary to the Board of Rites before setting out on his final appointment as military governor of Huai-yang, in Kiangsu Province, in which post he died in 1107, at the age of 57. He was buried in Tan-t'u, in Kiangsu Province, under an epitaph written for him by Mi Yu-jen.

Mi Fei was married and had five sons, of whom only the two eldest survived infancy, and eight daughters.

Mi Fei's writings were extensive. A collection of his poetry, the *Shan-lin chi*, has been lost, but still existing are the *Pao-chang tai-fang lu* ("Critical Description of Calligraphics in Mi Fei's Collection"), and *Hua shih* ("Account of Painting"), containing records of his own and others' collections of paintings, essays on aesthetic history, and criticism of paintings. There also exist three posthumous collections of his writings, *Pao-chin ying-kuang chi* ("Literary Works of Mi Fei), *Hai-yüeh ming-yen* ("Remarks on Calligraphy"), and *Hai-yüeh t'i-pa* ("Inscriptions and Colophons by Mi Fei").

Throughout Mi Fei's life, the workings of a highly individualistic temperament were evident. This was manifested in his fastidious attention to cleanliness, his preference for the clothes of ancient Chinese dynasties, and his love for strange rocks and ink stones, which he collected. To this eccentricity of personality must also be attributed his comparative failure in his official career. Despite his undoubted ability, his friendship with the leaders of both principal political factions, the poet-statesman Wang An-shih and Su Tung-p'o, and his family connections, he never rose to any office higher than that of a second-class secretary to the Board of Rites. The reason, his epitaph states, was that "he could not endure to bow and scrape before the world." An example of this upstanding independence is his attitude toward art collectors; he divided them into two classes, busybodies and connoisseurs—utterly deriding the former as illiterate and intent simply on ("Inscriptions and Colophons by Mi Fei").

His calligraphic style

As an artist, Mi Fei is best known for his calligraphy and landscape painting. His calligraphic style rejected what was inaccessibly unusual or flamboyant and was formed by patient and catholic study of the great Chinese calligraphers of the past. His theoretical writings in the *Shu shih* and *Hai-yüeh ming-yen*, which contain some of the most penetrating remarks written on the subject of calligraphy, indicate that Mi Fei believed in a respect for historic styles suffused with the calligrapher's own creative talent. Most of all he valued spontaneity and self-expression, eschewing the contrived and saccharine.

Mi Fei's paintings, such as "Tower of the Rising Clouds" (Freer Gallery of Art, Washington, D.C.) represented a break with the past. Before the Sung dynasty, landscape painting in China had depended essentially on line for its description of the world. Mi Fei, however, concerned with depicting the misty rivers and hills of the lake district of Honan Province, introduced a technique of extremely moist washes and horizontal texture strokes. Later known as "Mi dots" (*Mi tien*), this technique rendered a vivid impression of that rainy and cloud-clad Chinese region. This technique of "splashed ink" (*p'o-mo*) was Mi Fei's own invention; it attracted enthusiastic contemporary attention and remained a compelling influence throughout the history of Chinese painting.

MAJOR WORKS

The few well-documented works by Mi Fei are now all lost. Imitations and copies of works exist in the following collections: Yale University Art Gallery, Moore Collection, New Haven, Connecticut; Ōsaka Municipal Museum of Fine Arts, Ōsaka, Japan; Ogawa Collection, Kyōto, Japan. Extant paintings reflecting Mi Fei's style: Hanging scroll (Nakamura Collection, Tokyo); "Auspicious Pine Trees in the Spring Mountains" (National Palace Museum, Taipei, Taiwan); "Tower of the Rising Clouds" (Freer Gallery of Art, Washington, D.C.).

BIBLIOGRAPHY. LIN YU-T'ANG, *The Chinese Theory of Art* (1967), translations with introductions of aesthetic writings by various Chinese painters and critics, including Mi Fei; *Shodō zenshū*, vol. 15, pp. 26–36 (1954), reproductions of Mi Fei's calligraphy; OSVALD SIREN, *Chinese Painting: Leading Masters and Principles* (1956), critical biography and reproductions of some of Mi Fei's works; *Sung Mi Fei mo-chi*, ed. by the National Palace Museum, Taipei (1970), reproduction of Mi Fei's calligraphy; R.H. VAN GULIK, *Chinese Pictorial Art As Viewed by the Connoisseur* (1958), translations of important Mi Fei texts dealing with connoisseurship and seals; and *Mi Fu on Ink Stones* (1938); NICOLE VANDIER-NICOLAS, *Le Houa-ché de Mi Fou (1051–1107) ou Le Carnet d'un connaisseur à l'époque des Song du Nord* (1964), annotated translation of Mi Fei's *Hua shih* ("Account of Painting"); and *Art et sagesse en Chine: Mi-Fou (1051–1107) peintre et connaisseur d'art dans la perspective de l'esthétique des lettrés* (1963), a biographical study of Mi Fei; WAI-KAM HO, "Mei Fei," *Encyclopedia of World Art*, vol. 10, col. 84–90 (1965), an important biographical article, with an excellent bibliography.

(C.S.-y.)

Migrant Labour

Workers who move about systematically, seeking and engaging in seasonal temporary employment without becoming residents of the areas in which they work, are called "migrant labourers." Throughout the world, the wages, working conditions, and standard of life of migrant workers are low in comparison with other groups of workers, almost always the lowest. The union- and government-established labour standards available to industrial and regular agricultural workers generally do not exist for migrant labour; and where there are statutes their implementation is limited and irregular. Migrant labourers are the rural equivalents of the urban poor.

Migrant labour is primarily but not exclusively agricultural; it generally involves harvesting activities and is mostly manual, repetitive, easily learned, and demanding of almost no skill. Migrant workers pick cotton, cocoa beans, grapes, tea or tobacco leaves. They cut sugar cane and pineapples and gather artichokes, beans, carrots, sugar beets, or other cash crops. They also harvest corn, small grains, or wheat. Because all of these activities require physical strength and call for much bending and reaching, migrant labour is also called "stoop labour."

CHARACTERISTICS OF MIGRANT LABOUR

Criteria of migrant labour

Migrant labour must be distinguished from other forms of employment with similar characteristics and similar problems. Labour migration in Europe—often involving the crossing of national boundaries in search of employment or better employment—is different, since it usually involves a permanent or quasi-permanent resettlement of the worker and his family. Casual labour, in which a worker shifts, often daily, from employer to employer, is also different, because the casual labourer does not change his residence. Seasonal or part-time labour, offering the worker employment only on a seasonal or occasional basis, is again different, in that it does not involve either a change of residence or frequent changes of employers.

Association with industrialization. Migrant labour is associated with the developed or "middle" stages of industrialization and its technology and organization. While the demand for migrant labour stems, in part, from the seasonal nature of worker requirements in agriculture and related industries, it is not a phenomenon characteristic of agricultural societies. It was probably unknown in antiquity and is not significant today in the underdeveloped economies of Asia and Africa; high levels of demand for migrant workers are related to rapid increases in farm output, decreases in permanent farm employment, long periods of rural migration, increasing urban wages, and fairly rapid mechanization and commercialization of agriculture. The demand for migrant labour is particularly high in the case of market farming, where production is increasing or where numerous small farmers specialize in one or two cash crops and are joined in production or marketing associations.

The propensity to work as a migrant labourer stems from the unfavourable economic and social conditions in which the worker finds himself at what is called his "home base." Technological displacement, radical changes in sharecropping systems, long business recessions, drought and crop failures, and various forms of discrimination are factors inducing workers to enter the migrant stream. Lack of education, industrial accidents, and ill health also force many to become migrant. The supply of migrant labour thus develops in response to the push of need, not from a romantic desire to wander. In a minority of cases, particularly those in which a poor region, state, or nation is adjacent to a much richer one, workers respond also to the pull of better opportunities; in spite of its misery and degradation, migrant labour appears to them as a way to a higher standard of life in the future.

As national economies mature and move through the secondary (mostly industrial) stage to more developed stages, both the demand for and the supply of migrant labour decrease. The demand for it decreases because of mechanization; in the U.S., for example, the harvesting of wheat and cotton that required the employment of many migrants before World War II is now largely mechanized and done by regularly employed farmhands. The supply decreases as higher real standards of living and greater employment possibilities make migrant work less attractive. For a small number of employed workers in affluent areas, migrant labour provides secondary (or additional) employment opportunities.

In mature economies migrant labour contributes little to total agricultural output and a negligible amount to nonagricultural output. Nevertheless, the availability at the right time and place of migrant workers can be crucial, since without them large crop losses may occur. This may spur private employers, planning authorities, and governments to take remedial actions to improve the conditions of migrant labourers in order to assure their continued availability. If native migrant workers are in short supply, the gap may be filled by the temporary—legalized or tolerated—migration of foreign workers from poorer countries. In the Communist countries, where the use of migrant foreign labour is politically unacceptable, the flow of migrant workers is supplemented with organized brigades of urban youth encouraged to contribute their labour voluntarily to the common good.

The migrant flow. Most migrant workers move in well-established patterns or "streams." The migrants usually flow from the south to the north (in the Southern Hemisphere, however, they move in the opposite direction), as geography makes for later harvests in northern latitudes; also, the poorer areas with the labour surpluses tend to be in the south. Migrants make detours and side trips as they hear of better or new opportunities. The various "crews," families, or individuals that make up the migrant stream do not all follow the same pattern: some move from place to place, others make a big journey to a set of places with known jobs, while others make only one move and then go back to "home base."

ECONOMIC AND SOCIAL CONDITIONS OF MIGRANT WORKERS

In all employment relationships the matching of unemployed workers and job vacancies is a haphazard and complicated process, but the labour market for migrant workers is notably disorderly. One basic reason is that the employment relationship is ephemeral. To the worker, the identity of the employer is hazy, even unknown. To the employer, the typical workers have no individuality as he thinks of them often as alien drifters, "wetbacks," "Okies," etc. With minor exceptions, employers do not systematically recruit migrant workers and have no regular personnel specialists or procedures. Even though most labourers migrate year after year, they do not enjoy reemployment rights, are not organized in unions, and have little access to systematic means of job seeking; they tend also not to exploit the available services of public placement or other governmental agencies, limited as these are.

The role of the contractor

Middlemen, job brokers, labour contractors, crew leaders—*patrones* in Latin America, *sirdār* in the Asian subcontinent—arise out of the fundamental disorganization of the migrant labour market. Labour contractors, in addition to bringing the workers together, transporting them, supervising them, and dispensing their pay, search out the employers and negotiate wages and working conditions. In some areas the contractors have made themselves so essential that without them workers cannot find employment and employers cannot get workers.

Living conditions and wages. The conditions under which migrant labourers live are abysmal. Their misery has been formally documented by investigative bodies down the years: the U.S. Industrial Commission of 1901, the Royal Commission of Labour in India of 1933, the U.S. President's Commission on Migratory Labor of 1951, and others. Novelists, moved by the migrant workers' plight, have aroused sympathy with such masterpieces as John Steinbeck's *The Grapes of Wrath* or Mulk Raj Anand's *Two Leaves and a Bud*. Social scientists, particularly economists, however, have tended to ignore the issues posed by migrant labour. Governmental statistics on migratory labour are scarce, often contradictory, and approximate at best.

Migrant workers are often cheated even of the low

wages they earn. They work long hours and under exacting requirements. Total employment is always limited; in a year most migrants work less than one quarter of the working days. Child labour is widespread, and the children that do not work are not adequately provided for; they often do not go to school, as in most places schools are open only to residents. Housing amounts to hardly more than a roof over the head, often tar or tent. Tuberculosis, infant and maternal mortality, dysentery, enteritis, smallpox, and typhoid are all more prevalent among migrant labourers than among the general population. Literacy levels, social cohesion, and political participation rates are always exceptionally low among migrant workers, while crime and all other indexes of unacceptable social behaviour are high. Contributing to this picture of misery is the fact that the migrants, whether national or foreign born, are fundamentally alien to the community in which they work. The local population, if eager to see them come, is even more eager to see them go. They have difficulty in gaining access to local health and social services, are often deprived of their rights, do not enjoy easy recourse to the courts, and are abused by exploiters and sometimes the local police.

Difficulties of regulation. The fact that the migrant is "here today and gone tomorrow" makes the regulation of his working and living conditions difficult. Local, state, and national authorities generally acknowledge the existence of serious problems associated with migrant labour, yet in the face of these, generally claim they are unable to deal with them effectively. The International Labour Organization has consistently prodded governments to expand their legislation in this area. Efforts are being made to develop regulatory standards protecting workers during their journeys, to discourage migratory movements of undesirables, and to stabilize workers and their families near their place of employment.

The task of providing migrant workers with governmental protection and assistance is full of complex difficulties. Regulatory legislation ideally should be applied at the "home base" of the migrants as well as along the stream of migration and at the place of employment. But advances in one area often create difficulties in another due to uneven social and economic conditions. On the one hand, mechanization and permanent migration is reducing the number of migrant labourers, while, on the other, relocating them is difficult because their low educational attainments and lack of skills force them into urban poverty. Where legislation has been passed and improvements have been achieved, the major effect has been to force employers to provide more housing and sanitary facilities and to regularize the activities of the labour contractors.

PATTERNS OF MOVEMENT

The United States. In the United States in the 1960s about 400,000 people yearly were occupied in migrant work. In the previous decade the number was about 600,000 yearly; in 1940 there were 1,000,000; and it is possible that in the middle 1920s as many as 2,000,000 workers were in the migration streams. There are not many reliable statistics. Interpretation of scattered data suggests that the following trends and relationships have occurred since World War I: (1) the migrant labour population has been declining, albeit at a decreasing rate; (2) migrant agricultural workers have tended to become a larger proportion of all agricultural (nonfamily) workers; (3) foreign workers—while their participation varies with the rulings of the federal immigration administration and with requirements of foreign policy—constitute an increasing proportion of all migrant workers; and (4) the size of the national migratory labour force varies inversely, and the foreign labour force directly, with improvements in the total level of employment.

In the 1960s more than half of the migrant workers were 25 years or younger; fewer than one in ten was 55 years or older. Three out of four migrant workers were males, with males outnumbering females in all age groups. Almost half of the female migratory workers were between the ages of 25 and 54. In terms of areas of permanent residence a little less than two out of four migrant workers originate in the South. Over 55 percent of all migratory workers had eight years of education or less.

Three major patterns characterize the movement of migrant workers in the U.S. (1) On the East Coast, migration begins in winter when workers from Southern states move south to Florida to work on the citrus crop. From there, joined by others coming from Texas and Puerto Rico, they move north along the coast ending up in August and September gathering tomatoes and other vegetables on Long Island and potatoes as far north as Maine. Negroes constitute a large proportion of those on the East Coast. (2) The largest stream of migrants flows north and west from Texas in the spring to the north central, mountain, and Pacific Coast states. The crops they harvest are fruits and vegetables, sugar beets, and some cotton. The workers are natives of Mexican descent with whom some "wetbacks" (immigrants from Mexico) join; they travel mostly with their families and end work early in December. (3) A large number of Spanish-Americans work on harvesting vegetables, moving from Southern California northward through the Pacific Coast states. Thus some migratory farm work is done in almost every state, with most labourers coming from Texas and the Southeast. About half of the man-days of such employment occur in California, Michigan, Texas, and Florida.

These migrant workers suffer from unemployment, inadequate housing, little education, exclusion from normal community life, and the denial of services that others take for granted. In the middle 1960s they grossed an annual average income of about $1,700, working an average of 150 days if they worked in nonfarm activities as well. Those who worked only as migrant farm workers averaged $800 for about 82 days of work.

One of every five poor people in the United States is a migrant worker and entitled to the various governmental antipoverty manpower and social programs. Yet, an analysis of federal antipoverty expenditures from 1965 through 1968 shows that only one out of every 50 dollars was spent to help migrant workers, a pattern suggesting that there is considerable truth in the oft-quoted statement that more money is spent on migratory birds than on migratory labour.

In spite of this, it is probably true that the 1960s saw more improvements in the living and working conditions of migrant workers in America than any previous decade. A number of states and localities established special migrant committees to implement and expand legislation. "Crew registration laws" were increasingly adopted by the states, and their implementation is having beneficial effects. Most important, after years of inactivity attempts were being made by labour unions to organize migrants, with some success in California and Wisconsin.

Western Europe. Migrancy as described above does not involve very large numbers in Europe because of the longer history of urbanization. Migrant agricultural workers are still active in the rice harvest of northern Italy and in some of the vegetable harvest of Portugal and France.

The European migrant worker frequently moves across national boundaries in response to seasonal fluctuations in the labour demand of specific industries. This type of migration rarely involves the workers' families and generally calls for only one move. Thus Italian workers go to work in Swiss hotels, Polish workers go to mine English coal, etc. The problems of European migrants are less acute than those described above; they generally involve difficulties in acculturation and in qualifying for national health and social service programs.

Africa. Migrant labour is one of the most important problems of the developing nations of Africa, where migrancy is clearly an intermediate stage of the movement from rural agricultural to urban industrial life. The phenomenon does not conform to the general characteristics described earlier. Workers migrate to urban rather than to agricultural employment; they work with one rather than many employers; and they often stay in one job as long as three years.

Migrancy in Africa is very clearly a response to the pull

of expanding industrial opportunity and the lure of city life rather than to a push away from the poverty of rural life; it is thus a step in the process of building an industrial labour force. The worker who moves does so alone and with considerable reluctance. He often goes because he needs cash for a particular purpose (*e.g.*, the "bride price"). He is not prepared to give up his contacts with the village, where there is security in old age, sickness, or periods of unemployment. African migrancy is nonetheless quite similar to the migration that occurs in the middle stages of industrialization. In the regions where the migrant workers come from one finds the decline of traditional subsistence agriculture, changes in ownership patterns, and drought; one rarely finds workers displaced by mechanization of production.

Migrant labour is also to be found in many other nations. India has probably more migrant workers than any other Asian country; these are involved mainly in the harvesting of tea, cotton, and rice. In Australia and the southern nations of Latin America migrants work on ranches more often than on farms and are engaged in wool shearing and meat processing.

BIBLIOGRAPHY. There are no classic or standard works on this subject. Most of what is known is derived from the reports of investigative and governmental commissions. For the U.S. the most recent and the most complete is *Migratory Labor in American Agriculture*, Rept. of the President's Commission on Migratory Labor (1951). More recent data may be found in *Domestic Migratory Farmworkers, Personal and Economic Characteristics*, Agricultural Economic Rept. No. 121 (1967), and other serial publications of the Department of Agriculture. For information on other countries, the reports of the International Labour Organization are important. In particular see *Migrant Workers*, Rept. 5 of the 37th Session (1954) and the 38th Session (1955) of the International Labour Conference; and *African Labour Survey* (1958), for good information on African migrancy.

For greater details one should turn to the series of reports of the Subcommittee on Migratory Labor of the U.S. Senate Committee on Labor and Public Welfare. They are available for almost every year since 1955. The journal literature is thin and generally narrowly focussed. The following are examples of the type of work available: E.P. EDWARDS, "Children of Migratory Agricultural Workers in the Public Elementary Schools of the U.S.," *Harvard Educational Review*, 30:12–52 (Winter 1960); and J.C. GLASS, "Organization in Salinas," *Monthly Labor Review*, 91:24–27 (June 1968).

(O.A.O.)

Migration, Animal

The term migration may be applied to the cyclic or periodic travels of an animal if it returns eventually to the original place of departure. Such movements can be contrasted with emigrations, which involve a change in location not necessarily followed by a return journey; invasions or interruptions, both of which involve the appearance and subsequent disappearance of great numbers of animals at irregular times and locations; and range expansion, which tends to enlarge the distribution of a species, particularly its breeding area.

The migration cycle is often annual and thus closely linked with the cyclic pattern of the seasons. The migration of most birds and mammals and many of the fishes are on a yearly cycle. In many cases (*e.g.*, salmon and eels) animals with a relatively long life-span return to the place of birth in order to reproduce and eventually die. In other cases, as in certain invertebrates, where the animal has a relatively brief life-span and reproduces rapidly, migrations may not occur in every generation. The daily movements of certain fishes and invertebrates have also been called migrations because of their regular occurrence. This type of movement, however, is not to be confused with migration in the strict sense.

Most migrations involve horizontal travel. The distance traversed may be a few miles or several thousands of miles.

Some migrations take a vertical direction and involve no appreciable horizontal movement. Certain aquatic animals, for example, move from deep water to the surface according to the season. Certain birds, mammals, and insects migrate altitudinally in mountainous areas, going from the upper zones, where they breed, to the foothills or plains during seasons when the weather is severe and unfavourable. Such vertical travels involve essentially the same type of environmental change as horizontal, or latitudinal, migrations over long distances.

SURVEY OF MIGRATIONS IN ANIMALS

Lower invertebrates. Many marine invertebrates travel considerable distances during certain seasons. A large proportion of them, however—particularly planktonic organisms, plant and animal aquatic drifters—do not travel deliberately but are carried by ocean currents. Planktonic organisms also travel vertically in a daily rhythm. Very small or microscopic animals remain at great depths during the day and rise at dusk, concentrating in the upper layers of water during the night. Their predators, particularly fishes, follow them in their cycle. The daily activity of pelagic birds (*i.e.*, those living on the open sea rather than along the shore), such as petrels and shearwaters, which feed on planktonic crustaceans and squids, follows this same rhythm. Planktonic migration

A seasonal change of habitat, analogous to migration, is made by some Polychaeta (sandworms). Along the coast of Europe, clam worms (*Nereis*) live during the colder months in rock crevices and among algae. During the summer, however, they become planktonic and swim out some distance from the coast, where reproduction occurs. In the South Pacific, near Samoa and Fiji, the palolo worm (*Eunice viridis*) lives among coral reefs, where it develops posterior segments filled with genital (reproductive) cells. These are cast off, and the worm rises to the surface. The phenomenon occurs regularly on the first day of the last quarter of the October–November moon.

Some of the best known migrations among the invertebrates occur in crustaceans during the reproductive period, when some of them travel as far as 150 miles (240 kilometres). Generally in the crabs, females move into shallow coastal waters to mate and to lay their eggs. After the eggs have been laid, the females return to deep water.

Some freshwater crabs, such as the Chinese crab (*Eriocheir sinensis*), after remaining for three to five years in freshwater, migrate to brackish water, where mating occurs. Females with eggs externally attached then travel to the sea and remain a few miles offshore for several months during winter. The following spring they enter shallower water near the shore. Here the eggs hatch. Young crabs spend a year in brackish water and migrate upstream the following spring, settling in freshwater and growing to maturity.

Some crabs, such as robber crabs (*Birgus*) and land crabs of tropical regions (*Geocarcinus*), have adapted to life on land. They migrate to the sea to reproduce and then return inland and are followed at a later time by the young.

Insects. Migration among the insects is best known in locusts and butterflies; a great number of other insects, however, including some of the smallest, are migrants. Broadly speaking, insect migration is of three types—some insects emigrate on one-way journeys to breed; others migrate from a breeding area to a feeding area; still others migrate from breeding areas to hibernation sites. Types of insect migration

In the first type, adults with a life-span limited to a single season emigrate from their breeding site, deposit their eggs, and die. Such migratory flights can be very short or very long but, because they are always one-way journeys, cannot be regarded as migration in its strictest sense. The best known examples of such flights are those of locusts, particularly the desert locust (*Schistocerca gregaria*), a species found in tropical and subtropical countries of the Old World. The migratory, gregarious form arises from the solitary form as a result of various conditions—*e.g.*, lack of food, crowding.

The desert locusts breed only when and where seasonal rains permit; as a consequence of climatic conditions, therefore, the insects migrate from one breeding area to another. If the available food decreases and the numbers of insects increase too drastically in a particular area, migratory locusts develop. They differ from nonmigrato-

ry forms in colour, structure, behaviour, and physiology. Swarms numbering up to 10,000,000,000 individuals periodically invade territories in Africa, southwestern Asia, and southern Europe, covering areas as large as 400 square miles.

Other migrant insects travel beyond the limits of their breeding range and either die or return to the breeding range. The painted lady butterfly (*Vanessa cardui*) "migrates" in the spring, when its population becomes too large for local conditions, from the peninsula of Lower California in Mexico to the Mojave Desert in Southern California. Eggs are laid in the desert region, but the species does not become permanently established there and makes no return flight. Such movements are known in about 250 species of butterflies.

In the second type of migration—migration in the strict sense—insects migrate from a breeding area to a feeding area. In the feeding area the females develop mature ovaries and then return to lay their eggs in the place from which they came or a similar region.

Cockchafers (*Melolontha melolontha*), a species of beetle, leave the site where they emerge as adults and move to a feeding area, generally in a forested region, where maturation of the eggs takes place. They then return to the area where they developed from eggs and lay their own eggs. This process may be repeated several times during the life of the insect. Although the distances covered by the cockchafer may not be great, the regularity of the phenomenon is characteristic of true migration.

In the third type of migration, insects travel from their breeding areas to places where they hibernate or aestivate —*i.e.*, pass the summer in a dormant state. The place of hibernation or aestivation may be outside the zone where climate permits breeding. The following season, they return to the breeding place and lay their eggs. This type of migration, which can involve great distances, is made by insects with unusually long life-spans. The lives of these insects include a diapause, or period of dormancy during which development is suspended.

In warm countries the coccinellids, a family of beetles, and certain moths leave the hot lowlands and migrate to the mountains, where they aestivate and later hibernate. In spring they return to the breeding areas.

Migration of the convergent ladybug

One coccinellid, the convergent ladybug (*Hippodamia convergens*), lives in valley regions of California, where the eggs hatch in March or April and develop into adults one month later. In early summer they migrate to the mountains, particularly to the Sierra Nevada, where they may lay eggs if food is abundant and the weather warm. Generally, however, the adults gather in clusters and remain inactive until October, when rains initiate a period of activity, after which they travel to lower altitudes and hide in forest litter, passing the winter in a state of dormancy. As many as 30,000,000 ladybugs may congregate on a quarter acre. In spring they mate, fly back to the valleys, lay their eggs, and die.

The flight before diapause of some insect groups may cover thousands of miles. In North America, the monarch butterfly (*Danaus plexippus*) is a well-known example of a wide-range migrant with an extensive breeding range. The number of generations varies with the latitude; as many as five generations may occur each year in the south and only one in the north. In summer the insects travel northward to Hudson Bay. Individuals of the last generation of the year migrate southward in autumn to Florida, Texas, and California, where they hibernate after travelling nearly 2,000 miles (3,200 kilometres). They gather in sheltered sites, particularly on trees where they cluster on trunks and big branches. In spring part of the populations migrate back to the northern breeding areas. Some of the returning butterflies are members of the first generation that develops from the overwintered insects; others represent successive generations that develop as the insects progress toward more northern latitudes. The recapture of marked butterflies has revealed that they travel as much as 80 miles (130 kilometres) in one day. The longest distance recorded thus far for the complete flight of a migrant monarch butterfly is 1,870 miles (3,010 kilometres).

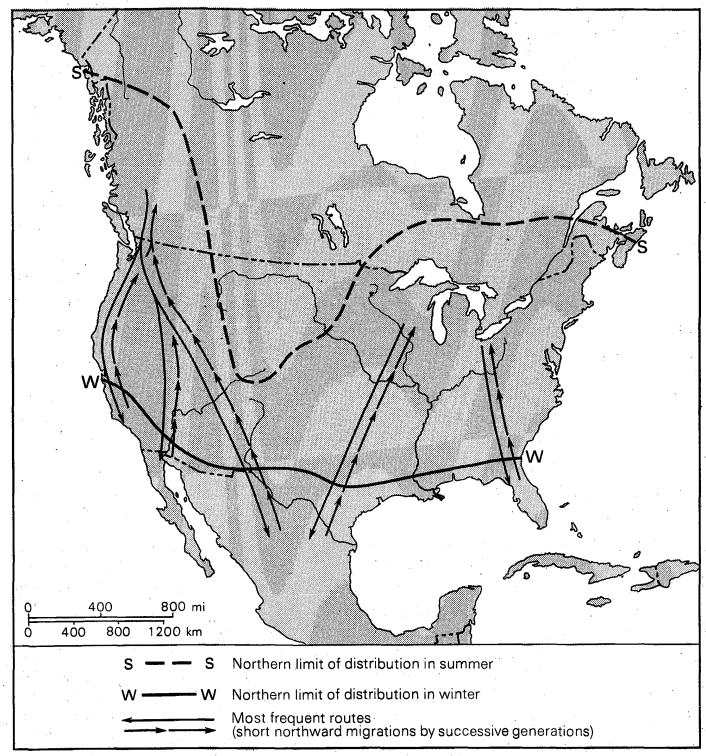


Figure 1: Migrations of the North American monarch butterfly (*Danaus plexippus*).
From M. Ricard, *The Mystery of Animal Migration*

Fish. Many species of fish wander annually through a particular area of the ocean. Some are true migrants, travelling regularly over great distances. Young fish usually leave the spawning grounds for areas where they develop into juveniles, before joining the adult stock at the feeding grounds. Adults move to the spawning grounds, then return to the feeding grounds. Migratory patterns of fish are related to oceanographic factors and to currents. Eggs, larvae, and young drift passively with the current, although migration of adult fish toward breeding grounds is usually against the current. Adult movements thus are directional rather than passive, and the fish respond to environmental conditions—*e.g.*, climate.

Three categories of migratory fishes can be distinguished: oceanodromous, anadromous, and catadromous.

Oceanodromous fish. Oceanodromous fish, which occur widely throughout the world's oceans, live and migrate wholly in the sea. They differ mainly from one another by the method and extent of their migration.

Herring migratory patterns

Herring (*Clupea harengus*), extensively studied because of their economic importance, are the best known of the oceanodromous type and can be classified into several populations, or local races, which do not mix freely. In addition, each has a particular migratory behaviour. In the North Sea, distinct groups spawn in different seasons and on different grounds: Buchan herring spawn in August and September off the coast of Scotland and migrate to the coast of southwestern Norway; Dogger Bank herring spawn in September and October in the central part of the North Sea and along the English coast and then migrate to the Skagerrak, an arm of the North Sea between Denmark and Norway; Downs herring spawn from November to January off the French coast, mainly between Dunkirk and Fécamp, then feed in summer in the middle and northern parts of the North Sea, sharing the feeding grounds with other populations. The diversity of migration and of reproductive seasons is closely connected with the annual cycle of oceanographic conditions in the North Sea.

Cod (*Gadus morhua*) have migration patterns similar to those of herring. The migrations of other fish cover even greater distances; in the Atlantic, for example, white tuna (*Germo alalunga*) are found in winter around the

Azores and the Canary Islands, where they spawn in spring. They then migrate northward to the Gulf of Gascogne and afterward to the waters around Iceland, arriving there in July. Populations of red tuna (*Thunnus thynnus*) occur throughout the Mediterranean Sea and the eastern Atlantic. In May and June they spawn in the western Mediterranean. During summer they spread northward to feed, finally reaching the Arctic Ocean. Similar migrations occur along the North American coast in the Atlantic and throughout the Pacific.

Anadromous fish. Anadromous fish live in the sea and migrate to freshwater to breed. Their adaptations to conditions of different habitats are precise, particularly with regard to salinity of the water.

Salmon (*Salmo, Oncorhynchus*) spawn in the cold, clear waters of lakes or upper streams. Eggs are laid in gravel beds. The young of the Atlantic salmon remain in freshwater for two to three years, sometimes as long as six; Pacific salmon sometimes migrate to the sea in their first year. Adult fish usually remain in the sea for two or three winters, but sometimes only one. Then, as grilse (adolescents) or as adults, they return to freshwater and spawn, after changes occur in colour and other external features. Some Atlantic salmon die in freshwater after a single spawning; others return to the sea.

The tagging of salmon has shown that European types may cross from Norway to Scotland, as well as the reverse. Pacific salmon are probably distributed over the Pacific Ocean and Bering Sea, between latitudes 45° N and 65° N with surface waters of 2° to 11° C (36° to 52° F).

Experiments in Canada and the United States, in which young salmon migrating to the Pacific have been tagged, have shown that a high proportion of the fish return to the river where they hatched. Tagging also has revealed that the same Atlantic salmon return year after year to a particular river. Adults reared from experimentally transplanted eggs return to the stream where they were hatched or grew, not to the stream where the eggs were laid. Aside from other means of orientation, such as reference to celestial features, topographical features are believed to play an important part in recognition of the original habitat. The sense of smell, or olfaction, however, has the most important role. Experiments have shown that migrating salmon are attracted to the waters of the stream in which they are going to spawn. Experiential imprinting at an early stage of development enables a grown fish to respond to waters that contain substances with a particular odour or that have a characteristic temperature.

Catadromous fish. Catadromous fish spend most of their lives in fresh water, then migrate to the sea to breed. This type is exemplified by eels of the genus *Anguilla*, numbering 16 species, the best known of which are the North American eel (*A. rostrata*) and the European eel (*A. anguilla*).

Eel migration

European eels and North American eels spawn in warm saline waters of the Atlantic, at depths of 400 to 700 metres (about 1,300 to 2,300 feet), in an area centred near latitude 26° N longitude 55° W called the Sargasso Sea. The pelagic eggs develop into leptocephali—transparent, leaflike forms with relatively small heads—that are carried by the Gulf Stream to the shallow waters of the continental shelves. When they are about two and one-half years old and about eight centimetres long (a little more than three inches), a metamorphosis occurs. The leptocephali are transformed into so-called elvers, which are bottom-dwelling, pigmented, and cylindrical in form. They arrive in coastal waters as glass eels and begin to swim upstream in freshwater streams in spring. Their migration upstream is spectacular, as the young fish gather by millions, forming a dense mass several miles long. In freshwater the eels grow to full size, becoming yellow eels. They live as such for ten to 15 years before changing into silver eels, with enlarged eyes; they swim downstream to the sea, return to the spawning grounds (Sargasso Sea), and die.

The migration of these eels is not entirely understood, particularly their return to the Sargasso Sea. It may be

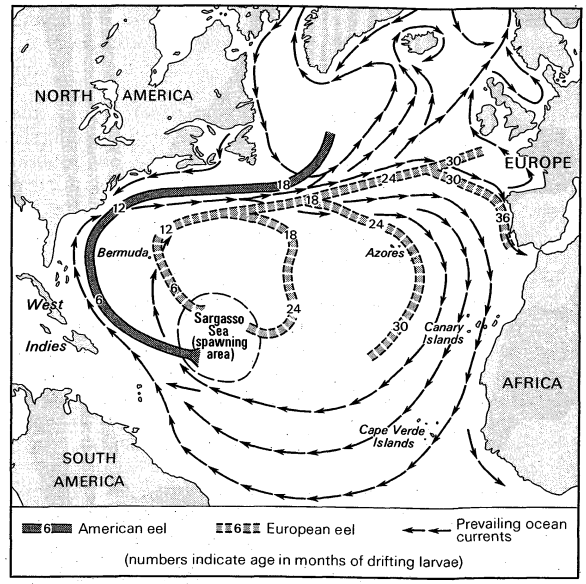


Figure 2: Migration of eels in the Atlantic Ocean.

that European eels and North American eels belong to the same species.

Reptiles and amphibians. The range of seasonal movements of most reptiles and amphibians is probably very limited. Generally incapable of travelling any great distance, they respond to unfavourable conditions by lapsing into a state of lethargy. This type of response makes it possible for them to stay in one place for an entire year.

The only migration-like movements of reptiles and amphibians are made during the reproductive period. Frogs and toads then concentrate in particular areas such as ponds and lakes; thousands travel toward these sites from year to year. After reproduction, the animals disperse and again settle over their usual range.

Turtle migration

The South American river turtle, or arrau (*Podocnemis expansa*), migrates along rivers in large masses that may impede the passage of boats. The turtles gather on sandbars of large rivers to lay their eggs. In the Galápagos Islands, giant land tortoises (*Testudo elephantopus*) stay chiefly in the upper humid zone, where food is abundant, but go down to the dry zone to lay their eggs. Despite their great body weight and slow pace, they travel some 30 miles (50 kilometres) across rough country.

Sea turtles, on the other hand, migrate over long distances, lay their eggs on special beaches, and then disperse over a wide area. Green turtles (*Chelonia mydas*), who deposit their eggs on the coast of Costa Rica in Central America, disperse through the Gulf of Mexico and the West Indies. Green turtles that have been tagged on Ascension Island, halfway between Africa and South America, have been recovered on the coast of Brazil, 1,400 miles (2,300 kilometres) away.

Birds. Migration is most evident among birds. Most species, because of their high metabolic rate, require a rich, abundant supply of food at frequent intervals. Such a situation does not always prevail throughout the year in any given region. Birds have thus evolved a highly efficient means for travelling swiftly over long distances with great economy of energy.

The characteristics of migratory birds do not differ greatly from those of nonmigratory forms; many intermediate types exist between the two groups. All transitional forms, in fact, may be manifested in a single species or in a single local population, which is then said to undergo partial migration.

In addition to regular migration, nomadic flights may also occur. This phenomenon takes place, for example, among birds of the arid zones of Australia, where ducks, parrakeets, and seedeaters appear in a locality following infrequent and unpredictable rains, breed, and then move to other areas. Nomadism is a response to irregular ecological conditions.

In Europe. The populations of many northern and eastern European species of birds have pronounced migratory tendencies; the populations of western Europe, on the other hand, are more sedentary.

Some birds are nomadic in winter, others spend the colder months in the southwestern part of the continent or in the Mediterranean region. Many migrant populations migrate to Africa south of the Sahara. Geographical conditions determine several main routes. The Alps are an important barrier to migratory birds. About 150 species travel westward and southwestward; others travel southeastward.

Tits (*Parus*), goldfinches (*Carduelis carduelis*), and blackbirds (*Turdus merula*) are usually sedentary in western Europe; they are usually migratory, however, in northern Europe, where their flights resemble a short migration. Starlings (*Sturnus vulgaris*) are sedentary in western Europe, where large numbers gather from eastern Europe. Large flocks also pass the winter in North Africa.

Insectivorous (insect-eating) species, such as warblers, flycatchers, and wagtails, are highly migratory and spend the winter in the tropics, chiefly in Africa. They migrate to Sierra Leone on the west coast, Tanzania on the east coast, and all the way southward to the tip of the continent. Most of these migrants use different routes to cross the Mediterranean, chiefly in the western portion, although some migrate only southeastward. Golden orioles (*Oriolus oriolus*) and red-backed shrikes (*Lanius collurio*) go to East Africa by way of Greece and Egypt. Swallows—particularly barn swallows (*Hirundo rustica*) and house martins (*Delichon urbica*)—and swifts (*Apus apus*) pass the winter in Africa south of 20° N latitude, particularly in South Africa, in the Congo River region, and in some coastal areas of West Africa.

From J. Dorst, *The Migrations of Birds;* published by Houghton Mifflin Company

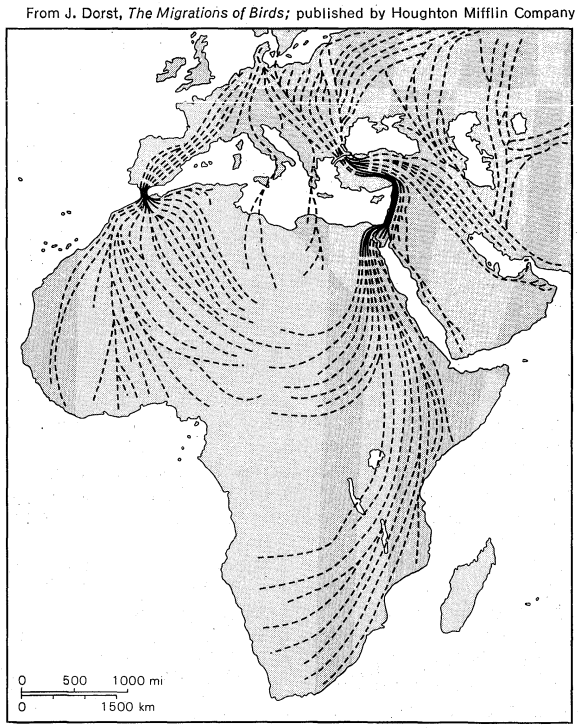


Figure 3: Principal routes taken by the European white stork (*Ciconia ciconia*) between nesting grounds in Europe and wintering grounds in Africa.

Migration routes of storks, ducks, geese, and swans

Among nonpasserines—*i.e.*, nonperching birds—one of the best known migrants is the stork (*Ciconia ciconia*), which migrates to tropical Africa along two well-defined flyways. The stork population nesting west of a line that follows the Weser River in West Germany flies southwestward through France and Spain, past the Strait of Gibraltar and reaches Africa by way of West Africa; the eastern population, by far more numerous, takes a route over the straits of the Bosporus, through Turkey and Israel, to east Africa. These well-separated routes are probably a result of the stork's aversion to long flights over water.

Ducks, geese, and swans also are migrants. These birds winter partly in western Europe and partly in tropical Africa. In Africa they are likely to spend the winter in lake and river regions from Senegal in western Africa to the Sudan in eastern Africa, where thousands of garganeys (*Anas querguedula*) and pintails (*A. acuta*) congregate annually. Some ducks leave their breeding grounds to molt (a process by which old feathers are replaced) in areas where they are most secure from predators during the time they are unable to fly; this is known as a molt migration. After molting, the ducks fly to their final winter quarters. Wading birds (Limicolae) also are typical migrants, most of them nesting in tundra of the Arctic region and wintering along the seacoasts from western Europe to South Africa.

In North America. North American birds must endure the same hazards of winter as European species. The geographical arrangement of the continent determines the main routes of migration, which run from north to south and include the Atlantic oceanic route, the Atlantic Coast route, the Mississippi flyway, the central flyway, the Pacific flyway, and the Pacific oceanic route. A great many birds pass the winter in the Gulf States, where the climate is favourable and food is abundant, but the principal wintering area extends through Mexico and Central America to Panama, which has the greatest density of winter bird residents in the world.

The ruby-throated hummingbird (*Archilochus colubris*) nests in southern Canada and winters in Central America as far south as Panama. Some of these birds fly nonstop across the Gulf of Mexico. Because of their food requirements, many American flycatchers (Tyrannidae), which are mainly insectivorous, have the same migratory behaviour as the hummingbirds. Others, like the phoebe (*Sayornis phoebe*), spend the winter in the Gulf States. Birds such as the American robin (*Turdus migratorius*) and several species of grackles assemble in the Gulf States in enormous flocks. The seasonal flights of the American wood warblers (Parulidae) are among the most spectacular on the North American continent. Some spend the winter in the Gulf States and in the West Indies; others, such as the blackpoll warbler (*Dendroica striata*), travel to Guiana, Brazil, and Peru by way of the West Indies. The spring migration routes of the Canada goose span the Continent of North America in an east–west direction from Hudson Bay as far south as Chesapeake Bay.

South America is winter quarters for several tanagers, such as the scarlet tanager (*Piranga olivacea*) and the bobolink (*Dolichonyx oryzivorus*); these birds migrate through the eastern United States and past Cuba to the swampy regions of Bolivia, southern Brazil, and northern Argentina. This area of South America is also winter quarters for the American golden plover (*Pluvialis dominica dominica*), which travels in an enormous loop over much of the New World. After nesting in the tundras of Alaska and Canada, the plover assemble in Labrador in easternmost Canada, then fly to Brazil over an oceanic route (the shortest possible route) about 2,400 miles (3,900 kilometres) long. Their return flight traverses South America, Central America, and the Gulf of Mexico, then follows the Mississippi Valley.

The role of wet and dry seasons

In intertropical regions. Birds of tropical regions migrate according to the rhythmic succession of wet and dry seasons—a profoundly influential factor on the annual cycle of animals and plants alike.

The migratory behaviour of birds has a unique regularity in Africa, where life zones are arranged symmetrically by latitudes away from the equator. Some migrants never cross the equator. The standard-wing nightjar (*Macrodipteryx longipennis*), which nests in a belt extending from Senegal in the west to Kenya in the east along the equatorial forest, migrates northward to avoid the wet season. The plain nightjar (*Caprimulgus inornatus*), on the other hand, nests in a dry belt from Mali in the west to the Red Sea and Kenya in the east during the rains and then migrates southward to Cameroon and the northern Congo region during the dry season.

Other birds migrate across the Equator to their alternate seasonal grounds. Abdim's stork (*Sphenorhynchus abdimii*) nests in a belt extending from Senegal to the Red Sea; after the wet season, it winters from Tanzania through most of southern Africa. The pennant-wing nightjar (*Cosmetornis vexillarius*), in contrast, nests in the Southern Hemisphere south of the Congo forests during the austral, or Southern Hemisphere, summer, then starts north with the onset of the rainy season. It spends its winters in savannas from Nigeria to Uganda.

In coastal and pelagic regions. Among the migrating seabirds, a distinction must be made between the coastal and the pelagic, or open-sea, species. Birds such as guillemots, auks, cormorants, gannets, and gulls—all common to the seashore—stay in the zone of the continental shelf. Except during the breeding season, they are dispersed over a vast area, often preferring specific directions of travel. Gannets (*Sula bassana*) nesting around the British Isles spread in winter along the Atlantic coast of Europe and Africa to Senegal, the young travelling farther than the adults. Pelagic birds, most of which belong to the order Procellariiformes (petrels and albatrosses), cover much greater distances and, from a few small nesting areas, roam over a large part of the oceans.

Wilson's petrels (*Oceanites oceanicus*), which nest in the western sector of the Antarctic (South Georgia Island, Shetland Islands, and South Orkney Islands), spread rapidly northward in April along the coasts of North and South America and stay in the North Atlantic during the summer. In September they leave the western Atlantic, travelling east, then southeast, along the coasts of Europe and Africa toward South America and their Antarctic breeding grounds, arriving there in November. These petrels thus travel in a great loop through the whole Atlantic Ocean, in a flight pattern correlated with the direction of prevailing winds. The same pattern is used by other seabirds normally carried by the winds. Albatrosses, such as the wandering albatross (*Diomedea exulans*) that nests on small Antarctic islands, circle the globe during their migrations. One such bird, banded as a chick at Kerguelen Island in the southern Indian Ocean and recovered at Patache, Chile, travelled in less than ten months at least 8,100 miles (13,000 kilometres)—perhaps as much as 11,200 miles (18,000 kilometres)—by drifting with the prevailing winds.

The significance of prevailing winds

In the Pacific, short-tailed shearwaters (*Puffinus tenuirostris*) nest in enormous colonies along the coasts of southern Australia and in Tasmania, then migrate across the western Pacific to Japan, remaining in the North Pacific and the Arctic Ocean from June to August. On the return migration they go east and southeast along the Pacific coast of North America, then fly diagonally across the Pacific to Australia.

Arctic terns (*Sterna paradisaea*), whose breeding range includes the northernmost coast of Europe, Asia, and North America, spend the winter in the extreme southern Pacific and Atlantic, chiefly along Antarctic pack ice 11,000 miles from their breeding range. American populations of the Arctic tern first cross the Atlantic from west to east, then follow the coast of western Europe. Arctic terns thus travel further than any other bird species.

Modes of migration. The migration flights of birds follow specific routes, sometimes quite well defined over long distances. The majority of bird migrants, however, travel along broad airways. A single population of migrants may be scattered over a vast territory so as to form a broad front hundreds of miles in width. Such routes are determined not only by geographical factors—*e.g.*, river systems, valleys, coasts—and ecological conditions but are also dependent upon meteorological conditions; *i.e.*, birds change their direction of flight in accordance with the direction and force of the wind. Some routes cross oceans. Small passerine (perching) birds migrate across 600 or more miles of sea in areas such as the Gulf of Mexico, the Mediterranean Sea, and the North Sea. American golden plover, wintering in the Pacific, fly directly from the Aleutian Islands (southwest of Alaska) to Hawaii, the 2,050-mile (3,300-kilometre) flight requiring 35 hours and more than 250,000 wing beats.

The speed of migratory flights depends largely on the species and the type of terrain covered. Birds in migration go faster than otherwise. Rooks (*Corvus frugilegus*) have been observed migrating at speeds of 32 to 45 miles per hour; starlings (*Sturnus vulgaris*) at 43 to 49 miles per hour; skylarks (*Alauda arvensis*) at 22 to 28 miles per hour; and pintails (*Anas acuta*) at 31 to 51 miles per hour. Although the speeds would permit steadily flying migrants to reach their wintering grounds in a relatively short time, the journeys are interrupted by long stops, during which the birds rest and hunt for food. The red-backed shrike (*Lanius collurio*) covers an average of 620 miles (1,000 kilometres) in five days as follows: two nights for migration, three nights for rest, five days for feeding.

Most migrations occur at relatively low altitudes. Small passerine birds often fly at less than 200 feet. Some birds, however, fly much higher. Migrating passerines, for example, have been observed at altitudes as great as 14,000 feet (4,000 metres). The highest altitude recorded thus far for migrating birds is 29,500 feet (9,000 metres) for geese near Dehra Dūn in northwest India.

Pelicans, storks, birds of prey, swifts, swallows, and finches are diurnal (daytime) migrants. Waterbirds, cuckoos, flycatchers, thrushes, warblers, orioles, and buntings are mostly nocturnal (nighttime) migrants. Studies of nocturnal migrants using radar on telescopes focussed on the moon show that most migratory flights occur between 10 PM and 1 AM, diminishing rapidly to a minimum at 4 AM.

Most birds are gregarious during migration, even those that display a fierce individualism at all other times, such as many birds of prey and insectivorous passerines. Birds with similar habits sometimes travel together, a phenomenon observed among various species of shorebirds. Flocks sometimes show a remarkable cohesion; the most characteristic migratory formation of geese, ducks, pelicans, and cranes is a "V" with the point turned in the direction of flight.

Gregariousness during migration

Mammals. Seasonal movements are not widespread among terrestrial species of mammals, because walking speed is relatively slow and energy consumption great. Marine and flying mammals have a much greater tendency to migrate, a tendency that is directly related to their locomotive powers.

Terrestrial mammals. True migration among mammals occurs mostly among large artiodactyls (split-hoofed animals) living in habitats with wide fluctuations of climatic and biotic conditions.

In North American Arctic regions, herds of caribou (*Rangifer tarandus*) settle during the summer in the barrens—rather flat wasteland with little vegetation. In July, the animals begin to move irregularly southward and spend the winter in the taiga, or northern forests, through which they wander freely with no general directional trend. Each herd seems to move in accordance with local conditions and without a well-defined pattern. The caribou again move northward as early as late February and return to the barrens. These migrations follow the same routes from year to year.

In former times, American bison (*Bison bison*) migrated regularly through the Great Plains. Herds of as many as 4,000,000 animals moved from north to south in fall and returned when spring rains brought fresh grass to the northern part of their range. Bison travelled over more or less circular routes and spent the winter in areas 200 to 400 miles from the summer range. Other North American mammals, such as elk (*Cervus canadensis*), mule deer (*Odocoileus hemionus*), and dall sheep (*Ovis dalli*) still migrate regularly in areas undisturbed by man.

Large African mammals migrate in accordance with the succession of wet and dry seasons, which can greatly modify the habitat. Some antelope remain in small areas throughout the year, but many species undertake seasonal movements over a large range. In the Serengeti region of Tanzania, plains animals, particularly wildebeests (*Connochaetes taurinus*) and zebras, travel more than 1,000 miles in their seasonal migrations. Herds spread outward during the rains and concentrate during the dry season

around water holes. Elephants (*Loxodonta africana*) wander great distances in search of the best food and water supply.

In southern Africa, hundreds of thousands of springbok (*Antidorcas marsupialis*) once migrated according to the rhythm of rainfall over their vast range. They moved in herds so dense that any animal encountered was either trampled or forced along with the herd. These huge migrations often resulted in enormous losses from starvation, drowning, or disease—natural methods for controlling overpopulation. Such movements, involving lesser numbers, still occur in parts of South West Africa and in Botswana.

Flying mammals (bats). A few bats native to Europe and Asia make short flights to winter quarters. In the Soviet Union, the common pipistrelle (*Pipistrellus pipistrellus*) and the frosted bat (*Vespertilio murinus*) withdraw to hibernating places at some distance from their summer range. In Germany, the large mouse-eared bat (*Myotis myotis*) leaves its winter quarters in Brandenburg in March or April and travels as much as 160 miles (260 kilometres) to its summer habitat in northwestern and northeastern Germany. It regularly returns to the same winter locality. Schreiber's long-fingered bat (*Miniopterus schreibersii*) changes its habitat in winter and moves over 100 miles or more according to a complex pattern. These local movements represent an adjustment to winter conditions and the search for more habitable caves.

Migratory bats in the United States

Other bats travel even greater distances. In the United States, the red bat (*Lasiurus borealis*), the large hoary bat (*L. cinereus*), and the silver-haired bat (*Lasionycteris noctivagans*)—three species that roost primarily in trees and shrubs—are true migrants with strong powers of flight. They summer in the northern U.S. and in Canada and winter in Georgia, South Carolina, Florida, and probably also in the southwestern states. The southward movement is made from mid-August to November. Migration flights occur at night, and during the day under favourable conditions. Large numbers follow the coast some distance from land, and all three species are found at sea far from the coast and in Bermuda. Fruit bats and flying foxes (*Pteropus*) native to the tropical regions of the Old World make regular mass migrations, following the seasons for fruit ripening.

Marine mammals. Antarctic whale migrate regularly to the tropics, a fact long known to whalers. By systematically marking whale by shooting into them steel tubes engraved with a serial number, man has obtained evidence of actual movements. A young fin whale (*Balaenoptera physalus*) marked in February in the Antarctic, at latitude 65° S, was captured two years later, in July, off the coast of South Africa, 1,900 miles north. During the austral (Southern Hemisphere) winter, whale migrate to areas rich in food, particularly the northwestern coast of Africa, the Gulf of Aden, and the Bay of Bengal. Antarctic whale—particularly humpbacks (*Megaptera novaeangliae*), a highly migratory species—can be divided into five distinct populations around Antarctica; each population migrates separately, and individuals usually return to their respective zones, though interchange may occur. The Antarctic population does not, however, migrate entirely into warm waters during the winter, and a segment of the population seems to stay behind at about latitude 50° S.

Northern whale have the same migratory habits as Antarctic whale. Northern blue whale (*Balaenoptera musculus*) migrate northward along the east coast of the United States, then through Davis Strait to Baffin Bay (north of Canada) or Spitsbergen to the waters off northern Scotland or the coast of Norway. They are believed to migrate southward along the same routes. Part of the North Pacific stocks of the northern blue whale winters in the Indian Ocean and in the seas bordering Indonesia.

Smaller cetaceans (porpoises and dolphins) migrate in the same way, as indicated by population fluctuations within a particular area; but little is known about their distribution and migration.

Noteworthy migratory habits occur among the pinnipeds (seals and walrus), some of which disperse over wide areas at times other than the breeding season. The harp seal (*Pagophilus groenlandicus*) lives in summer in northernmost Arctic waters, but reproduces in the White Sea (an arm of the Arctic Ocean extending southward into the landmass of the Soviet Union), in the eastern North Atlantic, and around Newfoundland, where young are born between January and April. The seal then returns to more northern latitudes. Northern fur seals (*Callorhinus ursinus*) reproduce only on the Pribilof Islands, off southwestern Alaska, from May to November, and the colonies then disperse into the open seas. The males stay in the Gulf of Alaska and off the Aleutian Islands; the females go farther south, to southern California, some 3,000 miles away.

NAVIGATION AND ORIENTATION

Migrants often return to breed in the exact locality where they were hatched or born. This journey homeward, particularly that of birds, may cover thousands of miles.

Homing experiments

Homing experiments have demonstrated the ability of animals to orient themselves geographically. Such experiments involve removing animals from a specific point (usually the nest), transporting them for various distances, and analyzing their speed and degree of success in returning. Starlings have returned to their nests after being transported 500 miles (800 kilometres); swallows have returned a distance of over 1,100 miles (1,800 kilometres). A Manx shearwater (*Puffinus puffinus*) returned from Massachusetts to Britain, 3,050 miles (4,900 kilometres) across the Atlantic, in 12½ days. Laysan albatrosses (*Diomedea immutabilis*) returned to Midway Island in the Pacific after being released at Widby Island, Washington; the journey covered 3,200 miles and took 10.1 days. Experiments with certain fishes and mammals have demonstrated similar homing ability.

It is apparent that homing animals use familiar landmarks; both random and oriented searches have been observed in birds and fish. Homing experiments with gannets observed from aircraft have demonstrated that, after release, the birds explore the region and hesitate as they apparently look for landmarks. Landmarks vary from topographical (*e.g.*, mountain systems, river systems, coastlines) to ecological (*e.g.*, vegetation zones) to climatic (*e.g.*, air masses differing in temperature and humidity, prevailing winds). Fish may orient themselves by using similar clues in the same way. Passive drifting is an important factor in the movements of larvae and young fishes, such as those of the eel, cod, herring, and plaice, and even in adults that are passive after spawning, such as herring and cod. As a result of drifting with the current, the movements of such fishes are similar from year to year.

Familiar landmarks and exploration do not, however, explain how migrants find their way along routes covering many hundreds or thousands of miles nor do the results of most homing experiments.

Birds. A compass sense has been demonstrated in birds; that is, they are able to fly in a particular constant direction, regardless of the position of the release point with respect to the bird's home area. It has also been shown that birds are capable of relating the release point to their home area and of determining which direction to take, then maintaining that direction in flight. The navigational ability of birds has long been understood in terms of a presumed sensitivity to both the intensity and the direction of the earth's magnetic field. It has also been suggested that birds are sensitive to forces produced by the rotation of the earth (Coriolis force); however, no sense organ or physiological process sensitive to such forces has yet been demonstrated to support this hypothesis.

Experiments have shown that the orientation of birds is based on celestial bearings. The sun is the point of orientation during the day, and birds are able to compensate for the movement of the sun throughout the day. A so-called internal clock mechanism in birds involves the ability to gauge the angle of the sun above the horizon. Similar mechanisms are known in many animals and are

closely related to the rhythm of daylight, or photoperiodism. When the internal rhythm of birds is disturbed by subjecting them first to several days of irregular light–dark sequences, then to an artificial rhythm that is delayed or advanced in relation to the normal rhythm, corresponding anomalies occur in the homing behaviour.

From M. Ricard, *The Mystery of Animal Migration*

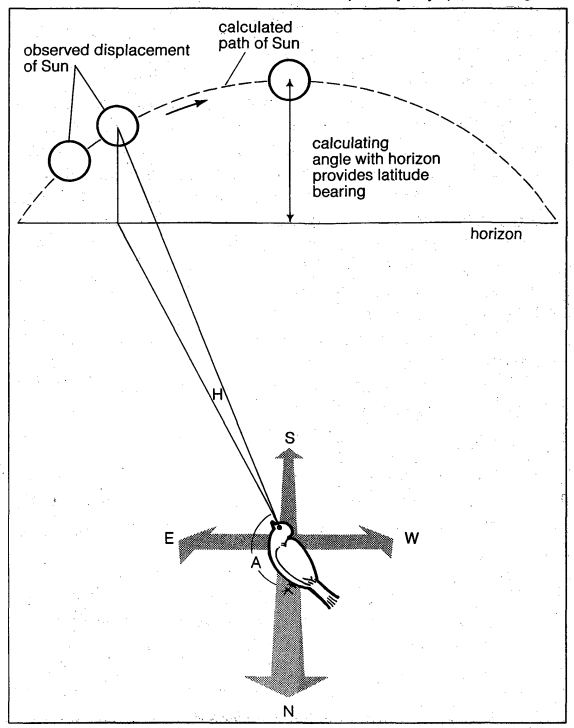


Figure 4: Matthews' hypothesis of solar navigation by birds. Birds can visualize the complete trajectory of the sun and determine its zenith and altitude at midday (angle H) in order to derive the latitude. Angle A is measured from the north to the point at which a vertical from the sun cuts the horizon. From this longitude can be derived.

Two theories have been formulated to explain how birds use the sun for orientation. Neither theory, however, has thus far been substantiated with proof. One theory holds that birds find the right direction by determining the horizontal angle measured on the horizon from the sun's projection. They correct for the sun's movement by compensating for the changing angle and thus are able to maintain the same direction. According to this theory, therefore, the sun is a compass that enables the birds to find and maintain their direction. This theory does not explain, however, the manner in which a bird, transported and released in an experimental situation, determines the relationship between the point at which it is released and its goal.

The second theory, proposed by British ornithologist G.V.T. Matthews, is based on other aspects of the sun's position, the most important of which is the arc of the sun —*i.e.*, the angle made by the plane through which the sun is moving in relation to the horizontal. Each day in the Northern Hemisphere, the highest point reached by the sun lies in the south, thus indicating direction; the highest point is reached at noon, thus indicating time. In its native area a bird is familiar with the characteristics of the sun's movement. Placed in different surroundings, the bird can project the curve of the sun's movement after watching only a small segment of its course. By measuring maximum altitude (the sun's angle in relation to the horizontal) and comparing it with circumstances in the usual habitat, the bird obtains a sense of latitude. Details of longitude are provided by the sun's position in relation to both the highest point and position it will reach—as revealed by a precise internal clock.

Use of celestial navigation by birds

Migrant birds that travel at night are also capable of directional orientation. Studies have shown that these birds use the stars to determine their bearings. In clear weather, captive migrants head immediately in the right direction using only the stars. They are even able to orient themselves correctly to the arrangement of night skies projected on the dome of a planetarium; true celestial navigation is involved because the birds determine their latitude and longitude by the position of the stars. In a planetarium in Germany, blackcaps (*Sylvia atricapilla*) and garden warblers (*S. borin*), under an artificial autumn sky, headed "southwest," their normal direction; lesser whitethroats (*S. curruca*) headed "southeast," their normal direction of migration in that season.

It is known, then, that birds are able to navigate by two types of orientation. One, simple and directional, is compass orientation; the second, complex and directed to a point, is true navigation, or goal orientation. Both types apparently are based on celestial bearings, either the position of the sun or the patterns of the stars, which provide a navigational "grid."

Other animals. The methods of directional orientation used by birds are similar to those used by other animals. Orientation to the sun has been demonstrated in various crustaceans, particularly in the sand flea (*Talitrus saltator*). Various insects, particularly bees and certain beetles (families Scarabeidae, Tenebrionidae, and Carabidae) use the sun to plot their course with remarkable accuracy.

Fishes also are able to use celestial bearings; salmon presumably use the sun. Experiments with the parrot fish (*Scarus*) have demonstrated a sun compass reaction that may also occur in other fishes. Localization of the sun is, however, much more difficult in water than in the air, because of the characteristics of light rays passing through water. Experiments suggest that topographical clues are also used by fishes to recognize their range, particularly their spawning grounds. Visual bearings in this respect have great importance. It is possible that chemical substances also provide clues; the role of olfaction in the salmon has been described (see above, *Anadromous fishes*).

Visible landmarks are used by mammals, at least for orientation within short distances. Scented trails are probably helpful within a limited area, proportionate to the size of the animal; olfaction plays an important role in the life of mammals. Some mammals, however, migrate over enormous distances and are able to return after being taken far away from their home territory; bats, for example, have returned 165 miles (265 kilometres) to their caves. Random exploration plays a part in such movements, but it is possible that some type of true navigation is involved in certain of these movements.

PHYSIOLOGICAL STIMULUS OF MIGRATION

Migration, like reproduction and other phases (as molting in birds), is part of the life cycle and depends on a complex internal rhythm that affects the whole organism, particularly the endocrine glands (glands of internal secretion) and the gonads. Migration must thus be viewed in relation to the entire annual cycle.

Each year birds return to particular areas to breed, and remain there until the members of the brood can care for themselves. There is no relation between the reproductive and migratory stimuli, yet the two phenomena, although independent, are nevertheless stimulated by the same factor.

A physiological study of certain migrants has revealed that metabolic patterns usually change prior to migration, and fats accumulate in the body tissues. The whitethroat (*Sylvia communis*) weighs an average of 12 to 13 grams (about 0.4 ounce) during the breeding season, 16 to 19 grams (about 0.6 ounce) in the autumn, and 20 to 24 grams (about 0.8 ounce) in the winter. Food consumption increases with the autumn molt, reaching a peak at the beginning of the migration season. These fundamental physiological changes, chiefly under the control of the thyroid gland, are correlated with migratory activity. Such fluctuations are not observed in nonmigratory species.

Variations in metabolism and related phenomena are controlled by an endocrine gland, namely the pituitary gland, which is located in the lower part of the brain and acts as a command post, sending out instructions in the

The role of the pituitary gland in migration

form of secretions called hormones. That the pituitary has a cycle independent of environmental factors is demonstrated by the regularity with which phases such as reproduction occur from year to year in the lives of some birds, and by the diverse response of various species and populations to the same environmental factors. That the pituitary is, however, influenced by environmental factors, such as variations in day length and the intensity of the sun, has been demonstrated experimentally.

Gonadal development and the deposition of fat, for example, are influenced by the pituitary, which responds to increasing day length in springtime by accelerating the rate of gonadal development. The pituitary thus governs the development of gonads and, in addition, affects all metabolic processes, including development of the thyroid gland, so as to prepare the animal physiologically for migration. If only the pituitary and variations in day length were involved, migration would be triggered at definite times, because the pituitary cycle is fixed, and photoperiodism is a highly predictable phenomenon; such a lack of flexibility, however, would inevitably cause migrant populations to suffer catastrophes because ecological conditions are irregular—meteorological events, such as the arrival of spring, and biological phenomena, such as flowering, foliation, hatching of insects, and availability of food, are highly variable from year to year. The pituitary thus serves only to prepare the bird for flight; the proper ecological conditions, on the other hand, are necessary to initiate it. The availability of food is an important factor. Temperature and weather conditions also have an influence—a sudden period of cold weather during autumn may induce the immediate departure of many migrants.

Sensitivity to changes in the weather and other environmental conditions varies markedly among species. Some, such as the woodcock, snipe, lapwing, starling, and lark, rely on surrounding conditions to initiate their spring and autumn migrations, and the patterns of their flight depend on temperature and barometric pressure. Others, such as the swift, cliff swallow, Baltimore oriole, and short-tailed petrel, are less weather dependent, and, since the dates of their arrival and departure are not regulated by the weather, they occur with remarkable regularity each year.

The factors that stimulate migration in animals other than birds are not yet well understood. Ecological conditions play a great part in the migratory activity of mammals, who react to general food shortage by moving to another region. Whale, for example, leave the Antarctic region as winter modifies the oceanographic conditions. Seals disperse when the food supply in the area of their breeding colonies is depleted. Environmental factors are of primary importance in the migration of fishes and marine invertebrates. Annual movements of water masses change physical conditions such as temperature and salinity; biotic conditions are influenced accordingly.

ORIGIN AND EVOLUTION OF MIGRATION

The origins of migration remain in the realm of pure conjecture; neither observation nor experiment has resolved the matter. The explanation, however, must be related to geographical and climatological factors that have prevailed since the Tertiary Period, which ended some 2,500,000 years ago. The great Quaternary ice ages, which came later, were very important in altering the distribution of animals over a large part of the world, but migrations occurred long before.

Migration, as it is now known among modern birds and mammals, probably appeared gradually by stages. Some animals changed their habitat only slightly, never leaving the same general region. The movements of other animals were more erratic, their dispersal being oriented toward the most favourable places. Such movements are the first stages of true migration—a phenomenon characterized by elaborate mechanisms—which gradually acquired stability through natural selection. At first, many populations must have perished rather than attempt to flee from unfavourable conditions. Only a fraction of such populations probably sought more favourable conditions elsewhere, but natural selection favoured the "migrants," and migratory tendencies were retained.

Causes of change of habitat

In some cases, original habitats were in present-day wintering areas, and animals developed a tendency to leave in spring in order to breed in other territories. Seasonal changes of weather and food supply in these newly settled regions forced the animals to migrate in fall, and they thus retreated to their former range. Among birds nesting in the Northern Hemisphere, hummingbirds, tyrant flycatchers, tanagers, orioles, bee-eaters, and swifts have distinct tropical affinities; in recent geological times these birds gradually spread northward as glacial ice receded and the continent became warmer. Other birds, such as plover, ducks, and geese, originally lived in what is now their breeding area. Gradual climatic changes forced them to spend their winters in regions far to the south. Migrations thus appear to be the consequence of invasions or emigrations, during which animals settle in new areas during a segment of the annual cycle.

Migratory birds use the routes by which their ancestors first invaded new regions after the glacial recession. The yellow wagtail (*Motacilla flava*) and the wheatear (*Oenanthe oenanthe*) settled in Alaska; they migrate annually into other parts of the Western Hemisphere but spend their winters in the warm regions of southeastern Asia and even Africa, probably following the migratory route of their ancestors. A typically North American species, the grey-cheeked thrush (*Hylochichla minima*), which has extended its breeding area to northeastern Siberia, returns to spend the winter in the central regions of South America.

ECOLOGICAL SIGNIFICANCE OF MIGRATION

There are many ecological implications of migration. The food resources of some regions would not be adequately exploited without moving populations. The sequence of migratory movement is closely integrated in the annual cycle of ecosystems characterized by productivity fluctuations. Migratory behaviour concerns only species located at specific trophic levels (zones of food availability) where maximal fluctuations occur both in breeding areas and in wintering regions. Migrant birds avoid equatorial forests where productivity is constant throughout the year, and food surpluses do not occur. They do congregate, on the other hand, in savannas where productivity varies with the seasons.

Such a coordinated sequence is particularly apparent in the case of birds migrating from the northern Arctic regions to tropical winter regions; both life zones are characterized by broad fluctuations in productivity. In the Arctic, vegetal and animal production is very high during the summer; ducks and waders nest in great numbers, exploiting these resources. As winter comes, food becomes scarce, and water birds migrate to the tropics, where the rainy season has caused food production to increase to optimal levels. Ducks and wading birds concentrate in the most favourable areas, remaining until spring, when productivity is lowest. By then the condition of breeding areas is again favourable for the birds. The life cycles of these birds are closely attuned with the cycles of their various habitats, and the sizes of bird populations are controlled by the capacity of both areas to sustain them.

Migration, then, has considerable ecological significance. It enables fast-moving animals to exploit fluctuating resources and to settle in areas where life would not be tenable for animals incapable of rapid travel. On the other hand, peaks of food production would be unexploited without the periodic presence of migratory populations.

BIBLIOGRAPHY. G.M. ALLEN, *Bats* (1939), a classic book on bat biology with a chapter on migration; F. BOURLIERE, *Vie et moeurs des mammifères* (1951; Eng. trans., *The Natural History of Mammals*, 3rd ed. rev., 1964), a modern review of mammal biology with a chapter on migration; J. DORST, *Les Migrations des oiseaux* (1956; Eng. trans., *The Migrations of Birds*, 1962), a modern review of all aspects of bird migration; D.R. GRIFFIN, *Bird Migration* (1964), a useful summary with a good chapter on orientation; F.R. HARDEN JONES, *Fish Migration* (1968), a standard book, with

many classic references; G.V.T. MATTHEWS, *Bird Navigation* (1955), an original attempt to explain bird orientation; R.T. ORR, *Animals in Migration* (1970), a modern review of migration patterns among animals; E.J. SLIJPER, *Walvissen* (1958; Eng. trans., *Whales*, 1962), a modern review of whale biology, including a chapter on migration; A.L. THOMSON, *Bird Migration*, rev. ed. (1942); and C.B. WILLIAMS, *The Migration of Butterflies* (1930), two classic books, still useful; A. WOLFSON, "Ecologic and Physiologic Factors in the Regulation of Spring Migration and Reproductive Cycles in Birds," in A. GORBMAN, *Comparative Endocrinology* (1959), an accurate review of the physiology of bird migration, somewhat out of date but still useful.

(J.P.D.)

Migration, Human

Migration is commonly defined as the permanent change of residence by an individual or a group. Since the definition does pose the problem of deciding what is "permanent," statisticians collecting migration data must usually rely on more or less arbitrary criteria. In statistics of international migration, for instance, a person is usually classified as an immigrant if he has stated his intention of settling in the country of entry for at least one year.

Various authors would like to include further limiting clauses in the definition. In particular, some would prefer to exclude all forced moves from the concept of migration, even though it is not always easy to say whether a move of residence is or is not forced. Similarly, some sociologists would like to limit the migration concept to moves from one social system to another, but the central problem here is that modern man participates simultaneously in various social groups. Some of these groups have their members concentrated in a small area, whereas others (*e.g.*, the Roman Catholic Church) have their members spread over the whole world. As a consequence, anyone who moves his residence leaves behind some of the social groups to which he belongs, while other groups to which he is affiliated are to be found also in his new place of residence. This means that if migration is to be defined as moving from one social system to another, the difficult decision must be reached as to how far group affiliations must change before one can say that there has been a move between different social systems.

Distinctions between migration and other moves by man

The criterion that migration be a "permanent" move differentiates migration from various other types of human moves. One of these is nomadism, the way of life of those having no fixed abode. The nomadic way of life may be brought about by economic necessity, as in the case of certain hunters and gatherers or of cattle-breeders who follow their livestock about. For other peoples, such as gypsies, nomadism is a tradition. A modern phenomenon akin to nomadism is migrant labour (*q.v.*). "Seasonal moves" differ from nomadism in that moves are made only twice a year, between winter homes and summer abodes. The Lapps offer an example of a people who have retained this mode of life up to the present.

Another type of human move is commuting—referring to the daily, or sometimes weekly, journeying to work of people who live a considerable distance from their jobs. Commuting has increased enormously in the developed countries over the past half century or so. Temporary moves are also made in tourism, which nowadays involves vast numbers of people. Quite often when moves to temporary homes are involved, tourism resembles migration, but its temporary character places it in a different category.

THE DYNAMICS OF MIGRATION

Motives for migration. The argument that man migrates because he is inherently restless—nomadic by nature—is countered by the argument that man is inherently sedentary, remaining fixed until some force impels him to move. The former proposition provokes the question as to why many people do *not* migrate; whereas the latter provokes the reply that many migrants were obviously by no means forced to move. Some further explanation for the migration phenomenon is needed, and this has been sought in the concepts of "push" and "pull." Migrants may be pushed by circumstances in their region of origin—for instance, exhaustion of natural resources. Conversely, they may be pulled by attractions in their region of destination—for example, better employment opportunities.

The "push-pull" factors in migratory patterns

Some investigators have tried to explain all migrations by push factors only, whereas others have stressed pull factors exclusively. It is undoubtedly true that push factors sometimes suffice to bring about migration (such as in the case of famine). In other instances, pull factors may be entirely responsible for a migratory movement (such as in gold rush). In most cases, however, both push and pull factors play a role. A prospective migrant weighs the circumstances in his region of origin and the attractions as well as the disadvantages of every potential region of destination. Such interregional differences thus usually form a prerequisite for migration (see below).

In highly developed countries, exceptions to the rule that interregional differences form a prerequisite for migration are rare. Only where economic development has not progressed very far do push factors alone sometimes suffice to bring about migration. Thus a dearth of basic means of livelihood may mean that moving elsewhere—anywhere—is seemingly the only solution. Then, migration may also be affected disproportionally by defective communication. Occasionally, rumours grossly exaggerate the opportunities in potential regions of destination with the result that a migration "craze" sets in.

There are indications that in many of the present developing countries, notably in Asia and the Middle East and more recently in Africa, this kind of push migration obtained after World War II. In these countries, according to some authors at least, agricultural crises produced rural–urban migration streams far in excess of what would be warranted by economic opportunities in the cities. The authors who adhere to this view speak of overurbanization.

Distance and direction of migration. Communication, as already implied, is important for migration, because obviously if interregional differences are to bring about any sizable migration, conditions in the region of destination must be known in the region of origin. The communication factor is principally responsible for the inverse relation between volume of migration and distance. This inverse relation was discovered by the British statistician Edward G. Ravenstein, whose first "law" in his classic "The Laws of Migration," published in 1885, stated that the great body of migrants proceed only a short distance. This has been proved true in numerous migration studies in several countries.

Not only geographical but also technical distance (by which is meant distance as influenced by the available means of transportation and communication) and social distance are important for migration. Aspects of social distance such as religions and language differences may seriously impede migratory movement.

Contrasts between international and internal migration

The difference in physical as well as social distance is the main distinction between international migration and internal migration (that is, migration within a country). International migrants not only often cover great physical distances but also face an unaccustomed cultural environment and a strange language. It is thus not to be wondered that the volume of internal migration usually far exceeds that of external migration. Within the context of internal migration, local moves can be distinguished from migrations over longer distances. The latter in the majority of cases are economically induced, whereas local moves are usually motivated by housing factors.

Ravenstein's second "law of migration" was concerned with the routes followed by internal migrants. In his momentous work, *The Growth of Cities in the Nineteenth Century*, published in 1899, the American statistician Adna F. Weber, paraphrased this "law" as follows:

> Since, then, the current of migration is toward the cities and yet the bulk of migration is for short distances only, we can see the manner of the movement; it is a migration by stages having for its object the satisfaction of the demands for more labor in the cities.

Migration by stages, then, is a process whereby a general migration over long geographical distances may take place

even though each individual migrant may cover only a short distance at each move. Since migrants tend to cover not only short geographical distances but also short social distances, there is also what may be called migration by steps—that is, migration from rural areas to the big cities via small- and medium-sized towns. The processes of migration by stages and migration by steps were first recognized around 1900, and recent studies show that they still play an important role in internal migration. A new "step" has, however, been added to the migration by steps—the migration from the cities to the suburbs.

Demographic factors. In former times, an entire tribe or nation sometimes migrated as a unit. Today, migration is usually an individual or family affair. This raises the question whether the migrants in a population differ, on the average, from the nonmigrants, and, if so, in what way.

The demographer Dorothy S. Thomas, reviewing studies concerning migration differentials published up to 1938, concluded that the only generalization about this subject that can be considered definitely established concerns age. There is a great preponderance of adolescents and young adults among migrants, particularly migrants from rural areas to towns, as compared with the age ratios in the nonmigrating population. No doubt the likelihood of improving living conditions through migration is greater for young people than for their elders, who are already established in their jobs and residential neighbourhoods.

Influence of age and sex factors

Migration is not selective by sex. Whether males or females migrate more to a certain region depends on the relative opportunities for either sex in that region. There is a tendency, however, for single women to cover shorter distances than men when migrating, possibly because in most countries the pursuance of a career, which may necessitate long-distance migration, is still considered less important for women than for men.

Influence of socio-economic factors

With regard to family status and family composition, a general finding is that married couples without children or with only very young children migrate much more than do couples with older children. Especially in the case of migration to the suburbs, there is a preponderance of families with small children.

Many studies in Germany, the United States, and Sweden have indicated that migrants from rural to urban areas tend on the average to have higher intelligence, education, and socio-economic status. This finding has puzzled and sometimes alarmed social scientists since before the turn of the century. In the 1950s, however, it was recognized that this positive selection is a consequence of the fact that only in the towns is work available for the highly skilled.

A recent phenomenon is the selective migration of highly trained persons—engineers, physicians, scientists—from less developed countries to more developed countries (and sometimes from the less affluent developed nations, such as Great Britain, to the more affluent, such as the United States). This is called "brain drain." Opinions differ as to the extent of the brain drain, and as to its consequences for developing nations.

Migratory selection according to ethnic origin or race is evident in many societies. This results from cultural or social factors. The cultural values of one ethnic group may be more conducive to migration than that of another. More often, however, migration differentials by ethnic group or race are a direct or indirect consequence of discrimination. The wanderings of the Jews form a clear example. In the United States, differences in educational opportunities as well as residential segregation explain differential migration as between whites and nonwhites.

TYPES OF HISTORICAL MIGRATIONS

Over the ages there have been many migratory movements, which various authors have tried to classify under a few broad headings.

Primitive migrations. Probably the oldest type of human migration is that resulting from such push factors as a people's inability to cope with natural forces or their flight from a stronger foe. Though termed "primitive" migration, this type of migration is not confined, however, to "primitive" peoples.

If one takes it that *Homo sapiens* made his appearance on earth in one particular place, primitive migration is the process by which he eventually populated the whole globe. This process of "dispersion" took place almost entirely in prehistoric times, so that not much is known about the routes followed. It is likely that the American Indians reached their country from Asia by way of the Bering Strait, while Oceania also appears to have been peopled from the West. The Americas and Oceania seem to have been the last parts of the world to be settled.

Tribal migrations

In the primitive migrations of former times, usually peoples as a whole and not merely families or small groups migrated. In German this is called *Völkerwanderung*, which may be translated as folk, or tribal, migration.

For over 2,000 years, tribal migrations played an important role in European history—from the incursion of the Dorians in Greece in the 11th century BC to the Norman raids and conquests between 800 and AD 1200. The Germanic migrations of the 4th, 5th, and 6th centuries AD in particular gave rise to much change. The Germanic peoples had originally lived on the shores of the Baltic Sea, but for many centuries they had been moving southward in search of new agricultural land. In the second half of the 4th century, the easternmost of their tribes, attacked by Huns coming up from the Eurasian steppes, invaded the Roman Empire to find a new place to settle. Other Germanic tribes followed suit. Some of them moved hundreds of miles in the course of the next two centuries. Probably none of these tribes was larger than a few score thousand. Numerically these migrations were thus not very important, but their consequences for history were enormous.

Tribal migrations over long distances have also taken place in other continents. One further example is the dispersion of the numerous Bantu tribes in Africa, who between 700 and AD 1200 moved from the neighbourhood of Lake Victoria westward into the Congo area and between 1200 and 1800 moved southward, reaching as far as present day Natal and South West Africa.

Primitive-type migrations, however, also include certain more recent movements of individuals and families. The flight from hunger of large numbers of Irish and German farmers in 1846 and the years following is a case in point. In Ireland and in the lower Rhine area of Germany, millions of people had become dependent on the potato for their livelihood. When in 1845 and 1846 the potato crop failed twice in succession, emigration was for many the only solution.

The 20th century has unfortunately seen a sharp increase in the volume of a category of migrants part of whom also fall within this section—namely, refugees (*q.v.*) Those refugee movements—which are neither forced nor planned but are due to people either judging the political situation in their country of residence unbearable or fearing that it will become so in the near future—may be classified as primitive migrations. Examples are the flight of more than 700,000 Arabs from Israel in 1947–48 and population movements between India and Pakistan in the same years, which involved 14 million people. Certain other refugee movements, in particular the transfers of people dictated by governments, belong to the type of migration dealt with in the next section.

Deportations and exiles

Forced or induced migrations. "Forced" migrations refer to people moved either by the state or by some other social institution that wields power over them. The most stringent forms of forced migration are deportation and slave trade. Deportation has been employed by most governments, at one time or another, to get rid of criminals, political opponents, or other persons deemed undesirable. One widely known example is the movement of some 150,000 British convicts to Australia between 1788 and 1867. Another instance is the exile of 1,000,000 Russians to Siberia during the 19th century. In the course of World War II, the Nazis deported 7,000,000–8,000,000 people, including at least 5,000,000 Jews who were subsequently murdered. After the war, the Allies decided that all Germans living outside the new German borders should be transferred to within them; this involved 9,000,000–10,000,000 people. Meanwhile, the Soviet government de-

ported to Central Asia 400,000 Volga Germans and 600,000 members of other minorities accused of collaboration with the Nazis.

Slave trade was widely practised for thousands of years in Africa, the Middle East, and Europe, where slavery was for a long time a well established institution. From 1510 until late in the 19th century, African slaves were hauled to the Americas. Estimates as to the number of slaves who left the African coast run as high as 20,000,000; a sizable number, however, did not survive the Atlantic passage. In 1942 Hitler revived a form of slavery in Europe by forcing men from the occupied countries to work in Germany. Their number reached eight million in 1944 (see SLAVERY, SERFDOM, AND FORCED LABOUR).

Apart from deportations and slavery, there are instances in which social institutions induce people to move by less severe methods. The term forced migration is really too strong for these cases; it would be better to speak of induced migrations. A widespread modern form of induced migration is transfer of staff. In the military services as well as in certain other governmental agencies and in large, dispersed private companies, personnel may be moved from one place to another by order.

Migrations of adventurers and minority groups

Free migrations. When there are neither push factors nor political forces inducing people to emigrate, nor as yet a migration tradition based on intensive communication between regions of origin and destination, there may be migratory movements of adventurous or deviant individuals or groups. The 19th-century Atlantic migration, for instance, started with these "free migrations." At first, around 1830, students and other young intellectuals predominated among the migrants from continental Europe to North America. They were motivated by their alienation from European society during a period of political reaction. In the course of a few decades, this was followed by the emigration of small groups, mainly religious minorities such as Pietist communities.

The significance of free migration is not in its volume, which is seldom large, but in the example it sets. It is a pioneer movement.

Chain migration. Through the mechanism of chain migration, the pioneer movement noted above can easily lead to migration on a large scale. Chain migration is based on the effect of letters from early migrants to relatives and friends at home. Visits home also exert a strong influence. These personal contacts form a most potent means of communication. Once the first migrants have induced some relations to follow them, these too in turn write to relations and friends. Thus a migration chain is started that may operate for scores of years and result in the movement of thousands of people.

In the purest examples of chain migration, the later migrants settle in the vicinity of the first migrants. In such countries as the United States and Australia, this has produced numerous colonies of people living close together who all hail back to a small region, sometimes one town or village, in Europe. Several examples of such colonies having developed around a single original settler are known. Chain migration does not always result in such colonies, however; the later migrants may disperse after having reached the country of immigration.

Some colonies that had "proprietors" (capitalists to whom kings had granted lands overseas) were also heavily advertised to the very first of the prospective settlers. William Penn's success in colonizing his Pennsylvania in the 17th century, for example, was largely due to his own skill as a promoter, for he wrote enticing tracts and on preaching journeys in Europe described the opportunities offered by his colony. His successful promotions were imitated by others.

Mass migration. Chain migration played a key role in bringing about a new phenomenon during the 19th century—namely, long-distance mass migration. In many countries of Europe especially, emigration became a style, an established social pattern, an example for collective behaviour.

The Great Atlantic Migration

The largest of the intercontinental mass movements was that from Europe to North America: the Great Atlantic Migration, as it has been called. Between 1820 and 1970 about 41,000,000 people left Europe for North America. Five million went to Canada, 36,000,000 to the United States. The flow to the United States reached its peak in the last decades of the 19th century and the first two decades of the 20th century, when it was further stimulated by agricultural crises in Europe. Between 1880 and 1910, 17,000,000 Europeans entered the United States. After World War I, the stream diminished due to American immigration restrictions and changing conditions in Europe.

Two other migrations accounted for between 15,000,000 and 20,000,000 people each since 1820. One was the movement from Europe to South America. This stream was made up mainly of southern Europeans. The other was the migration from European to Asiatic Russia. The figure quoted includes 1,000,000 exiles during the 19th century and 1,000,000 deportees during World War II, but for the remainder this migration was a voluntary movement. In the three decades before World War I alone, more than 5,000,000 Russians settled in Siberia.

Other important intercontinental migrations include European migrations to Africa, Australia, and New Zealand, and Japanese, Chinese, and Indian migrations to the Western Hemisphere. It should be pointed out that most of these migratory movements are still going on. Some, in particular migration from Europe to Australia, have reached a high level only since World War II.

An important mass movement that is *intra*continental is the Chinese emigration to Southeast Asia, which started many centuries ago but rose to large proportions in the 19th century. The present number of Chinese, both original migrants and descendents, in Southeast Asia is usually given as 12,000,000. Malaysia, Thailand, Indonesia, Singapore, and Vietnam have the largest numbers of Chinese immigrants.

In the 20th century, a new type of international mass migration has sprung up, in which the moves are mostly within one continent. This is the migration of unskilled labourers from less developed to more developed countries. Examples are the streams from southern Europe and North Africa to northwest Europe; from Malawi, Zambia, Botswana, and Lesotho to Rhodesia and the Republic of South Africa; from Upper Volta to Ghana and the Ivory Coast; and from Mexico to the United States. Because most of these migrants return after a few years, this is a mixed form of permanent migration and temporary labour transfer.

Modern migrations. The development of extensive mass-communication systems in modern societies has made migration less dependent on the personal messages and individual promotions that had played such a large role in earlier chain migration. Modern migrants tend to get their information concerning the region of destination from newspapers, television, official agencies such as labour exchanges or immigration offices, and so forth.

Rural–urban migration

The principal current of modern migration all over the world is toward urban areas. Of course, cityward migration is not a new phenomenon. Various indications show that the earliest historical cities, those of Mesopotamia in the 4th millennium BC, experienced regular in-migration. Rome attracted migrants by the thousand as early as the 3rd century BC. In the later European Middle Ages, the development of trade and crafts, which tended to be concentrated in the towns, added impetus to cityward migration.

Only with the Industrial Revolution, however, did cityward movements begin to predominate in the pattern of migration. Particularly in Europe, North America, and Australia, and New Zealand did large-scale migrations toward the towns take place in the second half of the 19th century. London, for instance, gained 200,000 people by migration in the 1850 decade, Paris as much as 280,000 in the five years 1876–80. Similarly heavy rural–urban migration began in Latin America and the Soviet Union after World War I, and in most of the remainder of the world after World War II. A single figure may serve to illustrate the magnitude of this movement: in India net migration to the towns amounted to 5,500,000 people in 1951–60.

In the rich countries, however, urbanization nowadays

takes a scattered form: suburbanization has set in. Even before 1900, large towns started losing migrants to the suburbs. Net out-migration is now the rule for these cities. This results in a large increase in commuting.

FACTORS DETERMINING THE VOLUME AND DIRECTION OF HUMAN MIGRATION

Incentives to migration. The volume and direction of every migration stream are determined by numerous factors that may be classified as, on the one hand, incentives to migration, and on the other hand, obstacles and alternatives to migration. As has already been stated, the incentives usually take the form of interregional differences, though in some cases the push of circumstances in the region of origin or the pull of attractions in the region of destination may be sufficient.

A dwindling food supply, exhaustion of natural resources, or overpopulation are push factors that may result in migration. A pull factor leading to migration may be the discovery in one particular region of a coveted asset: for instance, gold or diamonds. A migratory movement based exclusively on push or pull factors can grow to large proportions very quickly, but it tends to endure only temporarily. As communication improves, either some alternative to migration is found or the migration reverts to the normal type based on true interregional differences. The 19th-century Atlantic migration, for example, showed a more or less push character up to about the time of the American Civil War, being largely the result of disorganization in European countries. Later, knowledge of the differences between the Old World and the New became the principal migration factor.

The interregional differences that nowadays determine volume and direction of migration streams may be divided into various categories.

Differential economic opportunities. Differences in the demand for certain types of labour or for labour in general have for a long time been the main migration incentive. Differential economic opportunities may come about as a result of discovery or exhaustion of natural resources, technical innovations, development of new trade routes or establishment of new frontiers, changes in the division of labour, or changes in taste.

Differential standards of living. Although one would expect differences in real income between various countries or regions to be relevant to the volume of migration, in statistical studies their effect is in fact usually found to be nominal.

Differences relating to housing and the residential environment. This factor has recently become more important in determining the volume and direction of migration. It forms the basis not only for the entire large-scale suburbanization movement but also for much of the migration of elderly people. Aspects of the residential factor are quality of dwellings and of residential neighbourhoods, climate, and scenery. Where an absolute shortage of dwellings prevails, such as was the case in many European countries after World War II, differences in the availability of dwellings also play a role.

Differences relating to social amenities. Such urban amenities as recreational facilities (theatres, cinemas, concert halls), institutes of higher education, and specialized hospitals certainly influence the migration decisions of some individuals. Statistically, however, they have been found to contribute only marginally to the explanation of the volume and direction of migration streams. Modern urbanization is clearly based on employment factors and not on "the lure of the city lights."

Cultural and political differences. Often a minority in a country prefers the cultural and political values that obtain elsewhere. This may be a reaction to more or less revolutionary cultural and political changes in their country—as was, for instance, the case with regard to the many Germans who left Germany during or shortly after 1933. It may also be, however, that the values and views of the minority itself have changed, while the dominant culture remains more conservative. Many Europeans left for North America from the 17th to the 19th centuries because of the religious freedom in the New World.

Obstacles to migration. Obviously for migration to take place conditions in the potential region of destination must be evaluated by the migrants as the more favourable. Interregional differences that are negatively evaluated act as obstacles to migration.

Vested interests. Irrespective of interregional differences there is a factor that is likely to exert a negative influence on migration. To a person who has established himself in a certain region, with personal property, contacts with relatives and acquaintances, and familiarity with local customs, the prospect of migration may be less attractive. These personal ties between the individual and his own region may be termed "vested interests." It is impossible to measure the inverse relation of vested interests to the volume of migration streams, but studies of nonmigrants and their motives for not migrating suggest that vested interests may exert considerable influence.

Personal and cultural obstacles to migration

Defective communication. Another major obstacle to migration is lack of information concerning opportunities in potential regions of destination. This is a consequence of defective communication resulting from great physical, technical, or social distances or other barriers to communication. The latter include national or linguistic boundaries or historical partitions that have not yet lost their traditional influence. The volume of international migration is thus generally much smaller than that of internal migration, except when the social distance between two countries has been greatly lessened by prior migration and there is no language difference, as in the case of Ireland and England. If, on the other hand, a linguistic boundary exists *within* a country, its effect on migration is clearly discernable. In Belgium, for instance, the effect of the boundary between the Dutch-speaking and the French-speaking part of the country was found to equal that of a distance of 50 miles.

Transportation difficulties. Migration may be further curbed by difficulties of transport—to which, of course, distance may contribute. These difficulties may be technical or financial. In the case of intercontinental migration, the expense of the journey not infrequently limits the number of migrants, especially from the poorer countries. In recognition of this problem, some governments give financial aid to emigrants. Certain countries anxious to attract immigrants, in particular Australia and South Africa, also offer to defray part of the travelling expenses.

Emigration and immigration restrictions. Finally, among the obstacles to migration should be counted social norms that are opposed to emigration or, in the country of destination, to immigration. Such norms may be of an informal nature, but often they are formalized into official emigration or immigration restrictions. In the United Kingdom, for instance, laws against emigration were passed in 1719, 1750, and 1782. In Japan emigration was strictly prohibited from 1636 to 1868. Military considerations prompted many German states to restrict emigration severely in the 19th century.

Political restrictions on migration

Immigration restrictions are, broadly speaking, of two types: either foreigners need a permit to enter the country or they may freely enter but need a permit to take a job. The latter system is usually applied in Europe. As of Nov. 8, 1968, however, all migration restrictions have been abolished within the European Economic Community, so that nationals of any of the six Common Market countries may now freely settle and work in the other five. Within the European Free Trade Association, too, immigration restrictions have been slackened.

The English-speaking immigration countries outside Europe usually rely on an entrance visa. For a long time they were inclined to select immigrants on the basis of national preference, favouring Europeans with an emphasis on citizens of the United Kingdom. The United States institutionalized this preference in 1921 with the introduction of the quota system. The maximum of the nationals of any country to be admitted in any one year was fixed, under this system as revised in 1924, relative to the percentage of persons originating from that country resident in the United States in 1890. The quota principle was retained, with various alterations, until 1968, though in fact during the 1950s more immigrants were admitted outside the

quotas, under special laws, than were admitted within the quotas. As of 1968 the quota system was abolished and preference was given, within an annual grand total of 170,000, to relations of U.S. residents, refugees, and highly trained workers.

The Commonwealth immigration countries, Canada, Australia, and New Zealand, never had an official quota system, but their preference for Britons and other northern Europeans was unmistakable. In the 1960s these countries changed their policy to give preference to highly trained and skilled workers, regardless of national origin. Meanwhile, the United Kingdom in 1962 introduced severe restrictions on the entrance of citizens from the Commonwealth and colonies, who had formerly been freely admitted.

Migration models. Over the past four decades, demographers and statisticians have developed a number of analytical techniques and mathematical models to measure the influence of the various factors mentioned above on the volume and direction of migration. The more sophisticated of these models are highly successful in explaining variations in the volume of migratory streams. It is generally found that interregional differences relating to housing and the residential environment are mainly responsible for local moves and short-distance migration, whereas migration over longer distances is predominantly determined by differential employment opportunities. Other incentives to migration play a subsidiary role, except in those international movements that are principally politically motivated. In nearly all instances, the volume of migration streams shows a strong inverse relation to distance, indicating that communication is also a highly relevant factor.

The influence of distance on migratory patterns

That the same interregional differences may be evaluated quite differently by various persons is evinced by the fact that migration is nearly always a two-way movement. The streams in both directions between two regions often do not even differ much in size, at least in internal migration. To some extent return migration of migrants who have not been able to adjust to conditions at the destination also plays a role in bringing this about.

Alternatives to migration. Migration models do not usually encompass variables indicating alternatives to migration. Even so, it is clear that such variables do effect the volume of some migration streams. If the goals sought through migration can also be reached through other means, these alternatives may be chosen by all or some of the potential migrants.

Population control. Where, for instance, emigration is the result of overpopulation in relation to natural resources, population control is an alternative. Historically this has taken the forms of infanticide, late marriage, celibacy, and in modern times birth control.

Economic growth and development. An alternative to economically induced migration is of course economic growth and development. This is relevant for both external emigration and internal migration. In many European countries, regional economic development policies are now in force to combat migration from less developed regions to highly congested urban areas. In Britain, for example, such policies have as an objective the mitigation of congestion in the London area and the Midlands.

Social and economic reform. Social changes may sometimes be needed to curb emigration, especially if the migrants are motivated by wide distinctions between the social classes in their country of origin. Various measures, as for instance redistribution of landed property, may remedy this situation. Potential emigrants may consider "class struggle" or political action resulting in social and economic reform as alternatives to migration. The finding that in Italy before World War I those regions that did not show heavy emigration were characterized by a militant labour organization may be cited as an instance.

Commuting. One alternative to short-distance migration that is becoming increasingly popular is commuting. The decision to commute rather than migrate to a job may be based on various motives, of which vested interests—especially in the sphere of social relations—probably play the largest role.

ADJUSTMENT AFTER MIGRATION

Having migrated, the migrant must adjust to his new environment. In the case of international migration, the process of adjustment is often long, difficult, and painful. Not all migrants succeed in adjusting satisfactorily.

When the region of destination was uninhabited before the migrants arrived, they need only adapt to the physical circumstances. The same is true if the migrants drive out the original inhabitants or if these are relegated to certain limited areas, as in the case of the Indians in North America. There are indications that settlements in uninhabited areas tend to be innovatory, in the sense that new ideas that are latent or dormant in the community of origin are put into practice in the new land.

Assimilation and pluralism

In tribal as well as in individual migrations if the migrants are relatively numerous and from the same cultural background, the immigrants may impose their culture on the original inhabitants. Such a cultural conquest is, however, rare. Usually some elements of the native culture are retained. A cultural fusion then takes place: the native and migrant cultures become mixed. After the Germanic migrations, for instance, there were fusions of Germanic and Roman cultural elements.

A third possibility is that the native and migrant cultures, though affecting each other, retain their independence. In that case one speaks of cultural pluralism. South Africa and Rhodesia are countries in which cultural pluralism subsists as a political ideology.

In most cases nowadays, however, the immigrants in a country are individuals with varying cultural backgrounds. Assimilation is the process whereby these immigrants are absorbed into the culture of the native society. Even in the case of assimilation of individuals, however, the passage of culture elements is not a one-way affair: the immigrants also contribute in varying degrees to the dominant culture. The result of this "melting-pot" is not uniformity but rather the incorporation of a variety of persons into the society.

The success of assimilation depends on several factors. Among them are the migrant's personality, his cultural customs and values, his motives for emigration and expectations of the new land, and, on the other hand, the values and norms of the receiving society, especially the prevalence of prejudice and discrimination. Language is an important intermediary factor. The governments of modern immigration countries such as Australia or Israel actively support assimilation.

BIBLIOGRAPHY. Fairly extensive reviews of various aspects of migration may be found in J.A. JACKSON (ed.), *Migration* (1969); and C.J. JANSEN (ed.), *Readings in the Sociology of Migration* (1970). There are numerous books and articles that treat one or more of these aspects more thoroughly. Some examples are: On international migrations, A.T. BOUSCAREN, *International Migrations Since 1945* (1963); on internal migration in a highly developed country, H.S. SHRYOCK, JR., *Population Mobility Within the United States* (1964); and on migrations in a less developed part of the world, H. KUPER (ed.), *Urbanization and Migration in West Africa* (1965). The important international migrations have mostly found chroniclers. The development of European mass migration to North America is treated by M.L. HANSEN in *The Atlantic Migration, 1607–1860* (1940). The migration from European to Asiatic Russia is dealt with by D.W. TREADGOLD in *The Great Siberian Migration* (1957). Intra-European migrations in the period of the World Wars are analyzed by E.M. KULISCHER in *Europe on the Move* (1948). On immigration in Australia, see R.T. APPLEYARD, *British Emigration to Australia* (1964). Some classics in the field of migration are: On migration and distance, E.G. RAVENSTEIN, two articles entitled "The Laws of Migration," *Journal of the Royal Statistical Society*, 48:167–227 (1885), and 52:241–301 (1889); on differential migration, D.S. THOMAS, *Research Memorandum on Migration Differentials* (1938); on assimilation, S.N. EISENSTADT, *The Absorption of Immigrants* (1954); and for an excellent analysis of free, chain, and mass migration, see J.S. LINDBERG, *The Background of Swedish Emigration to the United States* (1930).

(H.t.H.)

Milan

The destiny of Milan (officially Milano), like that of many of the world's great cities, remains, from several

viewpoints, something of a historical paradox. There are powerful factors supporting the argument that Milan should have become the capital of a unified Italy, and this is the belief of many Milanese, in spite of the fact that Italy's unity was actually born in Turin, rather than in Milan, in 1870. The Milan of today is nevertheless the most industrious and vital city to have achieved prominence since the ancient land of Italy became aware of itself as a modern nation. Contemporary Milan is, without doubt, the richest city of Italy and one of the richest of Europe, so far as money is concerned. When the Milanese assert that Milan is the moral capital of Italy, they not only express the ancient regionalism typical of all Italy, and known as *Campanilismo* (a reference to the church bell of each city), they also refer to something intangible and yet authentic, for they are speaking of quality and values, historical as well as contemporary. And if the rest of Italy, Rome included, accepts this statement, or rather accepts the fact that the statement is made, it is because it is more than a simple claim. It is a claim justified by the contributions in every field—economic, cultural, and ideological—that the city of Milan, in modern times, and particularly since the unification of Italy, has made to the Italian nation. These contributions greatly exceed, even on a statistical basis, those made by all other Italian cities. When one remembers that, in the last century, a writer such as Stendhal, one of the giants of French culture, wished to proclaim himself "Milanese" on his epitaph, one must indeed believe in the fascination Milan exerted then, and still does, and of which the city is fully conscious. The fact, then, that Milan is at a distance from much of the rest of Italy, that it is peripheral in a geographic sense, does not explain its position of second city, a position it has always pathetically and vainly fought. Some of the greatest European capitals are peripheral in this sense.

Competition with Rome

This role was the consequence of the immense historical importance, of the enormous accumulation of myths and symbols, past and present, which conferred on Milan's antagonist, Rome, an inevitable prestige, since it became the heart of a future anticipated in the collective fantasies of the Italian people of yesterday and today. This character is fundamental because a capital is not simply the centre of such and such a government or such and such administrative offices. Milan was, in the 19th century, the most European among Italian cities, but it was not strong enough to become the centre of Europe. On the other hand, Europe was still, during the last century—and perhaps this remains true today—too feeble to seek and find its own centre. European myths, symbols, and values were, and perhaps still are, too weak to let Milan arise as the centre of Europe in the absolute order of things that a capital demands.

It was not by chance that Milan expressed its ideological greatness in the person of a poet, Carlo Porta (1776–1821), who wrote in Milanese—that is to say, in dialect. Yet that poet risked obscurity, both in his own country and abroad, eager as he was to give of his utmost self, aware that this coincided with the finest aspirations of his cocitizens over the preceding century. Rome can absorb the values and language of a renascent Florence and integrate these values throughout the centuries, as the Rome of today demonstrates. In modern Rome, cinematography, which could never take root in Milan, in spite of a thousand attempts, employs the mystery of the physiognomy and of the light of the city to recreate those classic elements, which, in Italy, may be called antique. This Milan cannot do. The majority of its intellectuals, of its writers and artists, at least until the end of the post-World War II era, abandoned the city to rush to Rome. Milan thus remained essentially an economic centre, succeeding, however—alone among Italian cities—in keeping alive an ideological inquisitiveness and a spirit of polemic, which involved not only these two cities but all the others in Italy. One may well hazard the conclusion that Milan's destiny had an unfortunate outcome. In any other country of the world—that is, in any country free from the overwhelming presence of Rome—Milan would have been the capital.

HISTORY

Foundation and early growth. Founded by the Gauls around the year 600 BC, the earliest settlement on the site of Milan became, in ensuing centuries, the capital of a Celtic tribe known as the Insubres. At the time of the Roman conquest, 222 BC, Mediolanum, as it was then called, was already one of the most powerful cities of the region on the Roman side of the Alps known as Cisalpine Gaul. Under the Roman emperor Augustus (63 BC–AD 14), it became a part of the 11th region of Italy, acquiring increasing prestige and economic power until it became the second city of the Western Roman Empire, Rome being the first. In the 3rd century AD, following the partition of the Empire instituted by the emperor Diocletian, it was assigned as residence and main administrative centre for one of the two emperors. The emperor Constantine the Great declared it the seat of the Vicar of Italy. In the year 452, Attila the Hun devastated the city, and, in 539, the Goths destroyed it. The city, however, did not entirely perish as a result of these barbarian incursions, and by the second half of the 10th century city life was surging with renewed vigour. Under the Carolingians (the region was incorporated into the dominions of Charlemagne in 774), life in Milan showed increased vitality, particularly through the late-9th-century efforts of Archbishop Ansperto da Biassono, who rebuilt and strengthened the fallen walls of the city. Under Ariberto da Antimiano (1018–45), the political power of the archbishopric reached its apogee. This assumption of temporal power by the archbishops, dating from around 1000, can be considered as the origin of the subsequent greatness of Milan.

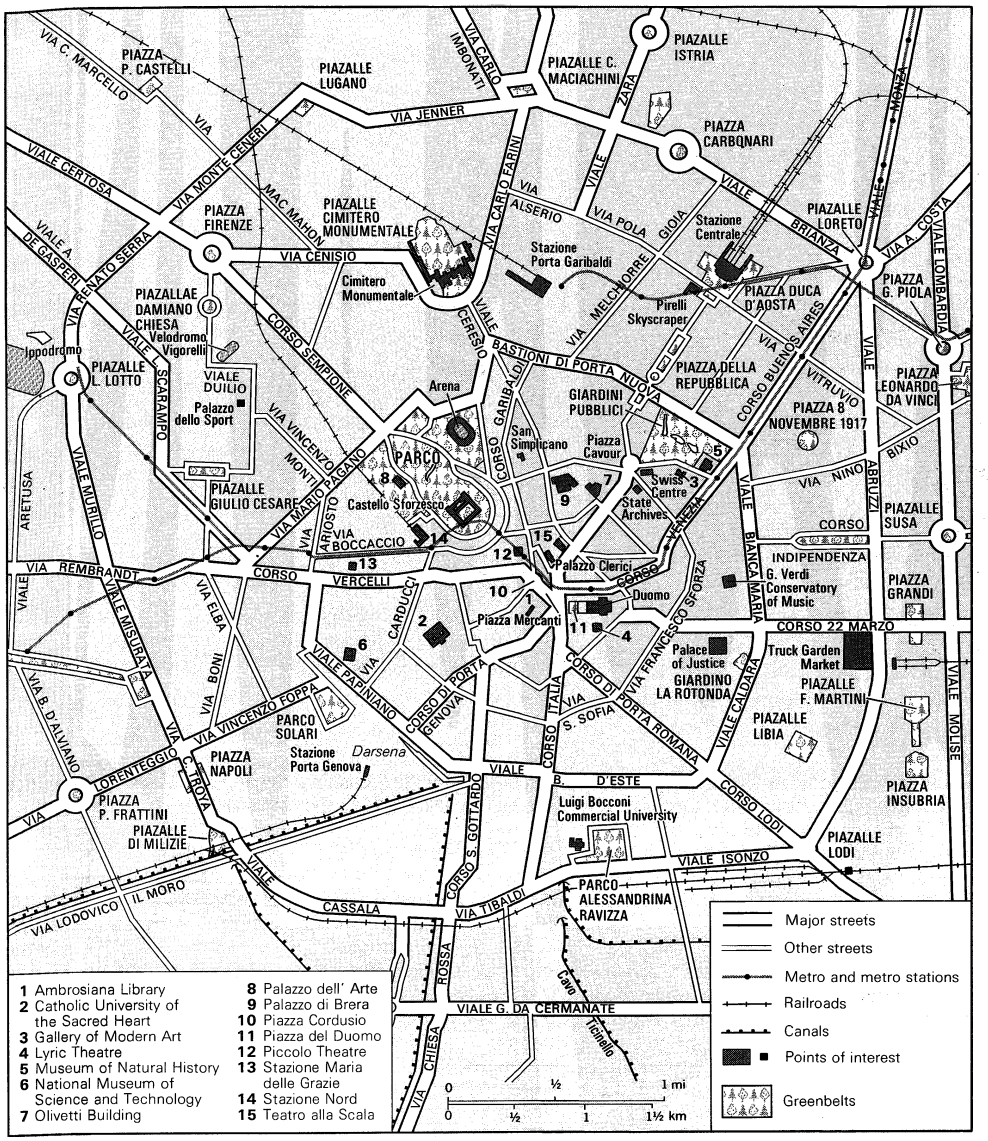
Early settlement

In 1045, however, as a result of tensions engendered by the authority of the archbishops, and because of the increasing growth and stability of the city as a whole, Milan constituted itself as a commune (*comune*), with permanent and autonomous governmental structures. In the resultant struggle for primacy among the cities of Lombardy, Milan became involved in a series of long battles against its less prosperous neighbours—Pavia, Cremona, Como, and Lodi. In 1111 the Milanese razed Lodi, and, after a bitter struggle lasting from 1118–27, Como was destroyed. This was the pretext for the intervention of Frederick I Barbarossa, who decided to bring Milan under the direct authority of the central imperial power of his Holy Roman Empire. The city held out until 1162, when it yielded after a nine-month siege. Its fortifications were then razed, and the destruction of the city was such that the Milanese were forced to seek refuge in the surrounding countryside. The war blazed on until 1183, the year of the Peace of Constance, although Milan, rebuilt in 1167 under the auspices of the newly founded Lombard League, succeeded in playing a major role in the defeat of the German forces of Barbarossa at the Battle of Legnano in 1176. Its privileges rewon, the city attained a splendid economic florescence over the next hundred years.

Establishment of the commune

Feudal conflicts. In the early years of the 12th century, however, the new industrial classes, in particular the guilds of the woollens and armaments workers, increased constantly in power and influence. The feudal nature of the relationship between the archbishop and his allies meant that the archbishop had to make enormous concessions to the emergent social and political forces among the citizenry in order to reinforce his own party, diminishing thereby the financial privileges of the church. Following the worsening of the relationship with Frederick II of Swabia, the Milanese proclaimed Pagano della Torre, a member of a family emerging as leaders of the less feudal of the city's power groupings, as their protector. The city forces were nevertheless defeated by the emperor in the Battle of Cortenuova (1237). In the shadow of the subsequent struggle between the Torriani family and another powerful Milanese family, the Viscontis, the Signorial era was born. The Torrianis, leaders of the new popular forces, took the name of Guelfs; the Viscontis, followed by the aristocracy, headed the Ghibelline faction. In 1277, Ottone Visconti, archbishop of Milan, utterly defeated the enemy in the Battle of Desio. His nephew Matteo succeeded him, and, starting in 1311, Matteo and his heirs

Dynastic struggles



The city of Milan.

reigned as supreme lords of the city and of the surrounding state, replacing the political forms of the commune. Under this lordship, or *signoria*, the industrial and mercantile economy underwent rapid development, giving birth to further powerful coalitions of economic interests. But in 1450, Milan found itself besieged again. Francesco Sforza (1401–66), a ruthless and ambitious general, now occupied the city and founded a new dynasty, basing his claim on his marriage to an illegitimate daughter of one of the Viscontis. A period of great prosperity then began for Milan, based on the power of the Sforzas and the introduction of the silk industry. It was the golden period of the Italian Renaissance, typified by the splendour of the Sforza Court. The dynasty of the Sforzas had but a short-lived enjoyment of power, however. In 1499, the Duchy of Milan fell into the hands of Louis XII, king of France, and also a distant descendant of the Viscontis. In 1500 Ludovico il Moro conquered the state but was defeated at Novara in the same year. The French continued to rule until 1513, when they were overthrown by Massimiliano Sforza, son of Il Moro, who had Swiss assistance. Francis the First, successor to Louis XII, reconquered Milan in his renowned victory of Marignano (now Melegnano), 1515. After a peace treaty signed in 1529, Milan was once more returned to the Sforzas.

The modern period. In 1535 the incumbent duke died unexpectedly, and Milan and the entire Milanese state fell under the domination of the Habsburg emperor, Charles V, who in 1540 invested his son—the future Philip II of Spain—with the duchy. Under the Spanish rule —which was to last until 1706—the political and artistic elite of Milan rapidly succumbed: the dramatic period of dynastic struggle, which was also a period of economic growth, was to be replaced by a long period of economic stagnation and political decline associated with unimaginative foreign rule. In 1630 the city was struck by the great plague. This catastrophe was to be vividly portrayed by the local author Alessandro Manzoni (1785–1873) in his *I Promessi Sposi* ("The Betrothed"), a historical novel that is one of the finest artistic achievements of Italian and indeed European literature. The end of the desolation and squalor of the period of Spanish domination began with the outbreak, in 1701, of the War of the Spanish Succession, following the death of Charles II of Spain. On Sept. 24, 1706, Prince Eugene of Savoy entered Milan as its first Austrian governor, and the city passed thus from Spanish to Austrian rule. Although the first half of the 18th century was marked by neglect and oppression, after the Treaty of Aix-la-Chapelle (1748), the new rulers, in collaboration with the wealthy commercial classes of Milan, were able to foster a half-century of enlightened, if despotic, growth. This was particularly marked in the cultural domain. It is during this period that such figures as Cesare Beccaria (1738–94), the outstanding criminologist and economist, and Pietro Verri (1728–97), the gifted administrator and man of letters, were active. These and other members of a Milanese group known as the Società dei Pugni accepted the innovations of the the-

Apartment buildings and downtown area of Milan (foreground); Duomo, the Milan cathedral (background).
By courtesy of Ente Provincial Per Il Turismo, Milan

oreticians of the French Revolution, and this in spite of Austrian censorship. Neoclassical architecture also flourished. When, on May 15, 1796, the republican army of France, with Napoleon Bonaparte at its head, entered Milan, it was greeted enthusiastically, particularly by the middle classes. In 1797, the constitution of a Cisalpine Republic was promulgated. In 1805, Milan became the capital of the Kingdom of Italy, under Napoleon, who was crowned in the city. A true awakening then occurred in the consciousness of the Milanese people, as their city prospered from its domination of most of the Italian peninsula.

The city under Napoleon

These hopes were dashed, however, by the invasion and re-establishment of Austrian authority that followed the collapse of the Napoleonic Empire and the 1815 Congress of Vienna. This alien control of the new Lombardo–Veneto kingdom was to remain for nearly 50 years. Influenced by the new currents of Italian unity and nationalism known as the Risorgimento, and smarting under the oppressive Austrian rule, the citizenry finally rose up in the "Cinque Giornate," the five days of March 18–22, 1848. In one of the most celebrated episodes of the city's history, Milan was liberated from the Austrians for several months until the rebellion was finally smothered. In spite of the fact that by August 6, 1848, the brutal occupation forces of the aging Austrian commander Joseph Radetzky were once more in firm control of Milan, the city decided to continue its resistance to the invaders. Young men crossed the borders of Piedmont and Sardinia to enter its army. It remained for the second War of Italian Independence to finally liberate Milan from foreign control, and a few days after the Battle of Magenta (June 4, 1859), Milan witnessed the triumphant entry of the two allies, Victor Emmanuel II and Napoleon III. The city—by now in the throes of an industrial revolution emphasizing metal products—was henceforth to be linked with the fate of the new Italian state, maintaining itself in a position of prime importance in the national economy. It also shared in the political development of the nation: on March 23, 1919, the formation of militant right-wing groups in the city marked the dawn of Fascism. Intense efforts at reconstruction healed the wounds of World War II, when the city suffered grave damage from intensive Allied bombing. The social and economic conflict that marked much of the city's development found expression in the existence of a strong Communist Party, whose influence continued unabated from the immediate postwar period through the ensuing decades of relative affluence (see ITALY AND SICILY, HISTORY OF).

Restoration and revolution

THE CONTEMPORARY CITY

Physical layout. Milan is set in the heart of the Po Basin of northern Italy, halfway across the immense plain spreading between the Ticino and the Adda rivers. The site is 400 feet (120 metres) above sea level. To the north lies the great sweep of the southern flank of the Alps. Between this semicircle of mountains and the course traced by the Po River to the south, there lies a zone that is arid toward the north, but swampy near the Po, where it turns into an expanse of marshy groves and rice fields. It is at the line of demarcation between these two areas, which are strongly differentiated, that Milan has risen, although these are now only swamplands marking the site of the ancient city. The earliest inhabitants reinforced their defenses by means of the small watercourses of the Seveso, the Nirone, the Lambro, and the Olona.

The physical setting of the city

Each period of historical crisis, advance, and consolidation has been reflected in the organic structure of Milan. For a thousand years the core of the city was located just southwest of the present cathedral, the Duomo (see below, *Cultural life and recreation*), and was made up of the rectangular, four-gated city of Mediolanum, with roads thrusting out from each gate to the surrounding countryside, together with an irregular outer defense consolidated in Carolingian times. This core has influenced the city plan down to the present. The period of dynastic struggle and the imposition of transalpine authority brought further changes. After the city was razed in 1162, an enlarged oval was constructed, the course of its outer walls still traceable in contemporary streets. Spanish domination brought the erection of still another outer ring resulting from their 16th-century reconstructions. This, too, can be traced in contemporary boulevards. Within the city centre, the main focus of activity centred on the Castello Sforzesco, a product of the 15th-century dynastic struggles, reinforced by the Spanish in the following century; the Piazza Mercanti, the centre of medieval economic activity; and the great Piazza del Duomo, laid out before the cathedral in 1489. Castle, cathedral, and a newer commercial area centred on the Piazza Cordusio, appropriate representatives of the motivating forces in Milanese life, have remained to dominate the modern city centre. In 1884, 1912, 1932, and again in 1950, city planners laid

The basic plan

down the basis of a more organic plan, bypassing the radial heritage, in order that new districts might enjoy wide streets and avenues intersecting at right angles. The centres of the newer suburban areas—Bollate, Novate Milanese, Cusano Milanino, Cinisello Balsamo, Sesto San Giovanni, and even Monza (15 kilometres [9 miles] away to the northeast)—are linked to the core of the ancient city by major arteries. Entire industrial districts have developed, particularly in the north and northeast and in the south and southwest. Unfortunately, while the periphery has been amply developed, the central nucleus has not benefitted from a similar transformation and the narrowness of the streets and squares creates great problems for modern traffic. In the urban design of modern Milan practically all industrial growth has been concentrated in peripheral areas. Industrial progress has reached such a level that a third of Milan's population is in some way connected with industrial construction. Heavy industries are concentrated outside the city boundaries, around Sesto San Giovanni. Since 1950 the city centre has fared badly under the influence of industrialization, with many old streets and buildings replaced haphazardly by massive blocks and skyscrapers containing apartments and offices. By the 1970s, the Pirelli Skyscraper, near the Central Station, and the Olivetti Building in Via Clerici were among the few structures of architectural note.

Demography. Milan's population has shown rapid increases over recent decades. In 1931 it had 1,115,841 inhabitants. In 1951 it was 1,274,245. In 1961 the figure rose to 1,599,008, and in 1970 it was 1,713,539. This increase has been due mainly to the flood of immigrants from the impoverished Italian south seeking improved conditions in the factories of the industrial north. During peak years, the annual total of such immigration has exceeded 70,000. Population pressure has resulted in the growth of slums and such shantytowns as Brianza, as well as in an expansion of the city itself. Milan now presses outward into the surrounding countryside with renewed vigour. This is particularly obvious in the northeast, toward Crescenzago, and in the south, toward Rogoredo and Vigentino. Milan's enormous urban expansion has contributed to its economic growth. It is the major economic centre of Italy. With an area of 2,762 square kilometres (1,066 square miles), the province of Milan covers but 0.92 percent of the national territory. The 247 communities that constitute it, however, make it one of the most densely populated provinces of Italy, with a density of 1,377 (1970) people per square kilometre. In addition, its 3,800,000 inhabitants, about half of them concentrated in the capital, represent nearly 7 percent of the population of the entire country and make it the most populous of the Italian provinces. The working population represent more than 50 percent of its inhabitants. Taking into account only the age range between 15 and 65, there is a working population of 75 percent—a very high average for a country such as Italy, which normally offers few work opportunities.

Immigration

Economic life. Milan, the most important economic centre of Italy, owes this fact to its geographical position. Located at the centre of the traffic routes of the Val Padana, it is on the borderline between the advanced agriculture of the fertile irrigated plains to the south and the limited agriculture of the north. In addition, it is backed by an impressive industrial development. Milan's rich and extremely populous surroundings, differing from those of other great Italian manufacturing centres, increase the city's economic importance and potential for expansion. A rich network of road and rail communications spreads toward the outlying areas and particularly toward the north, giving an economic advantage that other Italian cities do not possess.

If the province of Milan represents around 7 percent of the national population, its economic importance is proportionately nearly double; *i.e.*, its economic and financial potential is over 12 percent of that of the whole country. Today the mechanical industries predominate: car production, airplanes, motorcycles, major electric appliances, railroad materials, and other metalworking account for 45 percent of the manpower. Textile manufactures (cotton, hemp, silk, and artificial fibres) are situated in the province, while in the city itself the clothing and custom-made industries predominate. Chemical production is also considerable. Besides producing great quantities of medicinal products, Milan also serves other industries by the manufacture of dyes, soaps, and acids. A noteworthy contribution to the progress of the city economy has also been made by the graphic and publishing industries, as well as by food, wood, paper, and rubber products. Commercial activity is stimulated by industrial development since the manpower involved in commerce has increased from 46,000 in 1911 to 140,000 in 1951 and 190,000 in 1961. The greatest wholesale markets of Italy are also in Milan. Most important among these are the export trades; among them artificial fibres, cotton and wool goods, chemical products, and machinery. Naturally, this enormous economic complex has benefitted from the efficiency of the banks and of the stock market, the latter being the largest in Italy. Milan is the principal centre of exchange in Italy and one of the major economic areas in Europe.

Industrial products

Political and governmental institutions. The city, technically the commune of Milan, capital of the province (*provincia*) of the same name and of the Lombardy Region (*regione*), is under the jurisdiction of the Republic of Italy. The government of the Italian state is composed of a central administration and of a peripheral or local administration system, the first controlling the whole territory, the latter having a limited type of jurisdiction over a limited region. The most important division of the peripheral system is the prefecture (*prefettura*), at the head of which is the prefect, the highest authority of the executive power in the province of Milan. It supervises the activities of public administration. In its own area, the province performs the important functions of regulating health, sanitation, public works, welfare, education, agriculture, and commerce. The province of Milan, as all other Italian provinces, has a vice prefect, a counsellor of prefecture, and an administrative and provincial assembly whose members, as well as its president, are selected by the prefect. The commune of Milan is administered by a council of 18 members, chosen from an assembly of 80 members, and a *sindaco*, or mayor. In the elections of the communal council all electors registered in the province take part in the administrative elections. The assembly and the mayor are elected by the communal council. The *sindaco* is the head of the communal administration and represents the central government of the Republic of Italy. He can perform marriage ceremonies, register births and deaths, and maintain safety and order on behalf of each citizen, etc. The region and regional council, instituted in the early 1970s, are constituted as autonomous entities and have various prerogatives. Among these are the right to promulgate laws on matters relevant to the city so long as they do not infringe upon national interest or the interest of other Italian regions. The regional authorities are the regional council, the assembly, and the president of the assembly. In the commune of Milan, the conciliation judge presides.

Regional innovations

Transportation and services. Rounding off its character as a great centre of production and exchange, Milan is also a major focus of communications. The 26 lines of the state-run railroads are integrated within the city landscape by means of a carefully designed and executed overall design plan. The vast Piazza della Repubblica, for example, which contains the Pirelli Skyscraper and other high buildings, is located on land formerly occupied by a railway station. Trade goods are sent and received at the Lambrate-Ortica Station. Two other stations serve for the transportation of passengers, five more for merchandise. The largest railway loading site is the Central Station. Transalpine tunnels and other mainline connections link Milan with all parts of Europe. The road network converging upon Milan is also important, carrying, as it does, an unceasing flow of foreign and national tourists. Besides being the focus of a number of motorways, Milan is also the starting point for the famous scenic route known as the Autostrada del Sole, which runs down the spine of the Italian peninsula. The metropolitan transportation

service is one of the most active in Europe, having a four-line network and a length of over 18 miles. The central post office functions 24 hours per day. The Urban Streets and Sanitation Department is highly modernized, with a large personnel endeavouring to maintain both the cleanliness of the city and the health of its inhabitants, a task made difficult in some instances by overcrowding and bad housing conditions.

Cultural life and recreation. The most striking of the monuments to be seen in contemporary Milan is the Duomo, a triumph of the architecture of the early Middle Ages, and the third largest church of contemporary Europe, holding more than 20,000 people. Begun in 1386, it took five centuries to complete, and rises over the area occupied at one time by the churches of Santa Tecla and Santa Maria. The most imposing parts of the Duomo are its lateral aspects, its two top crosses, and the apse. In the latter, a powerful impression is made by the three immense Gothic windows of finely carved marble. The casing, of pink-tinged Italian marble, is to be found on all sides of the structure. At the lower level, it lends character to the small trilobate arches, capitals, and flowers; it also appears on the buttresses and, above them, runs along the crowning row of gigantic statues; above these, it covers the decorated water gutters and, finally, enhances the lacelike ornamental crest. The exterior of the cathedral is in fact covered with a remarkable profusion of turrets, pinnacles, and more than 3,000 statues. Within are 52 pillars, each over 80 feet tall and more than 10 feet in diameter and bearing, instead of capitals, a crown of statues within their niches.

The Duomo

The most notable of the city's many palaces is the Brera Palace, the present appearance of which dates from 1615. Its architect, Francesco Maria Richini, infused the whole Milanese Baroque with his severe style. The Pinacoteca di Brera, of international importance, contains a fine collection of north Italian painting. The building also contains the national library, and its beautiful courtyard is dominated by Antonio Canova's statue of Napoleon. On the Corso Garibaldi, or the road to Como, stands San Simpliciano, an extremely ancient church founded, according to tradition, by St. Ambrose, in the first centuries of the Christian era. Its apse contains the famed fresco, the "Coronation of the Virgin," by Ambrogio Bergognone (flourished 1480–1523). In Via Monte Napoleone, there are several handsome palaces, including the Bagatti-Valsecchi palaces. Leonardo da Vinci's fresco the "Last Supper," one of the most famous paintings of the Renaissance, is located in the dining hall of the convent of Sta. Maria delle Grazie. Beyond the Piazza Cavour (which contains a monument to the great Italian statesman after whom it is named as well as the imposing and ultramodern Swiss Centre) lie the public gardens, in which the Museum of Natural History stands. One of its most interesting exhibits is the Turati ornithological collection. The former Villa Reale, now the Galleria d'Arte Moderna, is also nearby. The Biblioteca Ambrosiana (founded 1609 and containing over 600,000 volumes) and the Civic Library with well over 200,000 volumes, and also a stamp collection, are also of significance. Other excellent libraries are found in the universities, the philological clubs of the Lombard Institute of Science and Letters, as well as in the Donati Foundation. The State Archives also contain an imposing collection. Milan is the seat of three universities: the Catholic University of the Sacred Heart, one of the best of its kind in Italy; the State University; and the Luigi Bocconi Commercial University. Other higher educational bodies include the Enrico Fermi Institute for Nuclear Studies and the Giuseppe Verdi Conservatory of Music.

Libraries and universities

Milan's La Scala, constructed 1776–78 by the leading Neoclassicist architect Giuseppe Piermarini, is one of the great theatres of the world. (Piermarini, incidentally, designed the Corso di Porta Romana, the first paved street of modern Europe.) The city also contains the Lyric Theatre, the Odeon, the Piccolo Teatro of the City of Milan, and the Stabile Theatre, home of one of the most important theatrical companies in Italy. There are also a very large number of movie theatres.

Milan is the leading centre of Italy in terms of the equipment, organization, and development of sports. The Sport Palace is a splendid example of modern recreational architecture. At San Siro, the Ippodromo, a large horse-racing arena, is one of the best in Europe. The Mirabello course is also excellent. San Siro has a fine football stadium, a Civic Arena, and numerous other fields for various sports, among which the modern Vigorelli, for cycling races, is well known. The Grand Prix motor-racing circuit at near-by Monza has an international reputation.

BIBLIOGRAPHY. The *Storia di Milano*, published by the Fondazione Treccani degli Alfieri, 16 vol. (1953–62), is the fullest account; but G. GIULINI, *Memorie spettani alla Storia di Milano*, 12 vol. (1760–71, new ed. 1854–57), is still useful for the Middle Ages. E. NOYES, *The Story of Milan* (1908), provides an introductory account in English. For descriptions of the modern city, see F. ELMO, *Milan and its Environs* (1955); ENTE MANIFESTAZIONI (ed.), *Snapshots of Milan* (1961); and A. TIRANTI, *Milan Today* (1958). For architecture, see C. ROMUSSI, *Milano nei suoi monumenti*, 3rd ed., 2 vol. (1912–13). Current information may be found in the monthly survey and statistical bulletin, PALAZZO DEL COMUNE (ed.), *Città di Milano*.

(A.L.)

Military Law

All states require a code of laws and regulations for the raising, maintenance, and administration of their military forces. All this may be considered the field of military law. The term, however, is generally confined to disciplinary military law, that part of the code that aims at and sanctions the maintenance of discipline in the force. In the past this was also known by the name of martial law, a term that now has the meaning of military enforcement of order upon a civil population either in occupied territory or in time of disorder. Members of armed forces do not cease under modern conditions to have duties as citizens and as human beings. All systems of military law must aim to ensure that the soldier is in no way enabled to escape the obligations of his country's ordinary law or of international law as recognized in various conventions.

Historical development. The object of the disciplinary code is to ensure that the will of the commander is put into effect. Military law therefore traces its origins to the prerogative power of rulers. In Rome, just as a sector of civil law developed from the imperium of the magistrates, so did military law derive from the imperium of those same magistrates in their capacity as commanders of the military forces. The Roman historian Tacitus indicates that military justice in the 1st century AD was generally regarded as somewhat rough-and-ready and heavy-handed and varying much with the individual commander. But it became more formalized and received treatment 400 years later in the Digest and Codex of the emperor Justinian. With the rise of the kingdoms of the Middle Ages, the maintenance of discipline was enforced by ordinances or articles of war issued by the sovereign or by a commander authorized by him at the beginning of each campaign. The earliest now extant are those of the English king Richard I in a charter of 1189 AD for the government of those going to the Holy Land.

Articles of war

With mercenary armies drawn from many nations in the wars of the 16th and 17th centuries, each national contingent tended to apply the articles of the supreme commander according to its own rules of procedure. The articles of war of Maurice of Nassau, prince of Orange, and Gustavus Adolphus had a considerable influence on the national commanders who served under them, when they came to command elsewhere. In the English Civil War, the ordinances of the royalist and the parliamentary commanders are thus in the most part literally the same and in the next reign formed the basis of Prince Rupert's code of 1672. The famed discipline of Cromwell's army was not due to any improved code but to the fact that the articles were rigorously enforced. On the continent of Europe, the articles of Gustavus Adolphus continued to be followed until supplanted by the codification of the 19th century, which established throughout those countries a generally similar system that, with revision and amendments, continues to this day.

With the introduction of a standing army in England in 1689, Parliament aimed to prevent this force coming under complete control of the sovereign by a series of mutiny acts, normally passed annually, to which the prerogative articles were subordinate. By a statute of 1717 the power to make articles was embodied in the act. In the United States in 1775 and again in 1806, articles of war were adopted that were modelled upon the mutiny acts and articles then in force in Great Britain. In the British army, the articles of war were replaced in 1881 by an annually renewed Army Act (reformed in 1955), although they continued in the Royal Navy until 1957. In the United States they were replaced by the Uniform Code of Military Justice in 1951.

Persons subject to military law. The jurisdiction of military law is not necessarily confined to offenses injurious to the discipline of the forces committed by members of those forces. It extends in various countries in varying degrees to all offenses committed by members of the forces and to offenses injurious to discipline committed by persons who are not members of those forces. In countries in which an obligation to military service exists, soldiers who fail to answer their initial call-up or report for duty are liable to military jurisdiction for such offenses as desertion or self-mutilation either because the military code makes such offenses applicable to them as a class of civilians (as in Belgium, France, and Italy) or because under the act introducing national service they are deemed to be enlisted on the dispatch of a calling-up notice (as in Great Britain). They continue to be liable for such offenses, even if not otherwise subject to military law, during authorized absence from the conscripted service or temporary reserve service. Reservists are also subject (as in Italy) to military jurisdiction for such offenses as treason, communicating with foreign countries, and revelation of official secrets. In Belgium and France, released soldiers remain liable for rebellion or offenses against superiors committed within one year and six months, respectively, of their release.

Civilian liability to military law

Civilians may become liable to military jurisdiction in a number of ways. In some countries, offenses in the nature of treason, rebellion, impeding military efficiency, or fraud in relation to service property are dealt with under the military code (as in Spain, Italy, and Turkey); in others, such offenses by civilians can only be committed in theatres of operations, in a war zone, or during a state of siege (The Netherlands, Norway). Under such circumstances in other countries nationals may be entirely subjected to the military code (Argentina). In other countries, civilians are liable for instigating and participating in military crimes (Germany, Italy). Austria recognizes no military offenses by civilians.

Among those who fall under military jurisdiction are prisoners of war. Sometimes, as in France, Belgium, and Luxembourg, they are expressly included among those to whom the ordinary military law applies; elsewhere, special regulations concerning their behaviour and trial must be passed. Under the Geneva Prisoners of War Convention of 1949, prisoners of war must be tried by a military court, except where the laws of the belligerent expressly allow a member of the belligerent's armed forces to be tried by a civil court for the same offenses. Prisoners of war must not be sentenced to any penalties other than those which might be inflicted on members of the forces of the detaining power for the same act.

Offenses against military law. The military law of the Anglo-Saxon countries and of countries deriving their military law from them, such as India and other independent members of the British Commonwealth, differs from that of the majority of the continental countries in that the latter divide military offenses into two classes: crimes that are the subject of judicial punishment and, second, breaches of discipline that are subject only to administrative action. The former group of countries (and a few others such as Spain), however, recognize no distinction, regarding all military offenses as crimes. Apart from offenses of a peculiarly military nature, such as mutiny, insubordination, desertion, and misconduct in action or in performance of service duties, when an act committed by a soldier constitutes an offense in the civil code, it will frequently constitute an offense of which military law takes cognizance. In the Soviet Union, Belgium, and The Netherlands, all civil offenses committed by soldiers, except very minor ones, are tried by military court; in France, only those committed on duty and in barracks; and in Italy, none. In Germany, Austria, and Scandinavia, in peacetime, all crimes, military or civil, are dealt with by civil courts. Great Britain, Canada, and other countries include as military crimes all actions committed by soldiers anywhere that would be offenses against the criminal law of their own country, although the most serious of these cannot be tried by a military court unless committed abroad, or in India at specified Frontier Posts. In the United States, because of the differences between the criminal law of different states, certain civil crimes are specifically made offenses against the military code. All countries have rules to prevent the double jeopardy of an offender being punished both by civil and military jurisdiction for the one act. Generally, when civil jurisdiction may be exercised, this takes precedence over military jurisdiction.

Summary punishment. In both Anglo-Saxon and continental systems, soldiers are subjected to penalties imposed summarily as well as to those imposed by courts. In the majority of countries, summary penalties can be inflicted only by officers not lower than the rank of captain, the commanding officer of the military unit being the principal source of discipline. The forms of punishment so inflicted are normally loss of privileges for a specified period, fines, or deprivation of liberty. Higher military commanders usually have power to deal summarily with officers (normally up to the rank of major), though in some countries these will not be liable to loss of liberty.

In Great Britain, the United States, and other common-law countries, there is usually a right of appeal against summary punishment awarded through the military chain of command and extending to the highest authority. In other countries, the appeal will lie to a tribunal. In still others, the soldier may have a choice between appeal through the chain of command or to judicial authority (Norway, Sweden). It is to be noted that in some cases the military superior has the power to increase the punishment (Norway, France).

In the Anglo-Saxon countries, offenses that are beyond a commanding officer's powers are dealt with by a service court (court-martial). In the continental countries, military crimes and similar offenses are also dealt with judicially. In the latter, however, there is an intermediate form of tribunal that deals with the more serious forms of breaches of discipline and may impose punishments affecting a soldier's career, such as loss of rank or appointment, dismissal, or forfeiture of pension rights. In Germany, military courts of service, which also hear soldiers' complaints, may impose career sanctions. Also to be noted are the unit assemblies in the Soviet Union, which are general assemblies of all military members of a unit. All those attending are invited to comment on the behaviour under investigation. An elected tribunal, normally of five, decides whether the public examination is in itself sufficient educational corrective of the aberrant behaviour or whether the offender should be reported to a competent commander for disciplinary measures.

Examination by unit assemblies in the Soviet Union

Courts-martial. Military courts follow a more formal judicial procedure than that observed in summary proceedings. Before trial by court-martial or its equivalents, there is always some formal investigatory procedure. This is normally conducted by a military magistrate and set in motion by a procurator who corresponds to the official responsible for initiating prosecution on behalf of the public in civil trials.

In Great Britain and in countries of the Commonwealth directly deriving their military law from it, the evidence is reduced to writing and considered by the accused's commanding officer, who may then apply for a court-martial to try the accused. In the United States, the preliminary investigation has affinity with civil proceedings before a grand jury.

Pending trial, all countries maintain a presumption of

the accused's innocence. He must be allowed full facilities for preparing his defense, and, where disciplinary requirements necessitate keeping him under arrest, this can only be done under the mandate of a magistrate, normally for limited periods only. When the accused's commanding officer is empowered to order arrest, the progress of the case must be reported at frequent intervals to a higher authority, who will consider whether the arrest should be suspended or modified. Courts-martial are generally composed, depending on the type of case, of not less than three, five, or seven judges. In general they consist of military officers, although in some countries other ranks may be included. In France, for example, underofficers may sit upon a court, and in the United States, if an accused enlisted man so requires, at least one-third of the court must be composed of enlisted men. In Israel, a private soldier may demand one judge of his own rank, and in the Soviet Union, the court includes popular assessors elected for two years by all the soldiers in a unit. Most countries include trained lawyers in their courts. In France, a civilian judge sits with six military members. In The Netherlands, there is a civilian president nominated by the crown for life, in Belgium a civilian judge on a three-year tenure, and in Italy a military magistrate. Each sits with four officers. In the Soviet Union, presidents, vice presidents, and members of military courts are chosen by the Presidium of the Supreme Soviet from citizens on active service aged 25 or over. In courts of first instance, one of these will sit with two popular assessors. In other countries, a court may have a legal adviser who has no vote as a member of the court, as in Great Britain and the United States.

Grades of courts-martial

In some countries there are grades of courts-martial with varying competence as regards persons whom they may try or punishment they may impose. In the United States and Great Britain, general courts-martial composed of not less than five officers with a legal adviser may deal with all persons subject to military law and pass any sentence authorized by the code; special courts-martial (U.S.) or district courts-martial (Britain) consist of at least three officers and have limited powers. In the Soviet Union there are inferior, intermediate, and superior courts. Although under the Anglo-Saxon system, in cases of minor importance, prosecution and defense may be conducted by regimental officers of no legal qualifications, in the majority of countries, the prosecution will normally be in the hands of a legally qualified official, known variously as commissioner, fiscal general, auditor, or military procurator.

A soldier being dealt with summarily, or by disciplinary procedure that is not regarded as judicial action, is not defended. In trials before military courts, all countries allow the accused to be assisted in his defense by an advocate, and in some countries this is compulsory. All countries permit the employment of qualified civilian lawyers. In Greece and the Soviet Union, the defense may be conducted by the family or friends of the accused. In the Soviet Union, additionally, such persons as representatives of syndicates and other social organizations may defend.

The stage at which a defender may operate varies. Normally, he may assist immediately after the first interrogation, when an accused is informed of his rights. He then has rights of intervention during the process of instruction and must be present at such features of it as the interrogation of the accused. In other countries (as in Belgium and Greece), the defender has no part in the instruction and appears only at the trial.

Appeals. Under the Anglo-Saxon system, the findings and sentence of a court-martial are of no effect until they have been confirmed by a military commander, the convening officer, or his superior. Both at this stage and subsequently, the proceedings are subject to review by the service authorities. The service authorities may be moved by a petition on behalf of the convicted soldier, and appeal from a court-martial may also lie to a superior court. In countries in which there is a hierarchy of more or less permanent military courts, appeal will lie to the tribunal next senior to that which tried the accused. In other countries, special courts of military appeal are constituted that may consist of military judges (Israel), mixed military and civilian lawyers (Italy, The Netherlands, Portugal), or purely civilian lawyers (United States and Great Britain). The majority of appeal courts concern themselves only with the legality of conviction rather than with questions of sentence. This is particularly so when there is recourse by way of appeal to the service authorities.

In continental countries, legal questions may be referable to the highest courts. In France, except when judgment has been given in default, the Court of Cassation is the sole court of appeal. In Belgium and Great Britain, final appeal may be made on legal grounds to the Court of Cassation and the House of Lords. Normally, appeal lies only with the defense, but sometimes, as in the Soviet Union, the prosecution has the power to appeal either against the original finding and sentence or on a question of law.

Appeals to ombudsmen

In a few countries, representations about conditions of service and applications for advice and help outside the normal service channels may be made through specific officials. In Norway, a military ombudsman was introduced in 1952. This official sometimes raises questions on disciplinary and penal offenses. The military ombudsman of Sweden, created in 1915, takes note of all sentences of military courts, reports on military prisons, and ensures observation of all regulations by persons concerned in military administration. In other countries in which there is a route of appeal via an ombudsman or similar civilian official (Finland and Germany, for example), there is a marked reluctance on the part of soldiers to take such a course, perhaps because of a general preference for avoiding civilian interference in military affairs. In the United States, there is a reluctance to take grievances to the inspector general of the army rather than going through the normal chain of command, mainly because of a fear of reprisal.

Wartime procedures. In time of war those countries that normally have an exclusively civilian jurisdiction over military offenders will usually introduce courts composed of soldiers. Civilian members of mixed tribunals will normally be replaced by soldiers. Elsewhere, military courts of less senior composition will have extended powers. In addition, the appeal procedure may be superseded or may be liable to be overridden by military commanders in the interests of the immediate imposition of discipline. This applies particularly in the execution of the death sentence.

Current developments. There is a tendency to allow the soldier off duty much greater freedom from discipline than in the past. The German code, for example, places great stress on the soldier's retention of his fundamental rights under the Basic Law, and, generally, the soldier's rights as a citizen are preserved as far as is compatible with the maintenance of discipline and efficiency. In democratic countries, soldiers everywhere retain the right to vote, if necessary by postal ballot, and to support the political party of their choice. Political activity within the service, however, cannot be reconciled with the maintenance of discipline, while the circulation of propaganda or organizing on behalf of political parties whose tenets support a hostile power or are subversive to military discipline must clearly amount to a disciplinary offense as well as a threat to security. The military commander must retain sufficient control over the private lives of his subordinates to ensure that discipline and good order do not suffer. Any orders reasonably calculated to maintain this object are lawful, even when they curtail the private actions of the soldier. Hence, however tolerant the civil legislation of a country might be, the use of drugs, which impair the soldier's efficiency, or the formation of homosexual relationships, which subvert discipline, must be suppressed in any fighting force, just as in all armed forces drunkenness is a specific military offense.

The question of how far a soldier may avail himself of superior orders as a defense to a charge of unlawful action is a topic that has been and continues to be of the greatest importance in relation to war crimes (see WAR,

LAWS OF). The difficulties presented by such cases lend support to the view, at present more ideal than practical, that all such offenses should be tried by some neutral international tribunal. It is perhaps only by the enforcement of such a rule that the soldier who refuses to obey an unlawful order can be protected.

BIBLIOGRAPHY. J.W. BISHOP, "Civilian Judges and Military Justice: Collateral Review of Court-Martial Convictions," *Columbia Law Review*, 61:40–71 (1961), and "Court-Martial Jurisdiction over Military–Civilian Hybrids: Retired Regulars, Reservists and Discharged Prisoners," *University of Pennsylvania Law Review*, 112:317–377 (1964); CLARENCE E. BRAND, *Roman Military Law* (1968), a short readable account of Roman military discipline and justice under the Republic and the Empire; *Manual of Military Law* (HMSO 1969), the official regulations governing British courts-martial and a history of English military justice; D. KARLEN, "Court-Martial Appeals in England," *Military Law Review*, 27:65–81 (1963); *Military Jurisprudence: Cases and Materials* (1951), published by the Lawyers Co-operative Publishing Company, includes the major decisions of the Supreme Court and other American courts up to its publication date; D.P. O'CONNELL, "The Nature of British Military Law," *Military Law Review*, 27:141–155 (1963); G.C. RYKER, "The New French Code of Military Justice," *Military Law Review*, 44:71–96 (1969); UNITED STATES DEPARTMENT OF DEFENSE, *Manual for Courts-Martial*, rev. ed. (1969), the official regulations governing the jurisdiction and procedure of American military tribunals (the Constitution and the Uniform Code of Military Justice are contained in appendixes); F.B. WIENER, "Courts-Martial and the Bill of Rights: The Original Practice," *Harvard Law Review*, 72:1–49, 266–304 (1958), an exhaustive historical survey; G.D. HENDERSON, "Courts-Martial and the Constitution: The Original Understanding," *Harvard Law Review*, 71:293–324 (1957); WILLIAM W. WINTHROP, *Military Law and Precedents*, 2nd ed., 2 vol. (1896, reprinted 1920), the classic treatise on American military jurisprudence; *Proceedings of the International Society for Military Law and the Law of War* (every 3 years beginning 1960); QAMAR HASNAIN, *A Guide to Indian Military Law* (1962); H.P. SECHER, "Controlling the New German Military Elite," *Proc. Am. Phil. Soc.*, 109:63–84 (1965).

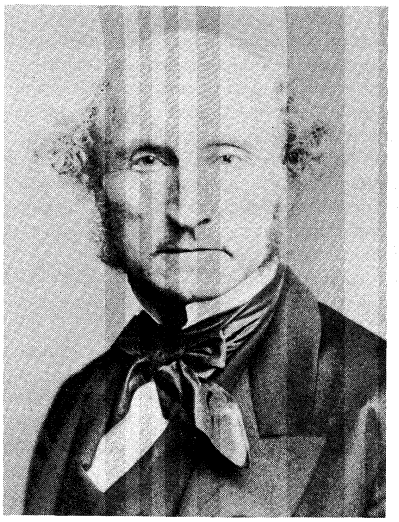
(W.E.S.)

Mill, John Stuart

John Stuart Mill was a British philosopher and economist, prominent as a publicist in the reforming age of the 19th century and of lasting interest as a logician and an ethical theorist.

Early life and career. The eldest son of the British historian, economist, and philosopher James Mill, he was born on May 20, 1806, in his father's house in Pentonville, London. He was educated exclusively by his father, who was a strict disciplinarian. By his eighth year he had read in the original Greek Aesop's *Fables,* Xenophon's *Anabasis,* and the whole of the historian Herodotus. He was acquainted with the satirist Lucian, the historian of philosophy Diogenes Laërtius, the Athenian writer and educational theorist Isocrates, and six dialogues of Plato. He had also read a great deal of history in English. At the age of eight he started Latin, the geometry of Euclid, and algebra and began to teach the younger children of the family. His main reading was still history, but he went through all the Latin and Greek authors commonly read in the schools and universities and, by the age of ten could read Plato and the Athenian statesman Demosthenes with ease. About the age of 12, he began a thorough study of Scholastic logic, at the same time reading Aristotle's logical treatises in the original. In the following year he was introduced to political economy and studied work of the Scottish political economist and philosopher Adam Smith and that of the English economist David Ricardo.

By courtesy of the Gernsheim Collection, the University of Texas at Austin



Mill.

While the training the younger Mill received has aroused amazement and criticism, its most important aspect was the close association it fostered with the strenuous character and vigorous intellect of his father. From his earliest days he spent much time in his father's study and habitually accompanied him on his walks. He thus inevitably acquired many of his father's speculative opinions and his father's way of defending them. But he did not receive the impress passively and mechanically. The duty of collecting and weighing evidence for himself was at every turn impressed upon the boy. His childhood was not unhappy, but there is no doubt that it was a strain on his constitution and that he suffered from the lack of natural, unforced development.

Effects of his education

From May 1820 till July 1821, Mill was in France with the family of Sir Samuel Bentham, brother of Jeremy Bentham, the English Utilitarian philosopher, economist, and theoretical jurist. Copious extracts from a diary kept at this time show how methodically he read and wrote, studied chemistry and botany, tackled advanced mathematical problems, and made notes on the scenery and the people and customs of the country. He also gained a thorough acquaintance with the French language. On his return in 1821 he added to his work the study of psychology and of Roman law, which he read with John Austin, his father having half decided on the bar as the best profession open to him. This intention, however, was abandoned, and in 1823, when he had just completed his 17th year, he entered the examiner's office of the India House. After a short probation he was promoted in 1828 to assistant examiner. For 20 years, from 1836 (when his father died) to 1856, Mill had charge of the British East India Company's relations with the Indian states, and in 1856 he became chief of the examiner's office.

In 1822 Mill had read P.E.L. Dumont's exposition of Bentham's doctrines in the *Traités de législation,* which made a lasting impression upon him. The impression was confirmed by the study of the English psychologists and also of two 18th-century French philosophers—Étienne Bonnot de Condillac, who was also a psychologist, and Claude Adrien Helvétius, who was noted for his emphasis on physical sensations. Soon after, in 1822–23, Mill established among a few friends the Utilitarian Society, taking the word, as he tells us, from *Annals of the Parish,* a novel of Scottish country life by John Galt.

Two newspapers welcomed his contributions—*The Traveller,* edited by a friend of Bentham's, and *The Morning Chronicle,* edited by his father's friend John Black. One of his first efforts was a solid argument for freedom of discussion in a series of letters to the *Chronicle* on the prosecution of Richard Carlile, a 19th-century English radical and freethinker. Mill seized every chance for exposing departures from sound principle in Parliament and courts of justice. Another outlet was opened up for him (April 1824) with the founding of the *Westminster Review*, which was the organ of the philosophical radicals. In 1825 he began work on an edition of Bentham's *Rationale of Judicial Evidence* (5 vol., 1827). He took part eagerly in discussions with the many men of distinction who came to his father's house and engaged in set discussions at a reading society formed at the home of English historian George Grote in 1825 and in debates at the London Debating Society, formed in the same year.

Public life and writing. The *Autobiography* tells how in 1826 Mill's enthusiasm was checked by a misgiving as to the value of the ends that he had set before him. At the London Debating Society, where he first measured his strength in public conflict, he found himself looked upon with curiosity as a precocious phenomenon, a "made man," an intellectual machine set to grind certain tunes. The elder Mill, like Plato, would have put poets under ban as enemies of truth; he subordinated private to public affections; and Landor's maxims of "few acquaintances, fewer friends, no familiarities" had his cordial approval.

Beginning of his mature thought

The younger Mill now felt himself forced to abandon these doctrines. Too much in awe of his father to make him a confidant, he wrestled with his doubts in gloomy solitude. He emerged from the struggle with a more catholic view of human happiness, a delight in poetry for its own sake, a more placable attitude in controversy, a hatred of sectarianism, and an ambition no less noble and disinterested but moderated to practical possibilities. Gradually, the debates in the Debating Society attracted men with whom contact was invigorating and inspiring. Mill ceased to attend the society in 1829, but he carried away from it the conviction that a true system of political philosophy was

> something much more complex and many-sided than he had previously had any idea of, and that its office was to supply, not a set of model institutions but principles from which the institutions suitable to any given circumstances might be deduced.

Mill's letters in *The Examiner* in the autumn of 1830, after a visit to Paris, where he made the acquaintance of the younger liberals, may be taken as marking his return to hopeful activity; and a series of articles on "The Spirit of the Age" appeared in the same paper in 1831. During the years 1832 and 1833 he contributed many essays to *Tait's Magazine, The Jurist,* and *The Monthly Repository*. In 1835 Sir William Molesworth founded *The London Review*, with Mill as editor. It was amalgamated with *The Westminster* (as *The London and Westminster Review*) in 1836, and Mill continued as editor (latterly as proprietor, also) till 1840. In and after 1840 he published several important articles in *The Edinburgh Review*. Some of the essays written for these journals were reprinted in the first two volumes (1859) of Mill's *Dissertations and Discussions* and give evidence of the increasing width of his interests. Among the more important are "Thoughts on Poetry and Its Varieties" (1833), "Writings of Alfred de Vigny" (1838), "Bentham" (1838), "Coleridge" (1840), "M. De Tocqueville on Democracy in America" (1840), "Michelet's History of France" (1844), and "Guizot's Essays and Lectures on History" (1845). The twin essays on Bentham and Coleridge show Mill's powers at their splendid best and indicate very clearly the new spirit that he tried to breathe into English radicalism.

Works on logic and political economy

During these years Mill also wrote his great systematic works on logic and on political economy. His reawakened enthusiasm for humanity had taken shape as an aspiration to supply an unimpeachable method of proof for conclusions in moral and social science; the French positivist philosopher Auguste Comte had some influence here, but the main inspiration undoubtedly came from the English scientist and mathematician Sir Isaac Newton, whose physics had already been accepted as a model of scientific exposition by such earlier British philosophers as John Locke, David Hume, Jeremy Bentham, and James Mill. But he was determined that the new logic should not simply oppose the old logic. In his *Westminster* review (of 1828) of Richard Whately's *Elements of Logic*, he was already defending the syllogism against the Scottish philosophers who had talked of superseding it by a supposed system of inductive logic. He required his inductive logic to "supplement and not supersede." For several years he searched in vain for the means of concatenation. Finally, in 1837, on reading William Whewell's *Philosophy of the Inductive Sciences* and rereading John F.W. Herschel's *Preliminary Discourse on the Study of Natural Philosophy*, Mill at last saw his way clear both to formulating the methods of scientific investigation and to joining the new logic onto the old as a supplement. *A System of Logic*, in two volumes, was published in 1843 (3rd–8th editions, introducing many changes, 1851–72). Book VI comprises his valiant attempt to formulate a logic of the human sciences—including history, psychology, and sociology—based on causal explanation conceived in Humean terms, a formulation that has lately come in for radical criticism.

Mill distinguished three stages in his development as a political economist. In 1844 he published the *Essays on Some Unsettled Questions of Political Economy*, which he had written several years earlier, and four out of five of these essays are solutions of perplexing technical problems—the distribution of the gains of international commerce, the influence of consumption on production, the definition of productive and unproductive labour, and the precise relations between profits and wages. Here for the most part Mill appears as the disciple of David Ricardo, striving after more precise statements and reaching forward to further consequences. In his second stage, originality and independence become more conspicuous as he struggles toward the standpoint from which he wrote his *Principles of Political Economy*. This was published in 1848 (2 vol.; 2nd and 3rd eds., with significant differences, 1849, 1852), and, at about the same time, Mill was advocating the creation of peasant proprietorships as a remedy for the distresses and disorder in Ireland. Thereafter, he made a more thorough study of Socialist writers. He was convinced that the social question was as important as the political question. He declined to accept property, devised originally to secure peace in a primitive society, as necessarily sacred in its existing developments in a quite different stage of society. He separated questions of production and distribution and could not rest satisfied with the distribution that condemned the labouring classes to a cramped and wretched existence, in many cases to starvation. He did not come to a Socialist solution, but he had the great merit of having considered afresh the foundations of society. This he called his third stage as a political economist, and he says that he was helped toward it by Mrs. Taylor (Harriet Hardy), who became his wife in 1851.

Activities during his married life

It is generally supposed that Mill writes with a lover's extravagance about Harriet's powers. He expressly says, indeed, that he owed none of his technical doctrine to her, that she influenced only his ideals of life for the individual and for society, and that the only work directly inspired by her is the essay on the "Enfranchisement of Women" (*Dissertations*, vol. 2). Nevertheless, Mill's relations with her have always been something of a puzzle.

During the seven years of his marriage Mill became increasingly absorbed in the work of the British East India Company and in consequence published less than at any other period of his life. In 1856 he became head of the examiner's office in the India House, and for two years, till the dissolution of the company in 1858, his official work kept him fully occupied. It fell to him as head of the office to write the defense of the company's government of India when the transfer of its powers was proposed. Mill opposed the transfer, and the documents in which he defended the company's administration are models of trenchant and dignified pleading. On the dissolution of the company, Mill was offered a seat in the new council but declined it and retired with a pension of £1,500. His retirement from official life was followed almost immediately by his wife's death at Avignon, France. He spent most of the rest of his life at a villa at Saint-Véran, near Avignon, returning to his house at Blackheath only for a short period in each year.

The later years. Mill sought relief by publishing a series of books on ethics and politics that he had meditated upon and partly written in collaboration with his wife. The essay *On Liberty* appeared in 1859 with a touching dedication to her and the *Thoughts on Parliamentary Reform* in the same year. In his *Considerations on Representative Government* (1861) he systematized opinions already put forward in many casual articles and essays. It has been remarked how Mill combined enthusiasm for democratic government with pessimism as to what democracy was likely to do; practically every discussion in

these books exemplifies this. His *Utilitarianism* (in *Fraser's Magazine*, 1861; separate publication, 1863) was a closely reasoned attempt to answer objections to his ethical theory and to remove misconceptions about it. He was especially anxious to make it clear that he included in "utility" the pleasures of the imagination and the gratification of the higher emotions; and to make a place in his system for settled rules of conduct.

Mill also began to write again on the wider philosophical questions that had occupied him in the *Logic*. In 1865 he published both his *Examination of Sir William Hamilton's Philosophy* and his *Auguste Comte and Positivism*, but in both writings his motives were largely political. It was because he regarded the writings and sayings of Sir William Hamilton as the great fortress of intuitional philosophy in Great Britain that Mill undertook to counter his pretensions. In dealing with Comte, Mill distinguished sharply between Comte's earlier philosophical doctrine of Positivism and his later religion of humanity. The doctrine he commended (as he had frequently done previously) because he regarded it as a natural development of the outlook of George Berkeley and Hume; the religion he attacked because he saw in it merely another attempt to foist a priestly hierarchy upon suffering humanity. It is noticeable that Mill's language in these books is much closer to the language of Bentham and James Mill than it had been since his boyhood, and it was as an act of piety that in 1869 he republished his father's *Analysis of the Phenomena of the Human Mind* with additional illustrations and explanatory notes.

Mill in Parliament

While engaged in these years mainly with theoretical studies, Mill did not remit his interest in current politics. He supported the North in the U.S. Civil War, using all his strength to explain that the real issue at stake in the struggle was the abolition of slavery. In 1865 he stood as parliamentary candidate for Westminster, on conditions strictly in accordance with his principles. He would not canvass or pay agents to canvass for him, nor would he engage to attend to the local business of the constituency. He was with difficulty persuaded even to address a meeting of the electors but was elected. He took an active part in the debates preceding the passage of the 1867 Reform Bill, and helped to extort from the government several useful modifications of the bill, for the prevention of corrupt practices. The reform of land tenure in Ireland (see his *England and Ireland*, 1868, and his *Chapters and Speeches on the Irish Land Question*, 1870), the representation of women (see below), the reduction of the national debt, the reform of London government, and the abrogation of the Declaration of Paris (1856)—concerning the carriage of property at sea during the Crimean War—were among the topics on which he spoke. He took occasion more than once to enforce what he had often advocated, England's duty to intervene in foreign politics in support of freedom. As a speaker Mill was somewhat hesitating, but he showed great readiness in extemporaneous debate. Elected rector of St. Andrews University, he published his "Inaugural Address" in 1867.

Mill's subscription to the election expenses of the freethinker and radical politician Charles Bradlaugh and his attack on the conduct of Gov. E.J. Eyre in Jamaica were perhaps the main causes of his defeat in the general parliamentary election of 1868. But his studied advocacy of unfamiliar projects of reform had made him unpopular with "moderate Liberals." He retired with a sense of relief to Avignon. His villa was filled with books and newspapers; the country round it furnished him with a variety of walks; he read, wrote, discussed, walked, botanized. He was extremely fond of music and was himself a fair pianist. His stepdaughter, Helen Taylor (died January 1907), was his constant companion after his wife's death. Mill was an enthusiastic botanist all his life and a frequent contributor of notes and short papers to the *Phytologist*. During his last journey to Avignon he was looking forward to seeing the spring flowers and completing a flora of the locality.

Mill did not relax his laborious habits or his ardent outlook on human affairs. The essays in the fourth volume of his *Dissertations* (1875; vol. 3 had appeared in 1867)—on endowments, on land, on labour, and on metaphysical and psychological questions—were written for the *Fortnightly Review* at intervals after his short parliamentary career. In 1867 he had been one of the founders, with Mrs. P.A. Taylor, Emily Davies, and others, of the first women's suffrage society, which developed into the National Union of Women's Suffrage Societies, and in 1869 he published the *Subjection of Women* (written 1861), the classical theoretical statement of the case for women's suffrage. His last public activity was concerned with the starting of the Land Tenure Reform Association, for which he wrote in *The Examiner* and made a public speech a few months before his death; the interception by the state of the unearned increment on land and the promotion of cooperative agriculture were the most striking features in his program, which he regarded as a timely compromise in view of the impending struggle between capital and labour in Europe. His *Autobiography* and *Three Essays on Religion* (1874) were published posthumously.

Mill died at Avignon on May 8, 1873. A bronze statue of him stands on the Thames embankment in London, and G.F. Watts's copy of his original portrait of Mill hangs in the National Gallery there.

Influence and significance. Mill was a man of extreme simplicity in his mode of life. The influence that his works exercised upon contemporary English thought can scarcely be overestimated, nor can there be any doubt about the value of the liberal and inquiring spirit with which he handled the great questions of his time. Beyond that, however, there has been considerable difference of opinion about the enduring merits of his philosophy. At first sight he is the most lucid of philosophers. Many people have spoken of the marvelous intelligibility of his writing. Usually, however, it is not long before doubts begin to creep in. Although the lucidity remains, its span is seen to be somewhat limited, and one sometimes has the uneasy feeling that he is being equally lucid on both sides of a question.

Evaluation of Mill

Oddly enough, however, this judgment has not led to any neglect of Mill. Little attention is now paid to Hamilton or to Whewell, but Mill's name continually crops up in philosophical discussions. This is partly due to the fact that Mill offers a body of doctrine and a set of technical terms on many subjects (notably on induction) that have proved extremely useful in the classroom. But a more important reason is that he has come to be regarded as a sort of personification of certain tendencies in philosophy that it is regarded as continually necessary to expound or expose because they make such a powerful appeal to serious minds. Thus he is or says he is a Utilitarian; yet nothing, it is pointed out, could tell more strongly against Utilitarianism than certain passages in his writings. Then again, he is said to be an Empiricist (although he says himself that he is not), and his theories of the syllogism and of mathematics are constantly used to demonstrate the fatal consequences of this way of thinking.

It is misleading to speak without qualification of Mill's Utilitarianism. Nor is it sufficient to add that Mill modified the Utilitarianism that he inherited from Bentham and from his father in one way and another in order to meet the criticisms that it encountered in Victorian times. He does, it is true, sometimes give that impression (as in his essay *Utilitarianism*); but elsewhere (as in his essay *On Liberty*) he scarcely attempts to conceal the fact that his premises are completely independent of Bentham's. Thus, contrary to the common belief, it appears to be very hazardous to characterize offhand the precise position of Mill on any major philosophical topic. He sometimes behaved with a reckless disregard of consequences more suitable to a Romantic than to a Utilitarian. He is thoroughly romantic, again, and thoroughly representative of his age in the eagerness with which he seeks out and endeavours to assimilate every last exotic line of thought which shows any signs of vitality. He himself claimed to be superior to most of his contemporaries in "ability and willingness to learn from everybody," and indeed, for all his father's careful schooling, there was never anybody less buttoned up against alien influences

than Mill. In his writings there can be discerned traces of every wind of doctrine of the early 19th century.

MAJOR WORKS

POLITICS AND ECONOMICS: *Essays on Some Unsettled Questions in Political Economy* (1844); *Principles of Political Economy*, 2 vol. (1848; 2nd and 3rd eds. with important differences, 1849, 1852); *On Liberty* (1859); *Considerations on Representative Government* (1861); *Utilitarianism* (1863); *On The Subjection of Women* (1869).

PHILOSOPHY AND RELIGION: *A System of Logic* (1843); *Examination of Sir William Hamilton's Philosophy* (1865); *Auguste Comte and Positivism* (1865); *Three Essays on Religion* (1874).

OTHER WORKS: Essays on "Bentham" (1838) and "Coleridge" (1840) in *Dissertations and Discussions*, 4 vol. (1859–75), also reprinted together with an introduction by F.R. LEAVIS (1950); *Autobiography*, ed. by Helen Taylor (1873).

BIBLIOGRAPHY. Mill's *Collected Works* are currently being published in a definitive edition of 22 vol., ed. by F.E.L. PRIESTLEY and J.M. ROBSON (1963–). The following have appeared: *The Earlier Letters 1812–1848* (vol. 12–13); *Principles of Political Economy* (vol. 2–3, 1965); and *Essays on Ethics, Religion and Society* (vol. 10, 1969). Each volume has full introductions, notes, and indexes. At present, for Mill's later letters it is necessary to refer to *The Letters of John Stuart Mill*, ed. by HUGH ELLIOT, 2 vol. (1910).

Biography: The standard biography is M. ST. J. PACKE, *The Life of John Stuart Mill* (1954), a fascinating and readable book (with bibliography), interpretations sometimes too colourful. Mill's own *Autobiography* has been edited several times, recently by JACK STILLINGER (1969). Stillinger has also edited *The Early Draft of John Stuart Mill's Autobiography* (1961), written in 1853–54, which includes a much more vivid and intimate account of Mill's relations with his father and mother; it suggests a more gloomy picture of his childhood than does the final version. (This is not a work that Mill was himself prepared to publish.) ALEXANDER BAIN, *John Stuart Mill: A Criticism with Personal Recollections* (1882), will never be superseded; an excellent, rambling account by Mill's closest philosophical disciple, it refers to many conversations and quotes from letters now lost. F.A. HAYEK (ed.), *John Stuart Mill and Harriet Taylor* (1951), is full of details about the writing of Mill's works from the 1830s to 1858 (with gaps for times when they were meeting together); the painting of Harriet Taylor should not be missed.

Comment and Criticism: R.P. ANSCHUTZ, *The Philosophy of J. S. Mill* (1953), a subtle and precise study presupposing a wide reading of the texts; K.W. BRITTON, *John Stuart Mill: Life and Philosophy*, 2nd ed. (1969), an introductory book covering Mill's life and work—sympathetic but critical; ELIE HALEVY, *La Formation du radicalisme philosophique*, 3 vol. (1901–04; Eng. trans., *The Growth of Philosophical Radicalism*, 1928; new ed., 1952), the standard comprehensive account of the school in which Mill was brought up, with issues fully analyzed and discussed (with bibliography); J. PLAMENATZ, *The English Utilitarians* (1949), a good account of the background and content of Mill's moral theory; J.M. ROBSON, *The Improvement of Mankind: The Social and Political Thought of John Stuart Mill* (1968), a readable book that offers an account of Mill's life and of his mature views on morals, scientific method, politics, and sociology; M. COWLING, *Mill and Liberalism* (1963), an eccentric account that imputes to Mill a strong strain of authoritarianism; A. RYAN, *John Stuart Mill* (1970), an attempt to show that a single constant theory of inductivism underlies all Mill's writings; PETER WINCH, *The Idea of a Social Science*, ch. 3 (1958), attacks attempts to construct a causal science of human conduct; F.H. BRADLEY, *Ethical Studies*, 2nd ed. (1967), essay 3 is an attack on Mill's hedonism; G.E. MOORE, *Principia Ethica* (1903; paperback edition, 1959), a very lively onslaught on Mill's Utilitarianism, often patently unfair; J.O. URMSON, "The Interpretation of the Moral Philosophy of J.S. Mill," *Philosophical Quarterly*, 3:33–39 (1953), on moral rules; B. MACMINN, J.R. HAINDS, and J. MCCRIMMON (eds.), *Bibliography of the Published Writings of John Stuart Mill* (1945, reprinted 1970).

(R.P.An.)

Millennialism

Millennialism (from the Latin word for 1,000) is a philosophy of history viewed from a Christian theological standpoint and a religious movement now associated with such modern Protestant sects as the Adventists, Jehovah's Witnesses, and certain segments of many Protestant denominations. There have been many millennial groups and individuals throughout church history. The term is derived from the imagery of the New Testament Book of Revelation (Rev. 20), in which the writer describes a vision of Satan being bound and thrown into a bottomless pit and of Christian martyrs being raised from the dead and reigning with Christ for 1,000 years. This 1,000-year period, known as the millennium, is viewed as a time during which man's yearnings for peace, freedom from evil, and the rule of righteousness upon earth are finally realized through the power of God.

NATURE AND SIGNIFICANCE

As a branch of eschatology—that area of theology dealing with last things, such as death, immortality, the Last Judgment, the end of the world, the rewards of the just, and the punishment of the damned—millennialism is concerned with the earthly prospects of the human community. Not limiting itself to the prospects of the individual in this world and the next, millennialism attempts to answer in vivid imagery such questions as: What will be the final end of this world? Will mankind ever fulfill the agelong dream of dwelling in an earthly paradise or will all men be destroyed in a cataclysm of fire brought on by their own folly or God's judgment?

Millennialism is thus the cosmology (study of order) of eschatology, its chronology one of future events, comparable to the historical record of the past.

Millennialism is found within both Christian and other traditions. During the 20th century, anthropologists, historians, and sociologists explored the millennialist aspects of non-Western cultures, finding many striking similarities to the millennialism within the Judeo-Christian tradition (see also TRIBAL RELIGIOUS MOVEMENTS, NEW; MESSIAH AND MESSIANIC MOVEMENTS). The millennial treatises produced by Jewish and Christian believers in the latter part of the Greco-Roman civilization—the Hellenistic period—particularly the books of Daniel and Revelation, provided the building materials from which the successive millennial structures were erected. In constant repetition the motifs, the leading characters, the symbols, and the chronologies of these works have arisen in the teaching of some prophet of the end of the world, each time taking on new significance from associations with contemporaneous events.

Basis in Daniel and Revelation

HISTORICAL AND INTELLECTUAL DEVELOPMENT

The four main periods in the Judeo-Christian millennial tradition are (1) the rise and development of apocalyptic millennial views from about 200 BC until AD 200; (2) the creation of a countervailing allegorical interpretation of the millennium under the Church Fathers Origen and Augustine and the dominance of that theory in Christendom until the 17th century; (3) the re-emergence of a belief in the literal fulfillment of millennial prophecies and the association of those prophecies with a progressive and optimistic philosophy of history from about 1650 to the 20th century; and (4) the recent emergence of sectarian groups that have given new expression to the earliest tradition of apocalyptic and pessimistic millennialism.

Apocalyptic millennialism. Millennialism is dependent on the concept of apocalypticism—the view that God's final intervention in history will be accompanied by dramatic, unusual events.

The Book of Daniel. Western apocalypticism is based upon the archetypal apocalyptic work in the Judeo-Christian tradition, the Book of Daniel. Daniel is the only apocalyptic book to be admitted to the Old Testament canon, just as the Book of Revelation is the only apocalypse included in the canon of the New Testament. There are many noncanonical apocalyptic works from both Jewish and Christian authors, among them the three books of *Enoch*, the *Second Book of Esdras*, the *Ascension of Isaiah*, and the *Apocalypse of Peter*. All of the apocalyptic works written during the first efflorescence of millennialism, including the Book of Revelation, owe much of their shape and style to Daniel. This Old Testament book stands in the succession of the Jewish prophets and was apparently influenced by Iranian religious thought, such as the Zoroastrian concepts of the Last Judgment, the battle between good and evil involving both men and angels, and a punish-

ment of fire for evildoers (see also ZOROASTRIANISM AND PARSIISM).

In the face of a threat to the existence of the Jewish faith and temple worship, a group of zealots revolted in 168 BC against the occupation forces of the Seleucid monarch of Syria, Antiochus IV Epiphanes (*c.* 215–163/164 BC) and against those of their Jewish countrymen who favoured reducing Judaism to the level of a hellenized Oriental (Greek and Syrian) cult. The author of Daniel constructed his work to give aid and comfort to the rebel cause, particularly to assure them that God was aiding them, that the end of their struggles was in sight, and that a new golden age was dawning. In a vision reputed to be seen by King Nebuchadrezzar, he depicted a series of four world monarchies, represented in one passage by parts of a giant statue and in another by mythological beasts, each empire embodying evil to a greater extent than the last. Man's empires will end with the fourth kingdom, which is crushed by a "stone . . . cut out by no human hand,"symbolizing the fact that neither its destruction nor the ensuing order are natural developments from forces latent in history. A figure called the Son of man, however, will institute a fifth, entirely righteous, just, and eternal kingdom.

Visions of the ages of the world

As in the Jewish prophetic tradition that preceded it, Daniel made predictions about the future, but, unlike the predictions of a prophet such as Jeremiah (late 7th to early 6th century BC), the outcome anticipated by the prophet was not the virtually inevitable product of antecedent forces but a total reversal of what might seem to be the likely outcome if God were not to intervene. The reversal of worldly expectations through a violent supernatural intervention in the course of history is one of the most characteristic features of apocalypticism and stands quite in contrast to the older prophetic style. Also essential to Daniel and subsequent apocalypticism is the immediacy of the message and the imminence of the deliverance that is promised—the promise of salvation now. Descriptions of this imminent salvation of cosmic proportions included vivid representations of historical figures who depicted the growth of evil and decline of goodness from past time down to the present, when all wickedness came into terrifying focus.

To these significant emphases in Daniel's apocalypse—the imminent and supernatural intervention of God in man's history and the reversal of the heretofore irresistible progress of evil and declension of good—might be added other characteristics that have proved to be influential. Numerology, mythological figures, and angelology, which have continued to play such a large part in millennial movements, were probably introduced as a result of the influence of Iranian thought. Other characteristics of Daniel, such as its pseudonymous authorship and the emphasis upon the esoteric, mysterious quality of the truths discussed, were probably due to the unique problems faced by the author in presenting these views to a 2nd-century-BC audience.

In the centuries immediately following the writing of Daniel, the apocalyptic world view influenced Jewish culture; the audiences whom Jesus addressed were acquainted with it; and the early Christian Church embraced the apocalyptic world view. The Apostle Paul frequently expressed apocalyptic expectations (I Thess. 4), and Mark 13, a passage often called "the little apocalypse," reflects the apocalyptic expectations of the Roman church at about AD 70.

The Book of Revelation. The Christian Church in the 1st century wrestled with a difficult problem. Jesus had promised the inauguration of a new age, the Kingdom of God, and yet life proceeded after his death in much the same way that it had before his birth, with the exception that Christian believers suffered severe persecution for their faith. The primitive church solved this problem through the paradox of the Second Coming of Christ: Christ has come and is coming again. They believed that the new age had dawned but would not be fully revealed until Christ's Second Coming in glory.

Like the Book of Daniel, the Revelation to John or Revelation was composed during a period of persecution. It was probably written during the last decade of the 1st century AD, and it reflects the persecutions beginning under the emperor Nero (AD 37–68), who seems to be portrayed as the Antichrist—the beast whose symbolic number is 666 (Rev. 13). After addressing letters to churches of Asia Minor, the author depicted his vision of a series of judgments—seven seals opened, seven trumpets blown, seven bowls poured out. The writer directed his attack against the Roman Empire, referred to cryptically as Babylon and as the great harlot. Christ was described as the executor of God's judgment, appearing not as the man Jesus but as an omnipotent king riding upon a white horse with eyes like a flame of fire and a mouth like a sharp sword "with which to smite the nations" (Rev. 19). In the Book of Revelation the assimilation of Jewish apocalypticism to Christianity was completed. Daniel's Son of man was replaced by Christ; many of the numerological formulas were repeated; and the dualistic universe of good and evil, Christ and Antichrist, was provided with a new and unforgettable set of characters. The essence of the apocalypse in Revelation remained what it had been in Daniel: the immediate, direct aid of God was to be momentarily expected, accomplishing the dramatic reversal of history that the believers' present desperate state demanded.

Apocalyptic symbolism

Apocalypticism in the early church. During the first hundred years of Christian history, this form of millenarianism, or chiliasm (from the Greek word for 1,000), was commonly taught and accepted within the church. Persecution of the church was intermittent, however, and apocalyptic zeal flagged without the pressure of opposition. Christian missionaries succeeded in converting large numbers of Roman citizens, and some of the antagonism toward the empire was dissipated. The appeal of millenarian thought was further limited by its association with the heresy of Montanism. In characteristic apocalyptic fashion, Montanus, the founder of the movement, was fascinated with the idea of dividing past and future into units of prophetic calculation. In AD 156, according to the 4th-century Christian antiheretical writer Epiphanius, Montanus declared himself the prophet of a third testament, a new age of the Holy Spirit. Phrygia in Asia Minor became the centre of this ecstatic and ascetic movement whose leaders claimed divine inspiration for their visions and utterances, the main theme of which was the imminent Second Coming of Christ. This concept of the third age, the new day of the spirit of God, has been one of the most consistently repeated features of millenarian history, reappearing, for example, in Joachim of Fiore's philosophy of history during the 12th century, in views of the 17th-century Quakers, and in the apocalyptic speculations of the Seventh-day Adventists of the 19th and 20th centuries. When persecution of Christians was renewed late in the 2nd century, Montanism began to appeal outside Asia Minor and found converts throughout the Roman Empire, including Tertullian, a North African lawyer and theologian. The church survived this persecution, however, and Montanism was stigmatized as a heresy.

The age of the Holy Spirit

The influence of Greek thought upon Christian theology undermined the millenarian world view in another, possibly more significant, manner. In the theology of the great 3rd-century Alexandrian Christian thinker Origen, the focus was not upon the manifestation of the kingdom within this world but within the soul of the believer, a significant shift of interest away from the historical toward the metaphysical, or the spiritual. The association of apocalyptic millenarianism with the Montanist heresy, the growing influence of Greek thought upon Christian theology, and the conversion of Constantine the Great and adoption of Christianity as the favoured religion of the empire—all combined to discredit millenarianism for centuries.

Allegorical millennialism. *Augustinian views.* In the new age of the church triumphant—*i.e.*, when Christianity became the accepted religion of the Roman Empire—Augustine (354–430), bishop of Hippo, gave definitive expression to the view that was to dominate Western civilization through the Reformation. In his

The City of the World and the City of God

City of God, a philosophy of history Augustine viewed the world as eternally divided between the City of the World and the City of God. All men owe allegiance to one or the other of these cities and will ultimately share the fate of that community. The City of the World is ruled by Satan, the prince of this world. He and all who pay homage to his city will suffer eternal punishment. The City of God is represented in the church, and for the church God has ordained salvation from the persecution of the City of the World and eternal bliss in the courts of heaven. Much of this conception is reminiscent of the apocalyptic world view, but the views of Augustine also contained distinct differences.

The dualism represented in apocalypticism is reflected with equal intensity in Augustine's two cities. Furthermore, Augustine remained as pessimistic as any millenarian about the future of the City of the World and the prospect of progress in this world. After his conversion to Christianity, Augustine, a former *bon vivant*, consistently favoured a world-denying and ascetic style of life. In fact, his disillusionment with worldly values was more thorough than that of the millenarians, for he rejected as carnal any expectations of a renewed and purified world that the believers could expect to enjoy. In this respect he differed sharply with the apocalyptic tradition. The millenarian, in contrast to Augustine, had no quarrel with the world as such except that he had found it controlled by his enemies. The millenarian believed that when the imminently expected saviour had defeated these foes, the righteous would share in an earthly paradise, a land of physical, not spiritual, benefits.

The literalistic descriptions of the judgments that were predicted for the wicked and the bliss foretold for the righteous found in such apocalyptic works as the Book of Revelation were interpreted allegorically by Augustine. He expected that ultimately the history of this world would end, but for him the millennium had become a spiritual state into which the church collectively had entered at Pentecost—the time of the reception of the Holy Spirit by Christ's disciples soon after his Resurrection—and which the individual Christian might already enjoy through mystical communion with God. In contrast to the apocalypticist's focus upon the contemporary world, Augustine, though just as influenced by his own cultural milieu, responded with a millennial eschatology that seemed almost oblivious of time. As far as the struggle with evil in this world is concerned, Augustine surrendered and abandoned the field. No imminent supernatural intervention in history was expected, and no dramatic reversal of the tide of battle was anticipated. Augustine

Realized eschatology

taught what has been referred to as "realized" eschatology. For him the battle had already been fought on the spiritual ground that really mattered. God had triumphed. Satan has been reduced to lordship in this world. In the present age the City of the World and the City of God have been forced to coexist. Eventually, even that small patrimony that Satan claimed would be taken from him, and God would become triumphant.

Medieval and Reformation millennialism. Augustine's allegorical millennialism became the official doctrine of the church, and apocalypticism went underground. During the late Middle Ages and the Reformation, millenarian views were frequently voiced, most often by rebels and radicals. Apart from these dissidents, however, the doctrine of Augustine remained unchallenged until the 17th century. The Protestant Reformers of the Lutheran, Calvinist, and Anglican traditions were not apocalypticists but remained firmly attached to the views of Augustine, for whose theology they felt a particular affinity. Many of the allusions of the Book of Revelation were viewed in a distinctly Protestant perspective by the Reformers—the allusions to Rome as the great harlot and as Babylon being transferred to the Roman church, and the pope being identified as the beast. Each of the three main Protestant traditions of 16th-century Europe, however, found support from the secular authorities in Saxony, Switzerland, and England and remained in the same position vis-à-vis the state as had the medieval church. The apocalyptics within medieval Christendom, as well as in the 16th-century Reformation, were those who believed that their only help was the Lord and for whom persecution was a reality and destruction an imminent threat.

Progressive millennialism. The Augustinian millennial world view, though it survived the Reformation, did not survive the intellectual revolution of the 17th century. Behind the development of science lay a profound reorientation of Western thought that involved, in the first place, the rehabilitation of nature. A part of Augustine's rejection of the world stemmed from the frustration felt by his generation in attempting to cope with the natural and social history of its time. By 1600 Europeans had gained confidence in their own abilities. Such philosophers as Francis Bacon announced the dawn of a new day and attacked the Augustinian reluctance to see anything but the work of the devil in attempts to control or understand natural processes. Secondly, European intellectuals were becoming far more interested in measurement and quantification. Allegory fell into disrepute when the medieval interpretation of the nature of the heavenly bodies was proved to be erroneous by the facts discovered by the use of the telescope. A new concern with calculation and literalism spread to biblical scholarship and resulted in the creation of the third type of millennialism found in the Christian tradition—progressive millennialism.

Early progressive millennialism. Joseph Mead (Mede), a 17th-century Anglican biblical scholar, was the pioneer in the movement. Ignoring the allegorical interpretation long associated with the book, Mead took a fresh look at the text of Revelation. He concluded that the Scriptures held the promise of a literal Kingdom of God. The work

Millennialist optimism

of redemption, he concluded, would be completed within human history on the stage of this world. The Book of Revelation itself seemed to contain a historical record of the progress of that Kingdom, and scholars other than Mead soon were speculating where in the prophetic timetable the modern millennialist might locate himself. Thus far, progressive millennialism appeared to be identical with the apocalyptic millenarianism of the early church, but there the similarity ended. The Kingdom would not be brought into being through any dramatic reversal of the historical process, nor did the progressive millennialists believe that the Second Advent of Christ would occur in order to rescue them from destruction. History did not need reversing for these early Enlightenment Christians (those who emphasized reason). They thought of the record of the past as the story of victory over evil and the conquest of Satan. They rejected the fundamental assumptions of the apocalypticist—*i.e.*, that victory would be snatched from the jaws of defeat only by a miraculous deliverance. For them it seemed that the progress of history had been continuously upward, that the Kingdom of God was coming ever closer, and that it would arrive, not without struggle, but on the basis of the same kind of effort that had always triumphed in the past.

In the 18th century the teachings of the progressive millennialists became dominant in many Protestant churches. The Anglican polemicist and commentator Daniel Whitby (1638–1726), in his *Paraphrase and Commentary on the New Testament* (1703), provided such convincing support to the position that he has often been credited with creating it. In America interest in the millennium had not been lacking among Puritan scholars, but it was the great revivalist Jonathan Edwards (1703–58) who first adopted progressive millennialism, giving detailed exposition of his views in his uncompleted *History of the Work of Redemption*. Edwards saw significance for millennialism in the discovery and settlement of the New World, and he anticipated the establishment of Christ's kingdom sometime near the end of the 20th century.

Later progressive millennialism. The association of the millennium with the role of the United States proved to be a volatile 19th-century mixture in the hands of Protestant ministers, and for much of that period millennialism fed the fires of nationalism and Manifest Destiny. In a typical utterance, a leading Presbyterian minister of the 1840s, Samuel H. Cox, told an English audience that, "in America, the state of society is without parallel in

universal history. . . . I really believe that God has got America within anchorage, and that upon that arena, He intends to display his prodigies for the millennium." The late-19th-century movement known as the Social Gospel, dedicated as it was to establishing the Kingdom of God here and now, manifested most clearly the continuing influence of progressive millennialism.

Because the advocates of optimistic millennialism were confident of the ultimate triumph of their cause, it must not be assumed that they took evil lightly. They thought of God's Kingdom as advancing, as Jonathan Edwards argued, but not without destruction. Though they were not apocalyptic, their view of history included the cataclysmic. In the American Civil War, for example, the anti-slavery writer Julia Ward Howe, in "The Battle Hymn of the Republic," described God's truth as "marching on." In Pres. Woodrow Wilson's crusade to make the world "safe for democracy" by the entry of the United States into World War I (1914–18), one can see the same idea barely disguised. According to the progressive millennialists, Christ's Second Advent would occur at the close of the millennium as its crowning event, and, as a result, their position has frequently been called postmillennialism.

Summary of millennial positions

Thus, the three variations of Western millennial thought are: the apocalyptic millenarianism of the early church, especially during its periods of persecution; the allegorical millennialism of Augustine, which dominated the medieval church, was accepted by some of the leading figures of the Protestant Reformation, and continues to receive acceptance among some Roman Catholics; and the progressive millennialism characteristic of Protestantism since 1650 and particularly influential in the United States. These three forms of millennial thought exhaust the catalogue of millennial variations but do not complete the historical survey. The form of apocalyptic millenarianism that achieved acceptance during the formative years of the church has had a history of almost innumerable resurrections, particularly as the expression of the dissident and rebellious.

MILLENARIAN GROUPS

Millennialist pessimism

For apocalyptic millenarianism, as indicated above, no one who is satisfied with his existence, who is wealthy, powerful, or well integrated into his social system will find much appeal in a world view that is focussed upon crisis and pessimistic about the future if God does not bring immediate miraculous deliverance. To many of the desperate men of the later Middle Ages, however, beginning to feel alienation from the simpler pastoral life that was fading away in the face of commercial and urban development, this apocalyptic outlook seemed directly relevant. Beginning as early as the 11th century, especially in those regions most affected by social change, thousands were swept into the camp of self-appointed messiahs, such as Tanchelm, a fiery heretic of the Netherlands. In the 12th century, Joachim of Fiore gave fresh expression to the Montanist millenarian vision of the third age of the spirit. During the Hussite rebellion in Bohemia in the 15th century, one group, the Taborites, gathered massive support around a savage attack upon clerical interests and the proclamation of the imminent return of Christ. These outbursts continued during the Reformation. In the first half of the 16th century, Luther, for example, vigorously opposed enthusiastic prophets of the millennial age and the coming of Christ, such as the radical Reformer Thomas Müntzer. The most famous 16th-century millenarian outburst came at the city of Münster in Westphalia, where refugees from persecution managed to gain control of the city, ejected unbelievers, and, while the community was besieged, attempted to establish a prototype of the New Jerusalem. Again in the 17th century, during the English Civil War, a few apocalyptic Puritans, enthused by the prospect of the millennium, attempted to set up the Fifth Monarchy prophesied in the Book of Daniel.

Other than groups such as the celibate Shakers in England and the United States, the 18th century was relatively free of apocalypticism until the time of the French Revolution; but that great cataclysm and the industrialization of the West encouraged a new wave of millenarian sects, among them the Catholic Apostolic Church, the Seventh-day Adventists, Jehovah's Witnesses, the Latter-day Saints (Mormons), and those premillennialists who contributed directly to the rise of the biblically oriented and conservative movement known as Fundamentalism. The appeal of these groups in the United States increased during times of national crisis or peril—the 1830s, 1870s, and the years surrounding the world wars of the 20th century, which were particularly marked by the prominence and success of apocalyptic appeals. Unlike many of the earlier reawakenings of millenarianism, however, these modern American apocalyptic groups possessed no revolutionary potential. They rather seemed to help their members to sublimate their frustrations in study or proselytizing.

CONCLUSION

Without the influence of the Bible, no connected millennial tradition would exist. The incorporation of Daniel and Revelation in the canon of the Old and New Testaments conferred a timelessness upon books apparently intended only as tracts for their own times. If the authors of Daniel and Revelation had been correct, judgment would have fallen upon the world in their own day. Their literature has survived, however, and become the basis for a whole tradition of expectations. As long as Christians accept the texts of the Bible as the Word of God, the tradition will endure.

BIBLIOGRAPHY. H.H. ROWLEY, *The Relevance of Apocalyptic* (1944), an introduction to the biblical books of Daniel, Revelation, and other noncanonical apocalyptic writings; J.A.T. ROBINSON, *Jesus and His Coming* (1957), a work helpful in understanding the eschatology of Jesus and the authors of the Gospels; N.R.C. COHN, *The Pursuit of the Millennium*, 2nd ed. (1961; rev. paperback ed., 1970), a discussion of millenarian movements during the Middle Ages; E.L. TUVESON, *Redeemer Nation: The Idea of America's Millennial Role* (1968), an analysis of the progressive millennialism of modern times; L.E. FROOM, *Prophetic Faith of Our Fathers*, 4 vol. (1946–54), a history of the development of prophetic interpretation, written to support Seventh-day Adventist theology.

(E.R.S.)

Miltiades

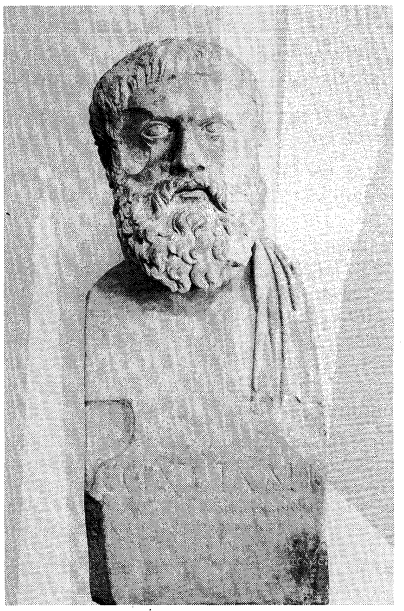
Miltiades, famous for his leadership in the Greek victory over the Persians at the Battle of Marathon, was born *c.* 554 BC in Athens. His family must have been extraordinarily wealthy; his father, Cimon, three times won the chariot races at the Olympic games, while his uncle, after whom he was named, was the founder of an Athenian semi-independent principality in the Chersonese. Miltiades the Elder died childless and arranged for his stepbrother's sons to inherit the dominions he had conquered. About 516, Miltiades left for the Chersonese (modern Gallipoli Peninsula), where he strengthened his authority by arresting his potential rivals and by surrounding himself, as tyrants were wont to do, with a heavily armed bodyguard of 500 men. He also married Hegesipyle, the daughter of a Thracian prince. Soon after, nonetheless, his authority was severely limited when Darius I of Persia expanded his power into Europe and reduced Miltiades to the rank of a Persian vassal. He led a contingent in the Scythian expedition of Darius (*c.* 513), and, according to Herodotus, he advised the destruction of the Danube bridge, which would have cut off Darius' retreat. Since Miltiades was not at once expelled by the Persians, the story is doubtful. It is a fact, on the other hand, that when the Ionian revolt against the Persians broke out in 499 BC, Miltiades decided to join the insurgents and to enter into friendly relations with the newly established Athenian democracy. Probably during the revolt Miltiades seized the isles of Lemnos and Imbros, which he eventually ceded to Athens. But the revolt collapsed by 494, and the year after, when Darius' fleet appeared off the Chersonese, Miltiades loaded five boats with his treasures and made for Athens. One of the boats, captained by Miltiades' eldest son, Metiochos, was captured. Metiochos was taken as a lifelong prisoner to

Relations with Darius

Persia, but Darius treated him honourably, married him to a Persian princess, and regarded their children as members of the Persian nobility.

Because of his fabulous wealth, his foreign wife Hegesipyle (who bore him a son, Cimon the younger, in *c.* 510), and his past as a "tyrant," Miltiades was bound to —and did—raise animosities in Athens. Prosecuted for his tyrannical role in the Chersonese, probably at the instigation of the rival clan of the pro-Persian Alcmaeonids, he was triumphantly acquitted as champion of resistance to Persian encroachments upon Greek freedom.

Fotofast



Miltiades, inscribed herm, *c.* 450 BC. In the Museo Nazionale di Ravenna.

The Persians were threatening Athens, which had supported the Ionian revolt with a punitive expedition, and Miltiades, who had first-hand experience of the Persians, was chosen, from 493 onward, as one of the ten generals of the Athenian land forces (unlike Themistocles, he was still thinking in terms of land warfare and of an agreement with Sparta, favoured by the Athenian landowners, the peasantry, and the rural middle class). In the summer of 490 BC the Persians landed at Marathon. The Athenians were faced with the choice of marching out and confronting them there or waiting for them at Athens: the decision was to be made by the Assembly. Miltiades was well aware of the power of the Persian cavalry, which, once out on the open plain, would wreak havoc. He was also anxious for a quick decision, because there were factions within Athens which would have welcomed a Persian victory in order to advance their own political ambitions. His arguments persuaded the Assembly, and the Athenian forces set out. A runner was sent to Sparta, to seek support of the Spartan army, but the Spartans replied that they would participate only at the conclusion of a religious festival six days later. A conflict then arose among the ten Athenian generals over whether to wait or to attack the Persians immediately. The deciding vote was to be cast by the *polemarchos* (supreme military commander) Callimachus, whom Miltiades was able to persuade to immediate action. The operational command of the army was to be held for one day in turn by each of the ten, but the four who had supported Miltiades surrendered their right to command to him.

Battle of Marathon

Occupying the foothills surrounding the bay, Miltiades waited for a favourable moment to attack. He chose a time when the Persian cavalry was nonoperational, either because it was being re-embarked for a possible direct attack on Athens or because of some other circumstance; the reason for its absence is uncertain. Charging a mile across the plain, Miltiades' forces engaged the Persian infantry, killing some 6,400 men (and capturing 7 ships) at a cost of 192 Athenian dead. The rest of the Persian force quickly embarked and put out to sea.

Following the defeat of the Persians at Marathon, Miltiades set out (spring 489 BC) with a fleet of 70 ships on an expedition to conquer those islands that had supposedly sided with Persia. His mission was not a success, and on his return to Athens there was an outcry of indignation, ably exploited by his rivals, the Alcmaeonids. Miltiades, dying of gangrene from a leg wound suffered in a mishap, was fined 50 talents, although the death penalty had been demanded. He probably died soon after in prison.

But the tragic outcome of his life did not cloud the judgment of Miltiades' historical role. His fellow citizens never forgot that it was to his initiative and leadership that they owed their victory over the Persians.

BIBLIOGRAPHY. The best writings on Miltiades in English are in H.T. WADE-GERY, *Essays in Greek History* (1958); E.S. GRUEN, "Stesimbrotus on Miltiades and Themistocles," *California Studies in Classical Antiquity*, 3:91–98 (1970); and the general works by C. HIGNETT, *Xerxes' Invasion of Greece* (1963); and A.R. BURN, *Persia and the Greeks* (1962). The only recent, though highly tendentious, biography of Miltiades is by H. BERVE, *Miltiades: Studien zur Geschichte des Mannes und seiner Zeit* (1937), who eventually qualified and severely modified his theory.

(P.T.)

Milton, John

John Milton stands next to Shakespeare among English poets; his writings and his influence are a very important part of the history of English literature, culture, and libertarian thought. He is best known for his long epic poem *Paradise Lost*, in which his "grand style" is used with superb power; its characterization of Satan is one of the supreme achievements of world literature. Milton's prose works, however, are also important as a valuable interpretation of the Puritan revolution, and they have their place in modern histories of political and religious thought.

Milton, engraving by W. Faithorne, 1670. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

Milton was born in London on December 9, 1608. His grandfather, an Oxfordshire yeoman, had been a staunch Roman Catholic who had disinherited his son, the poet's father, for turning Protestant. John Milton, Sr., going to London, made his way to prominence and a comfortable fortune as a scrivener, or notary, and through the collateral business of private banking or moneylending. Milton was to pay repeated tributes to his father's generous concern with his education. One debt was in the way of music. The father was a composer of some repute, and the poet's lifelong devotion to music is attested by the warmth of his own allusions and by his early biographers. Of his mother (died 1637) Milton said only that she was well esteemed and known for her charities. He had an older sister, Anne, and a younger brother, Christopher, who became a lawyer.

Education and early poems. Milton was educated at St. Paul's School, London. The conventional date given for his admission is 1620, but it may have been as early as 1615. In addition to his regular schoolwork in Latin, Greek, and—later—Hebrew, the boy had instruction at home, perhaps partly in modern languages, from private tutors. Milton was from childhood a voracious student; he traced the initial cause of his later blindness to his having, from his 12th year, rarely quit his books before midnight. Along with a couple of Latin exercises that have survived, his earliest attempts at verse, made when he was 15, were rhymed paraphrases of Psalms 114 and 136. Milton's closest friend, at school and later, was Charles Diodati, the son of a prominent physician of Italian origin, who went from St. Paul's to Oxford; a less intimate friendship, which also lasted beyond school, was with Alexander Gill, the son of the headmaster.

Years at Cambridge

On April 9, 1625, Milton matriculated at Christ's College, Cambridge; he received his bachelor of arts degree in March 1629 and that of master in July 1632. His experience at Cambridge can be partly gathered from his abundant Latin verse and his seven Latin prolusions (public speeches that were expected to display the speaker's learning and rhetorical and argumentative powers). Apparently in March 1626 he clashed in some way with his tutor and was rusticated. He compared his "exile" at home in London to that of his beloved Ovid and rejoiced in the opportunity to read the books (here Classical plays) "that are my life" and to see beautiful girls while strolling. On his return he was assigned to another tutor and graduated at the normal time.

Milton's nickname at the university, "the Lady," was apparently bestowed because of his handsome and delicate features and a purity of mind and behaviour that disdained the diversions of his coarser fellows. During his seven years at Cambridge he seems to have moved from some unpopularity to general respect and, among dons and cultivated students, to high esteem. He did not love the Scholastic logic that dominated the curriculum; then, as well as later, he denounced it as barren. In his last prolusion (1631–32?) he proclaimed the fervent creed and dream of a young Renaissance Humanist who was at once a Christian, a Platonist, and—whether or not he had yet read Francis Bacon—a Baconian.

Meanwhile, Milton had been learning his craft and sometimes revealing his inner self in writing Latin verse. Neo-Latin poets, such as George Buchanan or the contemporary Hugo Grotius (whom Milton later met in Paris), enjoyed European fame; and students commonly celebrated academic and national events in what was then the normal language of the university world. The young poet's sensuous instincts were revealed in these Latin poems and were further displayed, along with his mastery of Italian, in six Italian pieces (1630?), with which may be linked his first English sonnet, "O Nightingale."

Earliest extant English poems

Early in 1628 Milton wrote the earliest of his extant English poems (apart from the two psalms), "On the Death of a fair Infant," an elegy, in the Elizabethan vein, on his baby niece, Anne Phillips. In part of an academic prolusion in English couplets ("At a Vacation Exercise," July 1628) he declared his devotion to his native language, a style free from eccentricity, and exalted themes concerning nature and man. And in the Latin "Elegy VI," addressed to Diodati in the Christmas season of 1629–30, he praised the light verse kindled by wine and love but turned from that to celebrate the ascetic purity of the heroic poet. The elegy ended with a reference to a poem he had just written, his first great poem in English, "On the Morning of Christ's Nativity." Such a poem, composed shortly after his 21st birthday, may be taken as a kind of announcement of his poetical coming of age and future direction, both in its religious theme and in its mastery of conception and form and image and rhythm. Probably in the long vacation of 1631 Milton wrote the companion poems "L'Allegro" and "Il Penseroso." Less ambitious in theme than the "Nativity," they have their own complexity, concealed beneath a unique grace and charm. They may be said to embody—and excel—the urbane decorum of the lyrical and reflective pieces of Ben Jonson—a quality already manifested in the tiny "Song: On May Morning" (1629–30?) and the "Epitaph on the Marchioness of Winchester" (1631). Milton had lately (1630) also written the lines "On Shakespear," which were printed in the Shakespearean Second Folio, 1632.

Milton's scholarly and literary gifts had from childhood marked him out in the minds of his family and teachers for the ministry; in his later prose he said he had refused to "subscribe slave" in a church governed by prelacy, but the date of his negative decision is not known. As his academic career approached its end, the problem of an occupation would come up, and the poem "Ad Patrem" —though some scholars link it with *Comus* (1634)— may well have been written in 1631–32. In this piece, with a mixture of filial gratitude, firmness, and confidence in poetry and himself, Milton assumes or urges that he should not be pushed into some basely lucrative profession by a father who has fostered his literary pursuits and is himself a devotee of the muses.

Horton period (1632–38). On taking his master of arts degree in July 1632, Milton retired to his father's house —until 1635 at Hammersmith, then at the country estate at Horton, near Windsor—and proceeded to give himself the liberal education Cambridge had not provided. It was in these years that he laid the foundation or set the direction of his liberal thinking. He sought to digest the mass of history, literature, and philosophy, to gain the "insight into all seemly and generous arts and affairs" needed by the citizen-poet who would be a leader and teacher. At times he suspended his labours to visit London in quest of books or something new in mathematics or music.

Self-education

An important landmark in Milton's early development is the sober sonnet "How soon hath Time," written on his 24th birthday. During recent years he had become a figure in the small university world; now he writes this sonnet while his contemporaries are forging ahead, and he has had six months of studious and outwardly unfruitful obscurity at home. But any uneasiness about his present and future is quieted in the earnest dedication of his life to his great Taskmaster's will. What might be called the first pledges toward the fulfillment of this self-consecration were two short religious poems, "On Time" and "At a Solemn Musick" (1632–33?). These are early renderings of the beatific vision that always kindled Milton's imagination. Both, in different sequence, contrast the grossness of temporal life, the jarring discord of sin, with the eternity and harmony of heaven and good.

The same contrast is sounded in the opening lines and carried through the whole of the masque known as *Comus*. During 1630–34, perhaps in 1632, Milton had, at the invitation probably of the musician Henry Lawes, written "Arcades," a miniature masque of Jonsonian courtliness. This presumably led to a request from Lawes for another masque. *Comus* was presented on September 29, 1634, before John Egerton, earl of Bridgewater, at Ludlow Castle, in Shropshire, in honour of his becoming lord president of Wales. *Comus* was Milton's first dramatizing of his great theme, the conflict of good and evil.

If *Comus* was, in a way, a song of innocence, "Lycidas" (written in November 1637) was a song of experience— Milton's first attempt to justify the ways of God to himself and to men. His former fellow collegian, Edward King, was drowned in a shipwreck in the Irish Sea in August 1637, and Milton was asked to contribute to a volume of elegies; "Lycidas," signed "J.M.," appeared at the end of an undistinguished collection of pieces in Latin, Greek, and English (1638). The pastoral convention had, from its Greek beginnings, proved its value as a dramatic vehicle—or mask—for almost anything that a poet wished to say. Milton, working as usual within a venerable tradition, as usual re-created it. He had no reason to feel deep personal sorrow, but the drowning of a virtuous and promising young man, on the threshold of service in God's church, brought home the whole enigma of life and death, of the rightness of things in a world where such events could happen. What if his own talents—which during his years of study he has been nurturing—should be cut off? It is impossible to summarize the complexities and depths of the poem, its reverbera-

Composition of "Lycidas"

ting solidity of reference, its rich variety of pace and tone, the artistic control that dominates turbulent emotions and ends with the high serenity of victory won. "Lycidas" may be the greatest short poem in the language.

Italian tour (1638–39). In May 1638, a year after his mother's death, Milton set off—with one servant—on a visit to Italy. He sojourned chiefly in Florence, Rome, and Naples. Milton and some of his early poems—not those in English—were cordially welcomed among men of letters and patrons and their academies. The experience, which he later described at length in *The Second Defence of the People of England* (1654) and alluded to elsewhere, warmed his heart and nourished his self-confidence. (It may be remembered that at home he had very little literary acquaintance and, outside a small circle, no poetic reputation.) To his host in Naples, Giambattista Manso, marquis of Villa, who had been the patron of Torquato Tasso and Giambattista Marino, Milton wrote a bread-and-butter epistle, "Mansus" (1638–39), which is one of his best Latin poems, a graceful apologia for English culture and a no less graceful statement of his own poetic standing and hopes. He wrote also an epistle to a Roman poet, Giovanni Salzilli, and several epigrams in praise of the singing of the famous Leonora Baroni. Though he mingled happily with Catholics and was a guest at a grand entertainment given by Cardinal Francesco Barberini, the staunch Protestant did not remain silent on matters of religion. He made a call—later recorded in "Areopagitica"—on the astronomer Galileo, who was in semicaptivity because his views on the universe conflicted with the doctrine of the Roman Church. Milton felt obliged to forgo a visit to Sicily and Greece because of news of mounting political tension in England, although, as he says, he lingered some time longer in Italy. In August 1638 Milton's friend Diodati died. Milton had been informed of his loss while in Italy; on his way home he stopped to see Diodati's uncle, Giovanni Diodati, who was professor of theology at Geneva.

Middle period (1641–60). Milton returned to England in July 1639, settled in a house in London, and prepared to take in pupils; the first ones were Edward and John Phillips, the sons of his sister Anne. His elaborate pastoral elegy on Diodati, "Epitaphium Damonis," was written apparently late in 1640 and privately printed. This has commonly been ranked at the head of Milton's Latin poetry, though dissenters may ask if the fact of its commemorating his dearly loved friend has not heightened its intrinsic merits.

Milton had returned to England with plans for an Arthurian epic (plans set forth briefly in "Mansus" and "Epitaphium Damonis"); like other ambitious poets of the Renaissance, he hoped to write the great modern heroic poem. But he was also deeply anxious about the Puritan cause. In his denunciation of hireling clergy in "Lycidas," Milton had virtually declared his Puritan allegiance, and the years 1641–60 he gave almost wholly to pamphleteering in the cause of religious and civil liberty. There is an important personal passage in his fourth tract, "The Reason of Church-government Urg'd against Prelaty" (1642), that shows it was a heavy sacrifice to put aside his craving for poetic immortality and leave off his cherished studies to "embark in a troubled sea of noises and hoarse disputes," but he could not be deaf to God's summons. And, as his work went on, he was sustained by the conviction that in his many and varied defenses of liberty he was, in another way, fulfilling his epic and patriotic aspirations. His first five pamphlets (1641–42) were contributions to the attack made on prelacy in the Anglican Church by a group of Presbyterian divines (called, from their initials, the "Smectymnuus" group). The attack was directed chiefly against its episcopal hierarchy, *The Book of Common Prayer*, and ritual, as being a compromise with Rome. The group urged a return to the democratic simplicity and purity of the apostolic church. Milton's first tract was "Of Reformation Touching Church Discipline in England" (1641). This begins by assailing the Anglican service and ends with a vision of the new and grand Reformation. In the personal passage in "The Reason of Church-government" Milton explains his religious conception of poetry and the deferment of his great epic because of what he feels to be his public duty.

Allegiance to Puritanism

Notoriety came in 1643, with Milton's "Doctrine and Discipline of Divorce" (enlarged edition 1644), followed by three more tracts in 1644 and 1645 on the same theme. His discussion was presumably hastened by his own marital disaster. In June (?) 1642, several months before the outbreak of the Civil War, Milton had married Mary Powell, the daughter of a Royalist squire of Oxfordshire who owed money to his father. Success could hardly be predicted for the marriage of a scholar and poet of 33 to an uneducated girl of half his age from a large, easygoing household. The young wife, visiting her family a little later, declined—doubtless with their backing—to return. The shock must have been especially severe for a man who—as one may infer from the anguished cries that recur in the "Doctrine"—had approached marriage not only with high hopes (and an unblemished past) but with earnest prayers; and there was no release from such a tragic mistake. In the tracts, Milton argued that the sole cause admitted for divorce, adultery, might be less valid than incompatibility, and that the forced yoke of a loveless marriage was a crime against human dignity. Both religious and philosophic tradition and the way of the world (including the usually meagre education given to girls) kept woman inferior to man; Milton, without denying that view, upheld a personal and Puritan ideal of marriage as an active bond of mutual love and mental companionship, an ideal that exalted women far above the level they occupied, say, in Cavalier poetry. He was, however, attacked as a libertine by Royalists and Presbyterians alike. In 1645 friends brought about a reunion between Milton and his wife; and in 1646, when the Powells had been ruined by the war, he took into his house, for nearly a year, the whole noisy family of 10. Three daughters, Anne, Mary, and Deborah, were born in 1646, 1648, and 1652. A son died in infancy. Mrs. Milton died a few days after Deborah's birth.

First marriage

Milton's best known pamphlets (at any rate for modern readers) were published in 1644. "Of Education" is one of the last in a long line of European expositions of Renaissance Humanism. His aim was to mold boys into enlightened, cultivated, responsible citizens and leaders; its basis would be the study of the ancient classics, in due subordination to the Bible and Christian teaching. But he also gave notable emphasis to science. In "Areopagitica," on the freedom of the press, he writes as a scholar and poet and lover of books and reasserts above all his belief in the power of truth to win its way through free inquiry and discussion. The tract seems to have had very little effect in its own time.

On February 13, 1649, two weeks after the execution of Charles I, Milton's first political tract, "The Tenure of Kings and Magistrates," appeared. In it he expounds the doctrine that power resides always in the people, who delegate it to a sovereign but may, if it is abused, resume it and depose or even execute the tyrant. A month later he was invited to become secretary for foreign languages to Cromwell's Council of State. Hitherto a detached observer, Milton, in spite of his private studies, was doubtless eager to have a hand in the workings of government. He was not on the policy-making level, but he had the easy command of Latin needed for foreign correspondence. Also, as a publicist of demonstrated sympathy with the revolution, he was expected to continue his defense of the cause against the multiplying attacks on the regicides.

Secretary for foreign languages in the Council of State

Milton's first effort in this line was "Eikonoklastes" (October 1649), one of a number of answers to *Eikon Basilike*, a book edited from the King's papers by his chaplain, John Gauden; this ostensible self-portrait of the royal saint and martyr, catching the public in its first shock of horror over the execution, was one of the most potent pieces of propaganda in English history, and not even Milton could prevail against it.

During 1651 Milton was censor and supervisory editor of the chief Commonwealth newspaper, *Mercurius Politicus*, edited by Marchamont Needham. In this year he appeared on the European stage, with the publication of

his Latin *Defence of the People of England.* Charles II, in exile, had engaged Claudius Salmasius (Claude Saumaise), the most eminent of classical scholars, to arraign the regicides (*Defensio Regia pro Carolo I,* 1649). Milton was less effective in legal argument than in discrediting Salmasius by personal abuse; like some other crusaders, he tended to see opponents as monstrous enemies of a sacred cause who must be destroyed by any means.

If he was, then and later, uplifted by the vanquishing of a renowned antagonist, he was inevitably and profoundly depressed by the loss of his eyesight; it had been failing for years, and blindness became complete in the winter of 1651–52. Milton was only 43, and the great poem was still unwritten. Blindness reduced his strictly secretarial duties, though he continued through 1659 as a translator of state letters.

Onset of blindness

The Second Defence of the People of England—also in Latin, since it was also addressed to Europe at large—was much more worthy of its subject and its author. In it he celebrated the achievements of the Commonwealth leaders (though he was bold enough to warn Cromwell against one-man rule). In 1659 two more tracts on church and state were published. In "A Treatise of Civil Power in Ecclesiastical Causes" Milton argued for religious freedom (except for Roman Catholics, since Catholicism had shown itself a danger to national security). In "Considerations Touching The likeliest means to remove Hirelings out of the church" he reasserted the ideal of a clergy of apostolic simplicity of life.

His last political pamphlet, "The Readie and Easie Way to Establish a Free Commonwealth," was published in March 1660 and again, enlarged, in April. It was an act no less courageous than futile, since machinery was patently moving to bring back Charles II (he made his triumphal entry on May 29). Milton's pamphlet is a cry of incredulity and despair from the last champion of "the good Old Cause." The glories of the Commonwealth, to which he himself had given 20 years and his eyesight, are being swept away by a nation of slaves "now choosing them a captain back for Egypt." The Restoration was the last and heaviest of Milton's many disillusionments.

Estimate of the prose works

The large bulk of Milton's prose—which fills four times as many volumes as his poetry—is read only by scholars, but much of it is important for several reasons. In an age of great prose, Milton's, at its best, has a very individual if often undisciplined greatness, and "Areopagitica" at least is a classic document and possession of the race. Moreover, as the record of Milton's growth (a leftward growth, in religion and politics) and of his dreams and disillusionments, his prose works are the essential introduction to *Paradise Lost, Paradise Regained,* and *Samson Agonistes,* providing a bridge between the radiant idealism of his youth and the much-tried faith and fortitude of his later years. In particular, his treatise *On Christian Doctrine* held a central place in his thoughts and labours. He seems to have finished it by *c.* 1658–60 (it was first printed and translated by Charles Sumner in 1825). Its importance is that it expounds, with differences, the theological frame of *Paradise Lost.* Viewed in perspective, most of Milton's essential beliefs are those of traditional Christianity, but—always on authority he finds in the Bible—he departs from orthodoxy on some points: God created the world not out of nothing but out of his own substance; God the Father, the Son, and the Holy Spirit are not a coequal Trinity but a descending order; man's soul dies with his body, until revived at the resurrection (this "mortalist" heresy, which is put into the mouth of the despairing Adam, had other adherents in England). Milton's conception of the Trinity, which is too complex for summary, is apparent to alert readers of *Paradise Lost,* though the poem did not disturb generations of the orthodox. More basic for the poem is Milton's denial of predestination (which, of course, is not a heresy). Brought up, like most Anglicans of his time, as a Calvinist, he regarded himself as one at least until 1644, but his final belief was in the Arminian doctrine—the salvation not of a predestined few but of all believers, who constitute the true elect. Milton above all insisted on man's rational freedom and responsible power of choice.

Sonnets and other poems (1642–58). Milton's early poems, in English, Latin, Greek, and Italian, were published at the beginning of 1646 (dated 1645). During the 20 years given to public affairs he was mostly cut off from poetry but did write 17 occasional sonnets, versified a number of Psalms, and began the composition of *Paradise Lost.* Some of the sonnets are deeply personal: two on his blindness (1651?–55) and one on the death in 1658, some months after childbirth, of his second wife, Katherine Woodcock, whom he had married in 1656. The major sonnets have much poetical as well as autobiographical interest, and as a group they illustrate (with "Lycidas") both in texture and rhythm the beginnings of the grand style (which Milton may have learned partly from the 16th-century Italian poets Giovanni della Casa and Torquato Tasso) that was to have full scope in the epic. One is less conscious of sonnet structure and of rhymes than of a single massive unit that approaches a paragraph of Milton's blank verse.

Later life. The Restoration government executed the Commonwealth leader Sir Henry Vane the Younger and exhumed and hanged at Tyburn the bodies of Cromwell, Henry Ireton, and John Bradshaw. Milton himself, as a noted defender of the regicides, was in real danger. In the summer of 1660 a warrant was out for his arrest; he was kept in hiding by friends. In August the Act of Oblivion, granting pardon to most Commonwealth supporters, was passed. Milton was safe within its terms but was nevertheless taken into custody (and released on December 15). According to various early stories, his life was spared through the intercession either of the poet Andrew Marvell, who in 1657 had become a fellow secretary and was now a member of Parliament, or of the Royalist playwright Sir William Davenant, whose life Milton had earlier been the means of saving. It may have been decided that the blind writer was now harmless and that token proceedings against him would be enough.

Milton had given up the idea of composing a British epic. Instead he chose the most momentous event, next to the life and death of Christ, in the world's history—man's Fall from grace. It is not known when *Paradise Lost* was actually begun. Guesses have centred on 1655–58. Clearly, the lines on the poet's having fallen on evil days, in the prelude to book vii, were composed after the Restoration, and the whole may have been done pretty much in the order in which it stands. It was finished by 1665. The first edition of 1667 was in 10 books; this was reissued in 1668 and 1669, and in some of these issues Milton added the prefatory note on his use of blank verse and "The Argument." In the second edition (1674), along with some small revisions, books vii and x were each split into two; and the arguments, formerly grouped together, were placed at the head of the respective books. The main motives and events of *Paradise Lost* had precedent to a greater or lesser degree, though Milton handled them with powerful originality; like a Greek dramatist, he was reworking a story familiar in outline to his audience. His story, moreover, was one of unique truth, sacredness, and universal and eternal import, and it gave the poet the advantage of immemorial belief and association in the minds of his earlier readers. This advantage no longer operates in the same way—although, for modern readers, the fable still possesses at least the immemorial and universal import of archetypal myth.

Composition of *Paradise Lost*

As artist, Milton links himself, both proudly and humbly, with the ancients—especially Homer and the blind bards of Greek myth—but he regularly ranks his Christian theme above the themes of the pagan poets. As he had said long before, in explaining why he must postpone his epic, it was a work

> not to be obtained by the invocation of Dame Memory and her Siren daughters, but by devout prayer to that eternal Spirit who can enrich with all utterance and knowledge, and sends out his seraphim, with the hallowed fire of his altar, to touch and purify the lips of whom he pleases.

Milton's preface stressed the novelty and rightness of blank verse for a heroic poem. His manipulation of rhythm and sound is of course one of his supreme achievements. The continuous flow of his long sentences

and paragraphs is naturally unlike the dramatic blank verse of Shakespearean dialogue, and it builds up a continuous onward pressure. As a modern master of rhythm, T.S. Eliot, has said, Milton's blank verse is never monotonous. The pattern of sound is so wedded to the pattern of sense that each is essential to the other.

Paradise Regained and *Samson Agonistes*

In *Paradise Lost* (ix) Milton had spoken of "patience and heroic martyrdom" as themes unsung, though nobler than martial prowess, and this "better fortitude" was celebrated in *Paradise Regained* and *Samson Agonistes* (1671). *Paradise Regained* is a natural sequel to the long poem: Christ, the second Adam, wins back for man what the first Adam had lost. But Milton did not, as might have been expected, deal with the Crucifixion; instead, he showed Christ in the wilderness overcoming Satan the tempter, thereby proving his fitness for his ultimate trial and, in his human role, showing what man in general might achieve through strong integrity and humble obedience to the divine will. Although the poem has been found cold by the mass of readers and critics, it nevertheless has all the fire of Milton's religious and moral passion and his reverence for true heroism.

For some readers, the drama of *Samson Agonistes* (published in the same volume with *Paradise Regained*) is the most powerful and completely satisfying of Milton's major works. It is by far the greatest English drama on the Greek model. The action, up to the reported catastrophe, is wholly psychological; it is the process by which Samson, "Eyeless in Gaza at the mill with slaves," moves from preoccupation with his misery and disgrace to selfless humility and renewed spiritual strength, so that he can once more feel himself God's chosen champion. The drama must owe a great deal of its power to Milton's sense of kinship with his hero; he has been eyeless in London among a nation of slaves. But the restraint of Milton's art shows itself very clearly: there is nothing in the drama that does not belong to the story of Samson.

Altogether, if *Samson* was his last work, it was a grand testament. Like Samson, Milton was able to conquer despair or to sublimate it in his last three great poems. These expressed, not his earlier revolutionary faith in men and movements, but a purified faith in God and the regenerative strength of the individual soul.

Last years

The poet's final 16 years of life, during which these three works were finished or composed, were—outwardly and in part inwardly—peaceful, although, along with sombre or uplifting thoughts of past and present, there were concrete troubles: a frugal domestic economy necessitated by greatly diminished resources; blindness and what was sometimes a more severe affliction, the pains of gout; and a degree of friction with his daughters, due probably to faults on both sides.

Apart from the publication of books, the chief events of these years were Milton's marriage (1663) to a third wife, the young and amiable Elizabeth Minshull, who survived him, and the removal, during the plague of 1665, to a house at Chalfont St. Giles, Buckinghamshire (now a Milton museum).

Early biographers give pictures of a quiet daily routine. Milton got up at 4:00 AM and had the Hebrew Bible read to him. Meditation, reading, and dictation filled the time until midday dinner. His late and long poems were, of course, composed in his head, especially at night, as famous allusions in *Paradise Lost* indicate; when he was ready "to be milked," he would dictate, often with one leg flung over the arm of his chair. The taking of dictation, the correcting of copy, and reading aloud in various languages were services performed by paid assistants, his two nephews, his younger daughters, and friends and disciples. After dinner, says the 17th-century English antiquary John Aubrey, "he used to walk 3 or 4 hours at a time (he always had a garden where he lived)." In the evening there was reading of poetry, and he went to bed about 9:00, after a pipe and a glass of water. One refreshment was music, playing on the organ or viol and singing; "he would be cheerful even in his gout fits, and sing." And there were many, sometimes too many, callers—foreigners (who came to see the conqueror of Salmasius, not the poet) and English friends, old and young, such as men already mentioned, and the young Quaker Thomas Elwood, whom Milton helped with his Latin. In religion Milton had moved from the low-church Anglicanism of his parents to Presbyterianism to Independency to independence. In the latter part of his life, according to his early biographer John Toland, "he was not a professed member of any particular sect among Christians, he frequented none of their assemblies, nor made use of their peculiar rites in his family." But, as Dr. Johnson observed, suspending his Anglican hostility, "his studies and meditations were an habitual prayer." Milton died on November 8, 1674, "of the gout struck in," just before his 66th birthday. His burial in St. Giles', Cripplegate, was attended by "all his learned and great friends in London, not without a friendly concourse of the vulgar."

Reputation. Milton's reputation grew steadily after 1667 and was well established before Joseph Addison's papers on *Paradise Lost* appeared in *The Spectator* (1712); these were instrumental in extending the poet's fame to the Continent. His influence on 18th-century verse was immense. In the 19th century, two main streams of critical opinion are evident. On the one hand, the revolutionary Romantic poets Blake and Shelley launched the "Satanist" misinterpretation of *Paradise Lost* and made its author, like themselves, a rebel—their attitude is summed up in Blake's saying that Milton was of the devil's party without knowing it (in other words, that he had projected himself into Satan, who was the poem's real hero). On the other hand, other critics—also concentrating on the epic—threw overboard Milton's beliefs and ideas as long-dead fundamentalism and listened only to the organ voice.

The poet's influence waned during the Victorian age, and in the 20th century the new poetry and criticism launched by Ezra Pound and T.S. Eliot were strongly anti-Milton and pro-John Donne. But during the 1940s and 1950s a shift in critical attitudes took place, and dozens of books and hundreds of articles have been given to ideas and beliefs of the thinker, the publicist, and the poet, and have brought a new refinement of perception and analysis to the aesthetic study of Milton's poetry.

MAJOR WORKS

POEMS: *A Maske Presented at Ludlow Castle, 1634* [*Comus*], in an enlarged text, and "Lycidas" (1638) were both reprinted in *Poems* (1645), which included nearly all the other early pieces, "L'Allegro," "Il Penseroso," etc. The second edition (1673) reprinted the 1645 *Poems,* adding two early pieces and later sonnets; political reasons excluded four sonnets, addressed to Sir Thomas Fairfax, Cromwell, Sir Henry Vane, and the one to Skinner beginning "Cyriack, this three years' day." *Paradise Lost,* in 10 books, appeared in 1667 (revised in 12 books, 1674); *Paradise Regained* and *Samson Agonistes* were printed together in 1671.

PROSE: "Of Reformation Touching Church Discipline in England" (1641); "The Reason of Church-government Urg'd against Prelaty" (1642); "An Apology for Smectymnuus" (1642), three of Milton's five tracts against prelacy; "The Doctrine and Discipline of Divorce" (1643, enlarged 1644); "Of Education" (1644); "Areopagitica" (1644); "The Tenure of Kings and Magistrates" (1649); "Eikonoklastes" (1649), reply to John Gauden's *Eikon Basilike;* "A Treatise of Civil Power in Ecclesiastical Causes" (1659); "Considerations Touching The likeliest means to remove Hirelings out of the church" (1659); "The Readie and Easie Way to Establish a Free Commonwealth" (1660); *The History of Britain* (1670); "Of True Religion, Haeresie, Schism, Toleration, And . . . the growth of Popery" (1673); *A Brief History of Moscovia* (1682).

WORKS IN LATIN, GREEK, AND ITALIAN: (POEMS): *Poems* (1645) contained three Greek pieces and five Italian sonnets and a canzone. Most of the notable Latin poems are mentioned above; one other, "Elegy V" ("On the Coming of Spring"), was a fervently "pagan" celebration of awakening life and sexuality.

(PROSE): The seven Latin speeches delivered at Cambridge and Latin letters to friends (published 1674); *Pro Populo Anglicano Defensio* (1651); *Pro Populo Anglicano Defensio Secunda* (1654); *Pro Se Defensio* (1655); *De Doctrina Christiana* (published 1825).

BIBLIOGRAPHY

Editions: The complete works (with translations) were edited by Columbia University scholars, 18 vol. and a useful

2-vol. index (1931–40). H.F. FLETCHER edited a facsimile edition of the complete poems and manuscripts, 4 vol. (1943–48). The complete prose, with translations of the Latin works and elaborate commentaries, is being edited by D.M. WOLFE *et al.*, in 8 vol. (vol. 1–6, 1953–73). *The Student's Milton*, ed. by F.A. PATTERSON (rev. ed., 1933), has almost all the works in one volume, with some apparatus. The most comprehensive one-volume annotated edition is *Complete Poems and Major Prose*, ed. by M.Y. HUGHES (1957). Annotated editions of the complete poems are those of HELEN DARBISHIRE, 2 vol. (1952–55), with mainly textual notes; DOUGLAS BUSH (1965); JOHN CAREY and ALASTAIR FOWLER (1968); and J.T. SHAWCROSS (1963; rev. ed., 1971). There are countless editions of seleected poems. Selected *Prose* is edited by J. MAX PATRICK *et al.* (1967).

Bibliographies: D.H. STEVENS, *Reference Guide to Milton: From 1800 to the Present Day* (1930, reprinted 1967); J.H. HANFORD and C.W. CRUPI, *Milton* (1966); a full annotated reading list in *Milton's Epic Poetry*, ed. by C.A. PATRIDES (1967); CALVIN HUCKABAY, *John Milton: An Annotated Bibliography, 1929–1968* (1969); a descriptive survey in *English Poetry: Select Bibliographical Guides*, ed. by A.E. DYSON (1971). Current writings are recorded in the annual bibliographies of the Modern Humanities Research Association (1920 *et seq.*), in the *Publications of the Modern Language Association of America* (1922 *et seq.*), and in *The Milton Quarterly* (1967 *et seq.*).

Biography and criticism: Milton's utterances on himself and his work are collected in J.S. DIEKHOFF, *Milton on Himself* (1939; 2nd ed., 1965). The six early biographies, by JOHN AUBREY and others, are in *Early Lives of Milton*, ed. by HELEN DARBISHIRE (1932, reprinted 1971); some of them are in *The Student's Milton* (see above). DAVID MASSON's all-inclusive *Life* (6 vol., partly rev., 1881–96) still has its uses. The standard modern biography is WILLIAM RILEY PARKER, *Milton*, 2 vol. (1968). Four small critical biographies are those of J.H. HANFORD (1949), KENNETH MUIR (2nd ed., 1960), EMILE SAILLENS (1959; Eng. trans., 1964), and DOUGLAS BUSH (1964). *The Life Records of John Milton*, 5 vol. (1949–58), is an exhaustive compilation by J.M. FRENCH of documents and references. A compendious account of the life and work is *A Milton Handbook* (5th ed., 1970) by J.H. HANFORD and J.G. TAAFFE.

Early criticism is collected in *Milton: The Critical Heritage*, ed. by J.T. SHAWCROSS (1970), covering 1628–1731; the same editor's sequel (1972) covering 1732–1801; and *The Romantics on Milton*, ed. by J.A. WITTREICH (1970). Milton's Victorian repute is the subject of J.G. NELSON, *The Sublime Puritan* (1963). Twentieth-century criticism is surveyed in books by PATRICK MURRAY (1967) and K.L. SHARMA (1971).

The huge proliferation of criticism and scholarship since *c.* 1917 severely limits references. Some general surveys of the poetry and prose are: E.M.W. TILLYARD, *Milton* (1930; rev. ed., 1966); DAVID DAICHES, *Milton* (1957); DOUGLAS BUSH, *English Literature in the Earlier Seventeenth Century* (1945; 2nd ed. rev., 1962), ch. 12; and the modern biographies cited above. Collected studies by three eminent Miltonists are: M.Y. HUGHES, *Ten Perspectives on Milton* (1965); J.H. HANFORD, *John Milton: Poet and Humanist* (1966); and A.S.P. WOODHOUSE, *The Heavenly Muse: A Preface to Milton*, ed. by HUGH MACCALLUM (1972). Miscellaneous papers appear in the continuing *Milton Studies*, ed. by J.D. SIMMONDS (vol. 1–5, 1969–73).

Studies of major poems, early and late, are D.C. ALLEN, *The Harmonious Vision* (1954); M.H. NICOLSON, *John Milton* (1963); J.H. SUMMERS (ed.), *The Lyric and Dramatic Milton* (1965); JOHN REESING, *Milton's Poetic Art* (1968); BALACHANDRA RAJAN, *The Lofty Rhyme* (1970).

Early poems: The Latin, Greek, and Italian poems are treated by DOUGLAS BUSH and A.B. GIAMATTI in *A Variorum Commentary on the Poems of John Milton*, vol. 1 (1970). A.S.P. WOODHOUSE, DOUGLAS BUSH, and E.R. WEISMILLER deal with all the early English poems (and all the sonnets and psalms) in vol. 2 (3 pt.) of the same *Commentary* (1972). Three useful anthologies of essays are *John Milton: L'Allegro and Il Penseroso*, ed. by E.B. SAFER and T.L. ERSKINE (1970); *A Maske at Ludlow: Essays on Milton's Comus*, ed. by J.S. DIEKHOFF (1968); and *Milton's Lycidas*, ed. by C.A. PATRIDES (1961). CLEANTH BROOKS and J.E. HARDY, *Poems of Mr. John Milton* (1951), discusses the poems of the 1645 volume. Distinctive studies of major early poems are given in ROSEMOND TUVE, *Images and Themes in Five Poems by Milton* (1957). ALAN RUDRUM, *Milton: Comus and Shorter Poems* (1967), is more popular. A recent study of *Comus* is ANGUS FLETCHER, *The Transcendental Masque* (1971). The sonnets have been edited with commentaries by J.S. SMART (1921, reprinted 1966) and E.A.J. HONIGMANN (1966); see also the *Variorum Commentary* above.

Paradise Lost: Modern criticism is analyzed by IRENE SAMUEL in *Critical Approaches to Six Major English Works*, ed. by R.M. LUMIANSKY and HERSCHEL BAKER (1968). Among the scores of modern books, some are focussed on the theological and ethical "argument": J.S. DIEKHOFF, *Milton's Paradise Lost* (1946); D.H. BURDEN, *The Logical Epic* (1967); S.E. FISH, *Surprised by Sin* (1967). Some more general studies are: C.S. LEWIS, *A Preface to Paradise Lost* (1942); BALACHANDRA RAJAN, *Paradise Lost and the Seventeenth Century Reader* (1947); A.J.A. WALDOCK, *Paradise Lost and Its Critics* (1947), an attack that stimulated illuminating responses; ARNOLD STEIN, *Answerable Style* (1953); I.G. MACCAFFREY, *Paradise Lost as "Myth"* (1959); J.H. SUMMERS, *The Muse's Method: An Introduction to Paradise Lost* (1962); ANNE FERRY, *Milton's Epic Voice* (1963); L.L. MARTZ, *The Paradise Within* (1964); NORTHROP FRYE, *The Return of Eden* (1965); HELEN GARDNER, *A Reading of Paradise Lost* (1965); ALAN RUDRUM, *Milton: Paradise Lost* (1966); J.M. STEADMAN, *Milton and the Renaissance Hero* (1967) and *Milton's Epic Characters* (1968). Some special related works are: WATSON KIRKCONNELL (ed.), *The Celestial Cycle: The Theme of Paradise Lost in World Literature* (1952); J.H. SIMS, *The Bible in Milton's Epics* (1962); J.M. EVANS, *Paradise Lost and the Genesis Tradition* (1968); J.E. DUNCAN, *Milton's Earthly Paradise: A Historical Study of Eden* (1972); and the volume by J.M. STEADMAN forthcoming in the *Variorum Commentary* (see above, under *Early poems*).

Paradise Regained and Samson Agonistes: Paradise Regained and *Samson Agonistes* are treated in some books cited above. Others are ARNOLD STEIN, *Heroic Knowledge* (1957); B.K. LEWALSKI, *Milton's Brief Epic* (1966); G.M. CRUMP (ed.), *Twentieth Century Interpretations of Samson Agonistes* (1968); ALAN RUDRUM, *Milton: Samson Agonistes* (1969); *Calm of Mind*, ed. by J.A. WITTREICH (1971); *The Prison and the Pinnacle*, ed. by B. RAJAN (1973); and W. MACKELLAR's *Paradise Regained* and J.M. STEADMAN's *Samson* forthcoming in the *Variorum Commentary* (see above, under *Early poems*).

Other studies: Studies of metrics and style include ROBERT BRIDGES, *Milton's Prosody* (1921); S.E. SPROTT, *Milton's Art of Prosody* (1953); F.T. PRINCE, *The Italian Element in Milton's Verse* (1954); CHRISTOPHER RICKS, *Milton's Grand Style* (1963); E.R. WEISMILLER, in vol. 2 and later volumes of the *Variorum Commentary*.

Milton's political and religious thought is examined in some books already named (*e.g.*, A.S.P. WOODHOUSE, *The Heavenly Muse*, and *Prose Works*, ed. by D.M. WOLFE); in ARTHUR BARKER, *Milton and the Puritan Dilemma, 1641–1660* (1942); in volumes by WILLIAM HALLER (1938, 1955); and in MICHAEL FIXLER, *Milton and the Kingdoms of God* (1964).

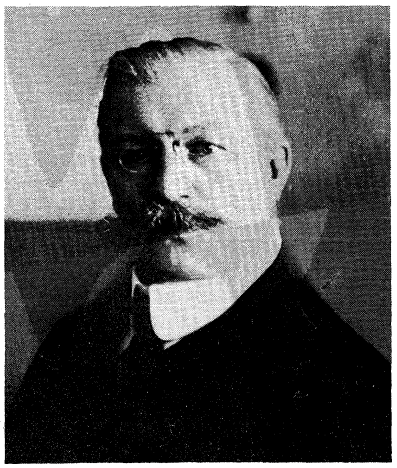
The nature and degrees of Milton's theological orthodoxy and "heresy" have caused controversy: see MAURICE KELLEY, *This Great Argument* (1941), and vol. 6 of the *Prose Works* (cited above); C.A. PATRIDES, *Milton and the Christian Tradition* (1966); and W.B. HUNTER, C.A. PATRIDES, and J.H. ADAMSON, *Bright Essence* (1971).

Some books on special subjects are: IRENE SAMUEL, *Plato and Milton* (1947) and *Dante and Milton* (1966); KESTER SVENDSEN, *Milton and Science* (1956); THOMAS KRANIDAS, *The Fierce Equation: A Study of Milton's Decorum* (1965); and MARCIA R. POINTON, *Milton and English Art* (1970). E.S. LE COMTE compiled *A Milton Dictionary* (1961), containing much precise information. A Milton encyclopaedia is being edited by P.E. BOYETTE *et al.* (D.B.)

Milyukov, Pavel Nikolayevich

The statesman and historian Pavel Nikolayevich Milyukov, after playing a significant role in the events leading to the Russian Revolution of March 1917, served as foreign minister in the Provisional government. His advocacy of constitutional monarchy and his insistence upon continuing the war with Germany proved unpopular. After two months in office he resigned and later went into exile. He remains one of the greatest of Russia's liberal historians.

Milyukov was born in Moscow on January 27 (January 15, old style), 1859. He entered Moscow University in 1877 and soon attained academic distinction as a specialist in Russian history. His three-volume *Ocherk: po istorii russkoy kultury* (1896–1903; Eng. trans. *Outlines of Russian Culture*, 1942) brought him nationwide renown. He was, however, to become even better known as a spokesman for imperial Russia's nascent liberal movement. His concern with contemporary political issues



Milyukov.
EB Inc.

led in 1895 to the loss of his teaching post, whereupon he went abroad, travelling widely and developing a deep admiration for the political values of more advanced democratic countries. On two occasions, in 1903 and 1904–05, he visited the United States, where he gave public lectures on Russian history and politics. As a convinced "westernizer," Milyukov saw tsarist absolutism as the main obstacle to Russia's progress. He strove to make Russian liberalism a more broadly based movement committed to a radical (but not Socialist) transformation of the empire's political and social structure. In particular, he stood for the rule of law, parliamentary government, universal suffrage, a broad range of civil freedoms, the expansion of popular education, and land reform. He was perhaps a better judge of abstract ideas than of human nature and was too inflexible to succeed in practical politics.

Political journalism

As a contributor to a clandestinely circulated journal, *Osvobòzhdeniye* ("Liberation"), founded in 1902, Milyukov did much to swing the moderate members of the zemstvos (local government bodies) to the left. During the revolutionary year 1905, he was active in forming the Union of Unions, a broad alliance of professional associations, and subsequently the Constitutional Democratic, or Kadet, Party. As editor of the Kadet daily newspaper *Rech* ("Speech") and a member of the party's Central Committee, he directed its tactics in Russia's first nation-wide election, which brought the Kadets success at the polls. The gulf between the newly created Duma, or national assembly, and the government could not be bridged, however, and in July 1906 the assembly was prematurely dissolved. Milyukov helped to draft the Vyborg Manifesto, appealing for a campaign of passive resistance to the government; this, technically speaking, was an illegal act. Many moderates thought that his tactics, which involved a measure of cooperation with the revolutionary Socialists, were too radical.

Subsequently Milyukov moved to the right. In the third and fourth Dumas (1907–17), he led the Kadets' depleted parliamentary faction in condemning bureaucratic abuses and championing progressive causes, notably the claims to cultural or political autonomy of national minority groups within the empire. In foreign policy Milyukov stood for cautious support to the Slav peoples of the Balkans, an area he knew well from personal observation. His attitude on this issue and on Russia's claims to the Turkish straits hardened after the outbreak of World War I. Milyukov adopted a patriotic "defensist" policy, which in 1915 led him to criticize the imperial regime for its inefficiency in organizing Russia's resources for the war effort, measured by the standards of its democratic Western allies. He helped to form a Progressive Bloc of moderate Duma leaders, who sought by gentle pressure from public opinion to secure a ministry enjoying the nation's confidence. When these overcautious tactics proved unavailing, Milyukov swung sharply to the left and, on November 14, 1916, delivered a powerful speech in the chamber implying that the government was guilty of treason. His charges had no real substance, but they greatly stimulated revolutionary sentiment, which soon exceeded the bounds Milyukov had expected or desired.

When Tsar Nicholas II abdicated under this pressure on March 15, 1917, Milyukov wanted to preserve the monarchy as a stabilizing force but found little support. The revolutionary tide was now running strongly against him. In the liberal Provisional government under Prince Georgy Lvov, which Milyukov helped to constitute, he towered intellectually over his colleagues but was soon outmanoeuvred by the Socialist leader Aleksandr Kerensky, who favoured a policy of concessions to popular demands, notably on the question of taking immediate steps toward a negotiated peace. As foreign minister, Milyukov clung to the wartime alliances and the demand for Russian possession of the Turkish straits. He resisted heavy pressure from inside and outside the government to send a note to the Allies redefining war aims, and when, on May 1, he yielded, the note was phrased ambiguously. Disturbances broke out in the streets of Petrograd, and Milyukov was obliged to resign (May 15).

Foreign minister

Thereafter he endeavoured to counter the new coalition government's leftward swing by rallying the moderates, but with little success. The Bolshevik seizure of power in November 1917 forced him to leave Petrograd for southern Russia, where he became political counsellor to the leaders of the White Volunteer Army. Subsequently he emigrated to Paris. In 1923, reverting to his earlier radicalism, he sought to unite the various currents of émigré opinion on a common republican democratic platform. This, too, proved a failure: he found that his compatriots preferred more extreme political doctrines. Undismayed, Milyukov held fast to his liberal principles and remained active in politics and scholarship until his death at Aix-les-Bains, France, on March 31, 1943.

BIBLIOGRAPHY. THOMAS RIHA, *A Russian European: Paul Miliukov in Russian Politics* (1969), is a thorough study that emphasizes the period 1905–17. Milyukov's works include *История второй русской революции*, 3 vol. (1921–24), a history of the Russian Revolution of 1917; *Россия на переломе; большевистский период русской революции*, 2 vol. (1927), Russia at the turning point; and *Outlines of Russian Culture* (abridged Eng. trans., 3 vol., 1942; orig. pub. in Russian, 3 vol., 1896–1903). His "Reminiscences," 2 vol. (1955), have in part been translated into English as *Political Memoirs, 1905–1917*, ed. by A.P. MENDEL (1967).

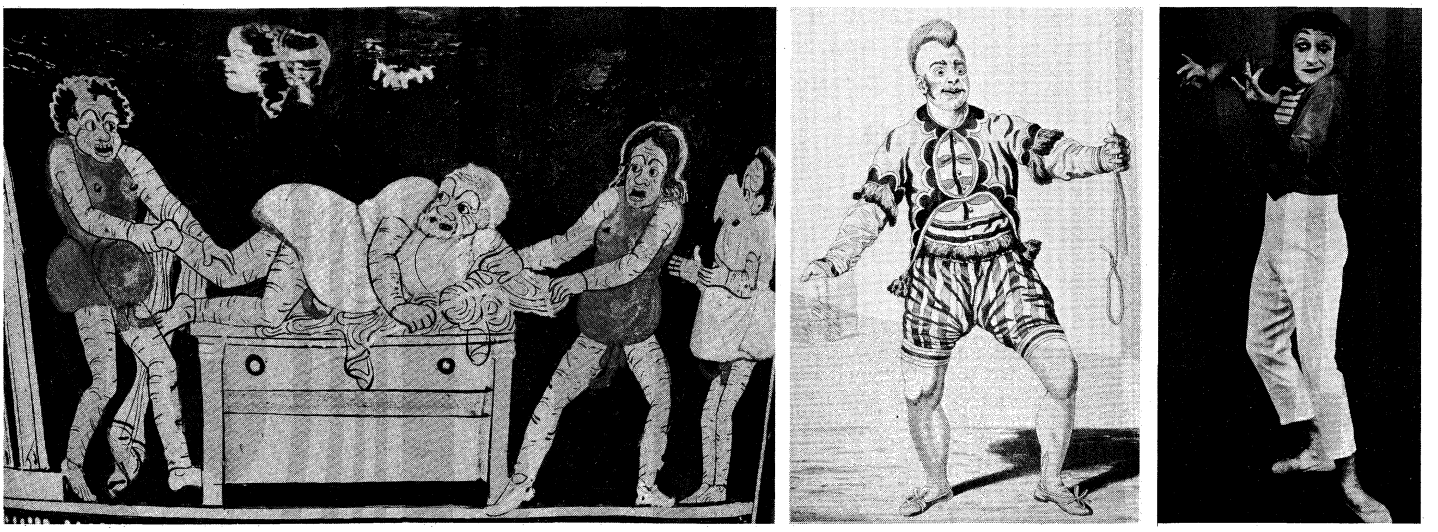
(J.L.H.K.)

Mime and Pantomime

The words mime and pantomime in the 20th century have often been used interchangeably and are associated mainly with forms of silent comedic acting. They are classic terms originally descriptive of two acting forms. The mime was the Greek comedic actor who spoke with the enactment of his role; the pantomimist was the Roman actor who performed his role with dance and song, or chant. As representation of emotions and actions by bodily movements and gestures in the telling of a story, pantomime has occurred in all primitive cultures, as in war dances, in animal mimicry, and in sacrificial rites. The stylized artificialities in Restoration comedy, the conventionalized symbolic acting of the Chinese opera or Japanese Nō drama, and the acting of comedic performers in the silent films, such as Charlie Chaplin and Buster Keaton, were in the realm of pantomime, as are any dance forms concerned with mood or story portrayals. Musical comedy and opera are pantomimic forms. Actors who speak their roles are also mimes.

In the dramatic theatre it is through mime that the player projects to the audience a visual understanding of his role and the nuances of his words. Not a literal imitator, the mime enlarges and emphasizes the character he performs, revealing his point of view of that character and all its meaning. His gestures are symbols that are a crystallization of human experience—a kind of visual code made up of gestures, smiles, glances, reactions to reflective thought, along with physical action the meanings of which are instantly understood by the audience.

This article attempts to sort out and describe the numer-



Comic pantomime.
(Left) A scene of Greek Old Comedy, 4th-century-BC vase painting by Assteas of Poseidonia. (Centre) Joseph Grimaldi as a clown in *Mother Goose;* print by S. de Wile, 1807, published by Richard Bentley, 1846. (Right) Marcel Marceau as Bip, a character of his own invention, playing an imaginary violin.
By courtesy of (left) the Staatliche Museen Preussischer Kulturbesitz, Antikenabteilung, Berlin; photographs, (centre) Radio Times Hulton Picture Library, (right) Columbia Artists Management

ous kinds of theatre representative of mime or pantomime during the course of theatrical history.

EARLY FORMS

Scraps of manuscripts and vase paintings of Greece from *c.* 600 BC, and fragments of the latter dating back two centuries earlier, attest to the early origins of mime as a theatrical form.

Greek forms. The earliest recorded form of the mime theatre was the *phlyake,* a farcical play on a theme of real life and mythology developed in Sparta, Megara, and other nearby Dorian centres of Peloponnesus. The *phlyakes* were played by stock figures clad in tights or short chitons, grotesquely padded front and back, and most wearing an exaggerated phallus. Like the tragic actors, they wore masks with enormous open mouths that amplified their words. It is believed that most of their dialogue was improvised, although by the 5th century BC their performances may have become more professional. The Athenian playwrights, including Aristophanes, looked down on these rustic plays much as the burlesque theatre was later regarded. The Athenians did not hesitate to make use of such techniques, however, analogies being obvious in the satyr plays that closed the tragic trilogies.

Dorian mimes

The earliest and most famous writer of Dorian mime plays was Epicharmus (*c.* 530–*c.* 400 BC). From 485 to 467 he won fame as a writer of mime in Syracuse, and Aristotle seems to suggest that he was the first dramatic poet of merit to write comic plots and to give literary refinements to the earlier, rude, improvised forms of mime. The few extant fragments and records show that his plays were adaptations of classical legends of the gods and mortal heroes, and political satire and debate played a large part in them. In one episode Odysseus rehearses to himself the story he will tell his comrades on his return home, not unlike Parolles in Shakespeare's *All's Well That Ends Well.* Since the mimes were comic, it must be assumed that the heroic characters were burlesqued. The plays contained not only proverbs but many puns designed to amuse the public. The carefully structured short verse mimes of Herodas (3rd century BC), much more of whose work survived, afford a clearer picture of the form. His mimes, scenes of contemporary life rather than mythological burlesques, were nonetheless broadly comic and often ribald in theme. Theocritus (*c.* 310–250 BC), a Syracusan and the first writer of pastoral poetry, is the most literary of them all. He and Sophron (fl. *c.* 430 BC), also a Syracusan, presented more lifelike mimes, with sharp topical comments and much proverbial wisdom.

The theatre of the Dorian mime comprised much that, by imitation or by coincidence, was to be characteristic of mime and pantomime well into the modern period. It was presumably a combination of words and actions with acrobatics and some dance, and it was comedic. Such stock characters as the thievish slave, the comic doctor, and others who represented aspects of contemporary life appeared repeatedly in the mimes. This theatre had no hesitancy in employing its gods and heroes as human, comedic characters. The mimes and mountebanks of Peloponnesus performed on crude platforms in public squares or outlying fields. The stage was reached by a short flight of steps, and a backdrop often had an opening with a painted window frame surrounding it. A minimum of furniture seems to have been used—a chair or throne, an altar, or a table.

The Dorian mimes were satiric social commentators aware of the temper of their audience. Their farce contained demonic figures and witches; animal masks sometimes appeared in the many licentious dances; and jugglers and acrobats were a part of this theatre. The "literary playwrights" of Athens could not ignore the influence of the mimes: the slave Xanthias tells the audience at the opening of Aristophanes' *Wasps,* "Do not expect anything great and wonderful from us. Do not look for merriment filched from Megara. We do not have a couple of slaves here to scatter nuts from a basket among the spectators nor have we Herakles defrauded of his dinner." Other writers mention the derision of Athenian audiences for the unbecoming jests of the Megarian comedy, and it seems apparent that they considered its stock characters "low comedy."

Indeed, the mime actor was much like the later burlesque and vaudeville comedian and knew what would delight his audience—lasciviousness, sentimentality, poking fun at authority. Throughout the entire history of mime there is a link between its form and the ethics of the day. The mimes appear to have symbolized the moral and religious codes of their period in reverse, and they performed to delight the popular mind.

Mime in Rome. Dorians from Sicilian Syracuse had been of the greatest importance in the form of the Greek mime, and, like the main body of Greek tragedy and comedy, this form was readily accepted on the Italian peninsula, encouraged as well by the ability of the wordless dancer or the broadly descriptive performances of the mimes to circumvent easily all language barriers. It becomes difficult, during the several centuries of cultural mixing about the Mediterranean, to separate pure mime from the main body of Roman comedy—*e.g.,* that of Terence and Plautus—so many elements did they share

with one another and with the Greek mime. The genre that arose apparently in the Campanian town of Atella, known as the *fabula Atellana,* which was a major force on the Roman comedy, bears striking resemblances in themes, character types, costuming, and staging to the Dorian mimes, and its relation to the sophisticated *fabulae* (*fabula,* a generic word for a play) of the Roman stage was similar to that of Megara to Athens.

Mimi and pantomimi

Roman mime, in the early centuries, was bound up with crude stage performances that included improvised farcical scenes, acrobatics, juggling, and the like, acted out by persons of low social status. The companies included women as well as men, and in playing, sex roles were not uncommonly reversed. About the time of Caesar a certain literary element was introduced and mimes were written for enactment by the *mimi*; still, they remained burlesques often so licentious that adultery was encouraged on the public stage. Another similar Roman institution was that of the *pantomimi,* masked dancers who executed their stories by extraordinarily skilled gesture, posture, and rhythmic movement alone. Pylades, a *pantomimus* of tragedy, and Bathyllus, a lighter performer, won acclaim in 22 BC; both played all roles in their dumb shows by changes of masks. A large chorus and orchestra accompanied these solo dancers, and their performances and those of others not only became popular events but also were welcomed into fashionable homes and were patronized by later emperors. The *pantomimi* played on themes from mythology and history rather than the everyday scenes enacted by the *mimi.*

Both the Roman mime and pantomime, the former largely on a low order of literacy despite finer writers like Lucan (AD 39–65) and Statius (*c.* AD 45–96), were so widely favoured that they virtually drove the more noted writers from the stage. Among the causes of this popularity may have been their appeal to the Germanic invaders, who found in them a pleasure not blunted by language barriers. The theatre historian Allardyce Nicoll believes that pantomime may have assumed the form of operatic dance associated in the 20th century with the Russian ballet. Despite the defense by Lucian (*c.* AD 125–*c.* 190)—whose dithyrambic praise of dumb shows, the dialogue *De saltatione* ("Concerning Dance-Acrobatics"), had great influence on dancing masters a millennium and more later and inspired the English pantomime—and by others, these forms continued several centuries into the Christian era in an aura of popularity and disrepute. Perhaps more than any other factor, they were responsible for the church's opposition to theatre, its excommunication of *mimi* in the 5th century and closing of theatres in the 6th.

Oriental dance-dramas. In the East the art of pantomime was developed long before it achieved definite form in the Western world. In India the union of music, song, and dance with character portrayals took place many centuries before Christ. Out of such native drama, with added dialogue, grew the *bhārata-nāṭya,* one of the traditional Hindu dance-dramas, first described by the sage Bhārata (*c.* 6th century BC), on the command of the god Brahmā to introduce dramatic entertainments. The two-hour performances feature a solo dancer backed by musicians and a singer; the dancer neither leaves the stage nor changes costumes. Sections of dance alternate with pantomime, the latter using conventionalized gesture-language. Pantomimic scenes derived from the mythology of Viṣṇu (Vishnu) are still sometimes enacted by the Bengali *jātrās,* folk pageants combining words and conventionalized mime, and the *rasas,* folk dance-dramas. In the 20th century, authentic classical dancing of India was introduced in the United States by Uday Shankar to the accompaniment of Indian music played on native instruments (see SOUTH ASIAN PEOPLES, ARTS OF: *The performing arts*).

Hindu mime

In the theatre of China and Japan pantomime acquired a role unknown in the West, becoming an integral part of the major dramatic genres. The conventions of gesture and movement, as well as the symbolism of the stage properties, in Chinese drama are immense in scope and mystifying to those unfamiliar with the traditional forms. On the Japanese stage dramatic narrative is inseparable from dance and pantomimic conventions, both in the popular Kabuki form and in the more esoteric Nō theatre (see DANCE AND THEATRE, EAST ASIAN; KABUKI THEATRE; NO THEATRE).

Medieval forms. Almost all of the theatre of the Middle Ages can be likened to a kind of travelling circus. Despite the opposition of the church, which instructed bishops to have no players at their tables and frowned on men playing women's roles, the Middle Ages saw the advent at the courts of those malformed figures, the jesters, whose twin-peaked caps with bells became symbols of their profession, and who later became figures in puppet shows and Shakespearean plays both comic and tragic.

The traditions of the *mimi,* together with the many other acts of the skilled clowns, dancers, improvisers of robust scenarios, and others who entertained both the rowdy public and the courts, kept alive the rough spirit of the Roman popular theatre. In the later Middle Ages those itinerant groups spawned such institutions as that of the troubadours, who were the earliest dancing masters at the European courts, and the public entertainments that led eventually to the Italian commedia dell'arte and provided the bases for the popular comedic theatre of Elizabethan England and other countries.

Major plays of the Elizabethan repertory contain evidence of the continuing popularity of "dumb shows," as they were called. Perhaps the most famous Shakespearean example is the dumb-show murder in *Hamlet,* and the clowning of Bottom and his fellows in *A Midsummer Night's Dream* is a delightful spoof of the melodrama of these shows. Pantomime persisted, too, alongside many other theatrical ingredients in the formal masques that became fashionable in the courts of the 17th century.

COMMEDIA DELL'ARTE AND ITS DESCENDANTS

The players who by the 18th century were calling their form commedia dell'arte (*i.e.,* professional comedy), to distinguish themselves from the free-wheeling styles of the amateur mimes and pantomimes and from the actors of literary comedy, employed a style that dated at least from the mid-16th century, when the "Italian comedians," the *teatro dell'improviso,* became popular in both the streets and the courts of Europe. At first improvising around scenarios containing a few key lines and later with fuller scripts, these actors had a large repertory of conventionalized, usually comic, pantomimic movements, called *lazzi.* They sought to rid such improvisatory theatre of free-for-all playing and to establish a more orderly form in which the actors could listen more closely so they might react more properly. In his (1727) history of Italian theatre, Luigi Riccoboni remarked that "the curse of improvisation is that the skill of the good actor is at the mercy of the person with whom he is conversing." By this he meant to scourge the mime-actor who thought of himself as a solo performer. The stock characters who appeared in the situation farces of the *commedia* were widely adapted, by Shakespeare and Molière among others, outside Italy for several centuries: Arlecchino (Harlequin), with his suit of many-coloured patches; Colombina his love; Pantalone, the graybeard with red fez, doublet-and-tights, and black robe; Dottore (Doctor), often bespectacled; the picaresque Scaramuccia (Scaramouche); the flamboyant Capitano; and the *zanni,* a half-mad but clever comic servant. For a fuller discussion of this form, see the article COMMEDIA DELL'ARTE.

The descendants of the *commedia,* which lost its own creative energies late in the 18th century, are numerous, and its characters held European stage into the 20th century. Prohibitions on spoken dialogue in all but royal-patent theatres in Paris and London led to the silent comic dancing and miming of Arlequin in Parisian fairs and to the English pantomimic tradition begun by the dancing master John Weaver at Drury Lane Theatre, London, in 1702. Within 15 years Weaver and John Rich of Lincoln's Inn Fields Theatre were competing in their so-called pantomimes, which became increasingly elaborate with their popularity. Weaver's shows were dramatic ballets with

Harlequin and his fellows

several performers (a form that France knew as *ballet d'action*), whereas Rich introduced grotesque scenes of Harlequin and Columbine, often pursued relentlessly by her father, Pantaloon; neither was the art of the solo *pantomimus* of Rome save in name. Audiences came to demand such pieces before or after (or both) even serious plays: Garrick was to remark, "If you won't come to hear Lear and Hamlet, I must give you Harlequin," and he did. Gradually Harlequin, with his magic wand, was replaced in eminence by the clown, an adaptation of the *zanni*. In the 19th century the "greatest" of the clowns, Joseph Grimaldi, with his clown in *Mother Goose*, helped bring to this pantomime the wealth of folklore, a repertory that was also tapped by others for "Red Riding Hood," "Cinderella," "Blue Beard," "Aladdin," and similar tales. This tradition evolved into the Christmas pantomine, still one of the most popular of English holiday entertainments. Among its continuing conventions is the role of "principal boy," which has been played by many of England's leading comediennes.

In Germany, too, Harlequin proved a threat to "serious" drama, and in France Harlequin's course roughly paralleled that in England. The French rejected the clown, however, developing in his stead the pale-faced, moonstruck Pierrot, a figure made popular first by Jean-Gaspard Deburau and carried through six subsequent generations of his descendants. Many other managers specialized in dumb-show melodramas and acrobatic spectacles under the term "pantomime."

MODERN TRENDS

Among the most significant developments in mime and pantomime since the late 19th century have been the strong impact made by persons experimenting with theories of body movement on the art dance (ballet and so-called modern dance as distinct from popular stage dancing and acrobatics), the re-emergence of mime and pantomime as silent theatrical performances by one or more mimes (as they refer to themselves), and an increased awareness of the value of absolute bodily control by actors in all dramatic media.

Mime in modern dance

Perhaps the major characteristic of the modern dance since its emergence as a recognizable trend around 1900 is its constant search for new methods of expression and communication through movement alone. Its pioneer theorists, including François Delsarte, Émile Jaques-Dalcroze, and Rudolf Laban, were no less theorists and experimenters with new forms and styles than were its practitioners, Isadora Duncan, Ruth St. Denis, Mary Wigman, Martha Graham, Doris Humphrey, Katherine Dunham, Merce Cunningham, Alvin Ailey, and Alwin Nikolais. Ballet, which had reached a nadir around the turn of the century, was itself given fresh artistic vigour in good part through its interactions with this new dance and through the influences of choreographers like George Balanchine and Jerome Robbins, who were willing to borrow or could move easily between the two forms. For more detailed information on this movement in dance, see the article MODERN DANCE.

In a constant teaching-learning relation with the modern dancers were others whose interest lay not only in dance but in pure mime as well. The theories and teachings of Étienne Decroux (1898–), a student of Dalcroze in Paris, were brought to the U.S. by Alvin Epstein. The leading figure in American mime was Angna Enters, who had studied with another Dalcroze pupil, Itō Michio, but had pursued her own belief in an eclectic approach: that each performer must evolve his own mode of expression, be it realistic and mimetically descriptive or totally abstract. In France, important work, in addition to that of Decroux, was done by Jean-Louis Barrault and by Marcel Marceau, who invented the white-faced character Bip, for whom he became famous. Bip is his translation into his own individual style of pantomime of the classic Pierrot of the commedia dell'arte. Marceau defines his work as the art of expressing feelings by attitudes and expressing words through gestures. Pierrot has been a subject for pantomimists ever since his invention, appearing in many versions. Deburau was noted for his pale-faced version, and Charlie Chaplin's little man is another classic example. The Russian play *He Who Gets Slapped* (1914) by Leonid Andreyev presents another form of Pierrot as a tragic circus clown.

The art of silent but precise communication of feeling was especially visible, indeed necessary, on the motion-picture screen, particularly in the "silent era" before the advent of sound pictures in 1927. Among many film actors who made silent pantomime an aesthetic delight were Chaplin and Ben Turpin, both of whom, interestingly, combined the broad comedy and the pathos of character inherent in the centuries-old traditions of the mime and pantomime.

Mime in modern acting

The realistic drama of Ibsen, Chekhov, and Shaw that came into vogue early in the century demanded from its performers an attention to fine detail in movement to replace the broad gesture of melodrama and the bravura acting of the romantic theatre. Following the pioneer theories and training of Konstantin Stanislavsky in Russia, schools and academies of acting began to include pantomime in their curricula, using it as a means both of communication and of achieving maximum control by the actor of all parts of his body. Playwriting itself shows an awareness of the dramatic effectiveness of wordless action. Samuel Beckett's plays depend at times as much on movement prescribed by the playwright as on his dialogue, and of note is the appearance of the mime Alvin Epstein in the New York productions of Beckett's *Waiting for Godot* and *Endgame*. Indeed, the whole "theatre of the absurd" movement after World War II showed in part a rejection of a literary theatre—as in Eugène Ionesco's *The Bald Soprano* ridiculing language as a means of communication—and looked to the rougher side of theatrical tradition for its inspiration.

Many reasons may be offered for the persistence of mime and pantomime as an inherent part of theatrical action. High among the reasons, however, is the fact that these words seem to describe, whatever the forms they connote at any given place or time in history, a theatre that is easily understood and enjoyed by all classes of society, from the least to the most literate and sophisticated. They both deal with the enacting, with or without words, spoken or sung, of aspects of human behaviour that are at once universal and yet related to the manners and modes of contemporary life. The forms may be of the commonplace or vulgar, but the spirit is of the human essence that all art seeks to discover.

BIBLIOGRAPHY. Books covering mime and pantomime within a larger context of drama and dance include: A. NICOLL, *Masks, Mimes and Miracles: Studies in the Popular Theatre* (1931, reprinted 1964); W. BEARE, "The Mime," in *The Roman Stage: A Short History of Latin Drama in the Time of the Republic*, 3rd ed., ch. 18 (1965); JEAN-LOUIS BARRAULT, *The Theatre of Jean-Louis Barrault* (Eng. trans., 1961); and W. SORELL (ed.), *The Dance Has Many Faces*, 2nd ed. (1966). A more exclusive focus on the history of these forms may be found in R.J. BROADBENT, *A History of Pantomime* (1901, reprinted 1965); A.E. WILSON, *The Story of Pantomime* (1949), *King Panto: The Story of Pantomime* (1935); and A. EPSTEIN, "The Mime Theatre of Etienne Decroux," *Chrysalis*, 11:3–13 (1958). The interaction between mime and the commedia dell'arte is among the subjects covered in M. SAND, *History of the Harlequinade* (1915, reprinted 1968); C.W. BEAUMONT, *History of Harlequin* (1926, reprinted 1967); and A. NICOLL, *World of Harlequin* (1963). The nature of mime and analyses of mime and dance movements are in RUDOLF VON LABAN, *Mastery of Movement* (Eng. trans., 1950; 2nd rev. ed., 1960); A. MORGAN, *An Hour with Delsarte: A Study of Expression* (1892); E. JAQUES-DALCROZE, *Eurhythmics, Art and Education* (Eng. trans., 1930); and A. ENTERS, *On Mime* (1965).

(A.E.)

Mimicry

Mimicry is a biological phenomenon in which two or more organisms, usually biologically unrelated, or only distantly related, resemble each other closely, the resemblance itself conferring advantage upon one or both of them. The type of natural selection and the basis of the selective advantage distinguish mimicry from other types of resemblance, especially convergence. In mimicry the

animate agent of selection (*e.g.*, a predator, a symbiont, the host of a parasite, depending on the type of mimicry encountered) interacts directly with at least two of the similar forms and is deceived by their similarity. The participants in the mimicry situation, therefore, normally must occupy the same geographical area. A convergent resemblance, on the other hand, results from the action of similar forces of natural selection (*e.g.*, temperature, food habits, predation) on unrelated organisms, which may be geographically or temporally isolated from each other.

In the most studied mimetic relationships the advantage is one-sided, one species (the mimic) gaining advantage from a resemblance to the other (the model). Since the discovery of mimicry in butterflies in the mid-19th century, a great many plants and animals have been found to be mimetic. In many cases the organisms involved belong to the same class, order, or even family, but numerous instances are known of plants mimicking animals and of animals mimicking plants. Although the best known examples of mimicry involve similarity of appearance, investigations have disclosed fascinating cases in which the resemblance involves sound, smell, light production (bioluminescence), behaviour, and even biochemistry.

A key element in virtually every mimetic situation is deception by the mimic, perpetrated upon a third party (sometimes called the receiver), which mistakes the mimic for the model. This third party, which exerts the selective pressure that makes mimicry advantageous, may be the collective potential predators upon the mimic, potential prey animals of a predacious mimic, or even one sex of the mimic's own species. In some cases, such as host mimicry by parasites, the organism deceived is the model.

Definition of mimicry

Because of the variety of situations in which mimicry occurs, a formal definition must rest upon the effect of certain key communicative signals upon the appropriate receiver and the resultant evolutionary effect upon the emitters of the signals. Mimicry may be defined as a situation in which virtually identical signals, emitted by two different organisms, have in common at least one receiver that reacts in the same manner to both signals because it is advantageous to react in that manner to one of them (that of the model), although it may be disadvantageous to react thus to the other.

The distinction between camouflage and mimicry is not always clear when only the model and the mimic are at hand. When the signal receiver is known and its reactions understood, however, the distinction is quite clear: in mimicry the signals have a special significance for the receiver, hence also for the sender, which has evolved the signals in order to be perceived by the receiver; in camouflage the sender seeks to avoid detection by the receiver through the imitation of the background or substrate or of a particular element of the environment. For further information on camouflage, see COLORATION, BIOLOGICAL.

BASIC TYPES OF MIMICRY

Batesian mimicry. In 1862 the English naturalist Henry W. Bates published an explanation for unexpected similarities in appearance between certain Brazilian forest butterflies of two distinct families. Members of one family, the Heliconiidae, are unpalatable to birds and are conspicuously coloured; members of the other family, the Pieridae, are edible to predators. Bates concluded that the conspicuous coloration of the inedible species must serve as a warning for predators that had learned of their inedibility through experience. The deceptively similar colour patterns of the edible species would provide protection from the same predators. This form of mimicry, in which a defenseless organism bears a close resemblance to a noxious and conspicuous one, is called Batesian, in honour of its discoverer.

Müllerian mimicry. Bates observed, but could not explain, a resemblance among several unrelated butterflies, all of which were known to be inedible. There seemed to be no reason for these species, each of which had an ample defense with which to back up the warning coloration, to be similar. In 1878 a German zoologist, Fritz Müller, suggested that an explanation for this so-called Bates' paradox might lie in the advantage to one inedible species in having a predator learn from another. Once the predator has learned to avoid the particular colour pattern with which it had its initial contact, it would then avoid all other similarly patterned species, edible and inedible. The initial learning experience of the predator often results in death or damage to the inedible individual that provided the lesson; there is thus some cost to the species that teaches the predator of its inedibility. Evidence indicates that there is little or no inherited recognition by predators; each individual learns of noxious or inedible species by sampling them. Other inedible species resembling the first, however, do not have to sacrifice individuals to teach this same predator, and the number of individuals sacrificed in educating the entire predator population is spread over all of the species sharing the same warning pattern. The tendency of inedible or noxious species to resemble each other is called Müllerian mimicry.

The value of similarity between models

Aggressive mimicry. In some situations, it is of advantage to a predator to resemble its prey, or a parasite its host. Aggressive mimicry, for which the phrase a wolf in sheep's clothing is an apt description, does not involve warning mechanisms. The mimic adopts certain of the recognition marks of its model, in order to secure advantage over the model itself or over a third species that interacts with the model. The model may be only a single stage of the life cycle, as in the case of parasitic cuckoos, the eggs of which resemble those of their hosts (see below *The occurrence of mimicry among plants and animals*), or it may be a prey of the mimic's victim, as in the case of angler fishes, which possess wormlike lures that bring other fishes within reach.

Automimicry. The phenomenon of automimicry involves the advantage gained by some members of a species from its resemblance to others of the same species. Males of many bees and wasps, although defenseless, are protected from predators by their resemblance to females that are equipped with stingers. Some butterflies are able to gain protection against predators through the ability to absorb, tolerate, and retain in the immature (larval) stage, poisons from the plants on which they feed. Individuals or even whole populations of such butterflies may fail to acquire such protection, as a result of feeding on nonpoisonous plants, but they are avoided by predators that have sampled protected individuals of the same species.

Other forms of mimicry. Many forms of mimicry do not fit neatly into any of the above categories. The roles of mimic, model, and receiver may be juxtaposed and multiplied to provide intricate and remarkable relationships, the unravelling of which may take years of study. One such case, studied in detail by a German zoologist, Robert Mertens, involves several species of poisonous and nonpoisonous South American snakes. The coral snakes (*Micrurus*), long recognized as dangerously poisonous, possess a brilliant red, black, and yellow ringed pattern generally believed to serve as warning coloration. Nearly identical colour patterns are found in several genera of nonpoisonous and mildly poisonous snakes, usually called false coral snakes or coral mimics. It was long assumed that the false coral snakes gained protection from their resemblance to the virulently poisonous ones, but Mertens has pointed out that the predator that samples a true coral snake will probably not survive to act as a selective force on the less dangerous species. It is quite likely that the role of receiver in this mimicry situation is played by the predator that has survived a painful but educational encounter with the moderately poisonous forms and therefore avoids both the highly poisonous true coral snakes and the nonpoisonous false corals. This form of mimicry, in which a mimic is more strongly protected than the model, has been called Mertensian mimicry.

Coral snake mimicry

WARNING SYSTEMS

The chemical basis for repulsion. Many plants are characterized by the production of unusually large num-

The origin of repellant chemicals

bers of metabolic waste products, often called secondary metabolites, that do not play a part in growth. Because plants do not have the excretory systems that would make it possible to eliminate such substances entirely, secondary metabolites—complex chemicals that include alkaloids, terpenes, phenylpropanes, resins, lignins, saponins, flavonols, and anthocyanins—are stored in the plant tissues. Many such substances are also found in animals that feed upon such plants, usually as a result of the inability of the animal's excretory system to get rid of the plant substances. Many animals produce substances similar to the plant secondary metabolites and store them in glandular pockets (*e.g.*, as in toads, salamanders, and some insects) or in musk glands (beaver, muskrat). Arthropods, particularly insects, are notable for the production of excretory substances that serve as means of defense. Millipedes of the family Glomeridae, for example, secrete a bitter substance (a quinazoline) that repels birds; similar substances, differing only slightly in molecular structure, are found in palms. The fact that a certain chemical substance is restricted to a specific function, such as sex attraction, does not necessarily mean that it was evolved from scratch for that purpose. It seems rather that natural selection follows the easiest course and makes use of substances already present, and sometimes widely distributed. If so, the appearance of such substances in other organisms is not too surprising.

Cardiac glycosides in milkweeds

Among the chemical compounds that protect certain plants from insects or other animals that might feed on them are the cardenolides, or cardiac glycosides. These substances have a highly specific toxic effect on the vertebrate heart and also activate the nerve centre in the brain that causes vomiting. Because the amount necessary to cause vomiting is about half the amount necessary to cause death through heart failure, an animal that samples a plant containing cardenolides is not killed but survives with the knowledge that the plant is inedible. Certain milkweeds (*Asclepias*) that contain cardenolides are the primary food of the larvae of danaine butterflies, including the familiar monarch and queen butterflies (*Danaus plexippus* and *D. gilippus*). The larvae consume the poison without ill effects and retain it through the pupal stage to adulthood. As adult butterflies, they enjoy protection from vertebrate predators.

There is, of course, no such thing as complete protection. Just as danaine larvae are able to eat the protected milkweeds, some predators are able to prey upon the protected butterflies. Birds of the Old World bee-eater family (Meropidae) specialize on bees, which are protected by their stings from many predators. Bee-eaters and a few other birds are able to eat bees because the horny beak protects them from being stung while the insect is being killed and because they have evolved behavioral mechanisms for removing the stinger (usually by wiping the insect on a perch) before swallowing the prey. Rabbits are able to eat the extremely poisonous mushrooms of the genus *Amanita* without ill effects. The larvae of the Florida feather moth (*Trichoptilus parvulus*) consume the insect-trapping glands on the leaves of the sundew (*Drosera*).

The evolution of warning systems. *The selective advantage of warning.* When an organism possesses a mechanism that provides protection from predators, there is a further advantage in preventing the potential predator from even sampling the protected organism. By the act of learning of the danger, the predator may well kill or maim the individual if, for instance, the protected species must be tasted for its inedibility to become known. Many protected insects are provided with tougher skins than their unprotected relatives, but the sampling by a vertebrate predator is almost sure to do some damage. Many noxious organisms have evolved warning (aposematic) mechanisms that serve to identify them clearly to a predator who has had prior experience with the same or similar species.

The value of not being tested

Warning systems often rely primarily on bright colours, but these may be supplemented by olfactory, acoustic, or behavioral means. The New World skunks, for example, have a prominent black and white pattern that renders them clearly recognizable to potential nocturnal predators. When threatened, skunks perform a highly stylized display dance, thus ensuring that the predator will see and recognize the warning coloration.

Acoustic warning signals are often favoured over visual ones because they allow the animal the option of remaining hidden. The rattlesnakes (*Crotalus* and relatives), which need protective coloration to avoid alerting their prey, are able to provide acoustic warning to large animals that threaten them. Many moths of the families Arctiidae and Ctenuchidae are bad tasting, but would be vulnerable to nocturnal predation by bats were it not for the emission of series of high-pitched clicks, audible to bats, made when the moths hear the bats' own ultrasonic navigational pulses. That the moth clicks actually do serve as warnings is borne out by the fact that captive bats ignore thrown mealworms (which they normally relish) when the mealworms are accompanied by recorded moth clicks. Several species of edible moths also produce clicks and may be regarded as Batesian mimics of the unpalatable species.

The role of the receiver. In some cases, the animal who serves as the receiver of the warning signal reacts by means of an innate system that exists independently of experience. Generally, however, a predator must learn the significance of the warning signal through experience. If the predator is a slow learner, or if the warning signal is not sufficiently distinct to avoid confusion with beneficial sensory impressions that the predator receives, several experiences may be necessary. Natural selection, therefore, will favour warning systems that are devoid of ambiguity. Experimentation has shown that certain birds and mammals, at least, are capable of acquiring and retaining knowledge of some aposematic mechanisms from a single experience.

Combination of warning systems with concealing coloration. It is of obvious advantage for an aposematic organism to be able to control the display of the warning system, partly to minimize the amount of sampling, with its concomitant liability of injury, by naïve receivers. Acoustic and chemical warning systems allow this. Many protected animals are coloured to match their backgrounds but provided with flash areas of warning coloration. Examples of these organisms are the tiger moths (certain of the Arctiidae), in which the hind wings are yellow or orange but are kept under the streaked brown fore wings until the moth is molested.

THE OCCURRENCE OF MIMICRY AMONG PLANTS AND ANIMALS

Batesian mimicry. The stinging Hymenoptera (particularly the bees, wasps, and hornets), well protected from most predators and usually equipped with conspicuous warning coloration, are mimicked by insects of many other orders. Ladybird beetles (Coccinellidae) and leaf beetles (Chrysomelidae) are inedible and are provided with prominent colours and usually with contrasting spots. A whole group of Philippine roaches of the genus *Prosoplecta* mimics these beetles, having undergone profound modification in accomplishing the similarity. To simulate the short, rounded form of the ladybirds, the large hind wings of the roaches are rolled and folded in a manner unparalleled in other insects.

Mimicry in the swallowtail butterfly

The order Lepidoptera abounds with Batesian mimics, the best known of which is a swallowtail butterfly, *Papilio dardanus*, a widespread African species. In many populations of this species the females are polymorphic; *i.e.*, a number of different types (morphs) of coloration are found, with each morph a mimic of a species of inedible butterfly of another genus (either *Danaus* or *Amauris*). In all populations, the males are nonmimetic, retaining the same yellow and black pattern throughout. The presence of polymorphism, coupled with the ability of the lepidopterist to breed and rear this species in the laboratory, makes this an apt species for the study of colour inheritance. Investigators have found that mimicry in *Papilio dardanus* depends upon the action of certain primary genes, the expression of which is switched on or off by modifier genes. The modifier genes reduce the number

of possible morphs to the restricted number of mimetic forms. The effects of modifier genes are not carried to the offspring when members of different geographical races are crossed. This finding suggests that each set of modifier genes is adapted to the gene complex in which it normally occurs and in which it probably evolved.

Müllerian mimicry. Müllerian mimicry often occurs in groups of unrelated species, all noxious or inedible and all possessing the same conspicuous warning coloration. Such groups, called mimicry rings, often have associated Batesian mimics. It is not always easy to evaluate the palatability of members of such rings, and thus to distinguish Müllerian from Batesian mimics. Parallel Müllerian mimicry rings are known from South Africa, Borneo, and the tropical Americas; each contains such unrelated insects as malacodermoid and longicorn beetles, butterflies, true bugs, and spider wasps. In South America inedible butterflies of many distinct nymphalid subfamilies (Danainae, Ithomiinae, Acraeinae, and Heliconiinae) share the same warning coloration. Certain species show a highly perplexing divergence from the usual mimicry principles, however. It is axiomatic that maximum protection is gained by Müllerian mimics when all individuals employ the same signal, a principle known as signal standardization. Two species of *Heliconius* (*H. melpomene* and *H. erato*) are polymorphic, however, with each morph in one species duplicated by one in the other and with the morphs of each pair having virtually contiguous geographical ranges. The Müllerian principle appears to be working here between species but not within them.

Mimicry rings

Aggressive mimicry. Examples of aggressive mimicry are abundant and varied; each has its own particular ramifications of basic mimicry principles. The examples cited below illustrate a few of the remarkable extremes in the evolution of mimicry.

Parasitic worms. The flukes (Trematoda) are a class of parasitic worms belonging to the phylum Platyhelminthes. One species, *Leucochloridium macrostomum*, resides principally in the intestine of songbirds. The eggs of the parasite pass to the outside in the feces of the birds, and are readily ingested by a terrestrial snail, *Succinea*, an inhabitant of waterlogged meadows and riverbanks. The parasite eggs hatch into the first larval form within the snail. The next stage, called the sporocyst, is strikingly green in colour and bears yellow-brown rings. The sporocyst develops in the snail tissues and carries several sacs of "spores," one of which is placed into each of the snail's tentacles, or eyestalks. The sac then begins to pulsate violently, at about 40 to 70 beats per minute. The tentacle of the snail becomes greatly enlarged and eventually is transformed into a transparent covering over the pulsating sporocyst. *Succinea* usually avoids light, but specimens with this parasite do not. When the snail appears with its conspicuous, pulsing eyestalks, birds mistake the eyestalks for insect larvae, bite them off, and eat them. Within the bird, the sporocyst then hatches into the final larval stage, which grows into an adult worm. In the meantime the snail's eyestalk regenerates, and the cycle is repeated when another sac passes into the new eyestalks. Because the sporocysts of other trematodes are neither brightly coloured nor mobile, it can be concluded that the colour and pulsation of *Leucochloridium* are adaptations for arousing the interest of the insectivorous birds. Under normal circumstances the host birds do not eat snails, so the sporocyst must imitate the bird's proper food, in order to be eaten and to complete its life cycle in the bird host. The process represents an unusual case of aggressive mimicry, for the sporocyst within the snail eyestalk deceives the bird into infecting itself with the parasite.

Alteration of the snail host

Another trematode, *Cercaria mirabilis*, is notable in that a different larvae form, called a cercaria, is particularly large. The appearance of this cercaria and its hopping mode of locomotion cause it to resemble a small, swimming crustacean or mosquito larva, with the result that fish mistake it for food and swallow it. Research on parasites of this kind is much easier when it is recognized that the larval stages often mimic the food of their respective hosts. Examination of the parasite often provides a suggestion as to the probable host.

Insect-luring plants. Newly hatched flesh flies (*Sarcophaga*), blow flies (*Calliphora*) and greenbottle flies (*Lucilia*) are attracted to glistening droplets or imitations of droplets. The grass of parnassus (*Parnassia palustris*) has flowers with five nectar petals, which bear glistening buttons, but no nectar. They attract flies, nevertheless, and reward them with nectar in two depressions on the upper surfaces of the petals. The insectivorous sundew (*Drosera*), on the other hand, presents a deceptive lure, consisting of glistening secretory droplets on the glandular leaves, which trap insects that are then dissolved with digestive juices. The pitcher plants (*Nepenthes, Darlingtonia, Sarracenia, Cephalotus*) have jug-like leaves, which may bear flowerlike markings near their openings. Some have a flap or hood that enhances the resemblance to a flower and prevents filling with rainwater. One form, *Nepenthes*, secretes nectar at the lip of the pitcher. A foraging insect landing on this apparent flower slips on the edge and falls in. A band of gland cells below the slippery region secretes an enzyme that digests protein. The lower part of the pitcher contains a watery mixture containing digestive fluid and insect remains.

Pitcher plants

From W. Wickler, *Mimicry in Plants and Animals* (© 1968); used by permission of McGraw-Hill Book Co., Inc.

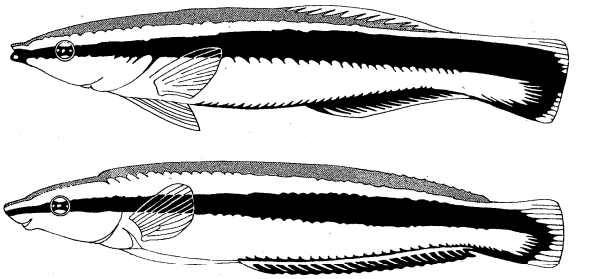


Figure 1: Differences and similarities between (top) the cleaner wrasse (*Labroides dimidiatus*) and (bottom) its mimic, the sabre-toothed blenny (*Aspidontus taeniatus*).

The cleaner mimic. One of the few cases of mimicry reported among vertebrates is that of a so-called cleaner fish. This example involves a particularly close imitation of a model involving shape, coloration, and behaviour. The model, a wrasse *Labroides dimidiatus*, of the Indo-Pacific Ocean, is known as a cleaner fish because it removes and eats externally attached parasites and, occasionally, damaged skin fragments from other marine fish. It occupies specific sites, or territories, on coral reefs, where, within a six-hour period, the individual cleaner may be visited by up to 300 other fish seeking its services. The other fish are attracted by the conspicuous black and white coloration of the cleaner and by its dancelike swimming pattern, in which the tail fin is spread and the posterior part of the fish oscillates up and down. The fish undergoing cleaning acts as though it were in a trance, while the cleaner fish cleans its body, including the inside of the mouth and gills. Even large predatory fish allow themselves to be cleaned, and the much smaller cleaner almost invariably emerges uninjured from their throats. It is quite apparent that the cleaners are protected from these predators although neither inedible nor capable of self-defense.

At the cleaning stations of the cleaner fish, there is often found quite another fish, the sabre-toothed blenny, *Aspidontus taeniatus*. It is similar to the cleaner fish in size, coloration, and swimming behaviour, and it even exhibits the same dance as the cleaner. Fish that have had experience with the cleaner position themselves unsuspectingly in front of this mimic, which approaches carefully and bites off a semicircular piece of fin from the victim and eats it. After having been repeatedly bitten in this way, fish become distrustful even toward genuine cleaners. Observations in the wild indicate that younger fish are the principal victims of the mimic, whereas older fish avoid it whenever possible. It is unlikely that the ability to discriminate between the mimic and the model develops automatically with age and is independent of experience; and this is borne out by the finding that adult fish kept in an aquarium and not previously exposed to the mimic

Mimicry by the sabre-toothed blenny

confuse cleaner and mimic just as younger fish do. If such adult fish are kept with the mimics for a certain length of time, however, they eventually avoid these and the genuine cleaners. The obvious conclusion from these experiments is that other fish cannot distinguish between the model and the mimic without having had experience with both. Evidently victims of the mimic seek out and learn characteristics that enable them to distinguish between reliable and unreliable cleaners—that is, between the model and mimic, respectively. The most successful individuals among the mimics, therefore, are those that most confuse their victims. As a result, the further development of the mimic is steered in the direction determined by the characteristics of the cleaner; for example, the model occurs as a number of local races within its area of distribution, each of which shows its own peculiarities in coloration, such as a small or large black vertical stripe at the base of the pectoral fins or an orange-red spot on the flanks. In every case, the local population of the mimic shows the same special coloration as does the model in that particular area.

One of the interesting and highly unusual aspects of the cleaner-mimic relationship is that the individual characters of the mimicry pattern, especially the behavioral ones, have been traced to their origins. Certain chance similarities include body size and shape, as well as swimming pattern. The drive in the mimic to approach other fish, for example, is a specialization of a more general pattern, observable in nonmimicking relatives of the blenny, involving searching for food on suitable surfaces. The basic colour pattern of light and dark horizontal stripes is characteristic of fish that swim in open water, but the actual coloration of the mimicking blenny has been selected for closer resemblance to the cleaner wrasse. Interestingly enough, the blenny alters its coloration with its motivational state and adopts the appearance of the cleaner only under the specific conditions of self-confidence and intent to attack. As a group, blennies tend to wriggle while swimming, to counteract a strong tendency to sink at the tail. The sabre-toothed blenny, however, when confronted with danger holds the body stiffly without wriggling, allowing the hindquarters to sink somewhat, and advances solely by the use of the pectoral fins, in the manner common to wrasses. Superimposed on this motion is a nodding of the blenny's head, typical of approach-retreat conflict behaviour in blennies, but resulting in a stimulation of the swaying dance of the cleaner. The combination of coloration and behavioral signals has a particular significance for the experienced visitor to the cleaner's station, and causes it to adopt the posture that invites cleaning. In so doing, the visitor gives the mimic an opportunity to take a bite from a fin.

Fireflies. A form of aggressive mimicry that relies entirely on behaviour occurs in certain North American fireflies (Lampyridae). Males of these familiar night-flying beetles emit light bursts in flight according to highly specific patterns. The females, usually stationary, respond to the flash patterns of males of their own species with specific patterns of their own. The flying male responds to the appropriate female signal by approaching, landing, and courting. Most adult fireflies are short-lived and do not feed at all, but females of the genus *Photuris* have been found to feed on other beetles, including males of the genus *Photinus.* Upon perceiving a flashing male *Photinus,* the female *Photuris* responds with a flash that mimics the female *Photinus.* As the male *Photinus* approaches, the female *Photuris* even reduces the intensity of her flashes, to resemble more closely the weaker signals of the smaller female *Photinus.* The hapless male, after landing, is seized and eaten by the *Photuris.* In response to males of her own species, of course, the female *Photuris* gives a flash response quite different from that of *Photinus.*

Host mimicry by parasites. Another form of mimicry, sometimes considered an extension of aggressive mimicry, is mimicry of a host by its parasite. Most of the best known examples occur among birds and represent some of the few known instances of mimicry in that class of animals.

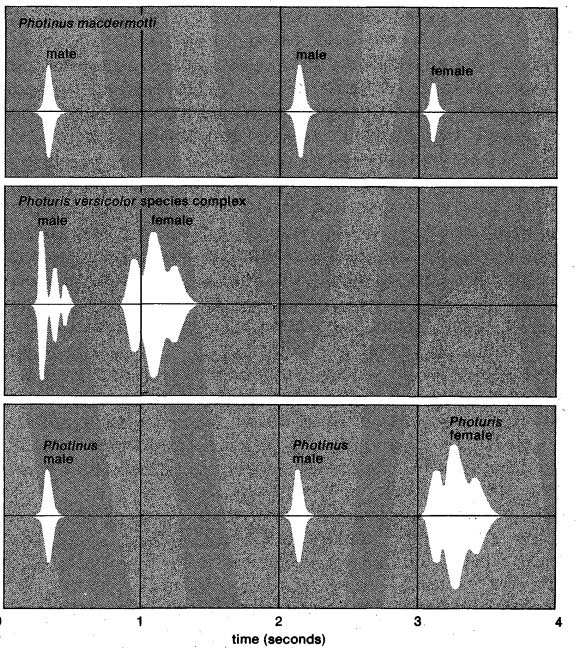


Figure 2: Comparative pulse rates of two genera of fireflies, *Photinus* and *Photuris,* showing (bottom) mimetic response of *Photuris* female to *Photinus* male.
By courtesy of James E. Lloyd

Cuckoos. The European cuckoo (*Cuculus canorus*) is a brood parasite—*i.e.,* it lays its eggs in the nests of other birds, which act as foster parents for the young cuckoos. The most frequent foster parents are various species of small songbirds. Although the eggs of the various host species span a great range of colours and spotting, there is a striking correspondence in appearance between the eggs of the cuckoo and those of the host. A female cuckoo normally deposits only one egg in each host nest and removes one of the hosts' eggs, leaving the same number of eggs as were previously there. Most small birds react unfavourably if they perceive a foreign egg in the nest and either abandon the nest, build another nest right over the first, or eject the strange egg. It is known that each female cuckoo consistently lays eggs of one colour pattern and must therefore parasitize only a few host species. Some cuckoos do lay in the nests of the wrong hosts occasionally, because of inability to find the correct nest, and most of such eggs are probably removed by the hosts, providing strong selection tending to eliminate the defects. One survey has shown that of 1,642 cuckoo eggs laid in the nests of the correct (matching) hosts, only 8 percent were lost, whereas of 298 in the nests of the wrong hosts, 24 percent were lost. It is logical to conclude, therefore, that a cuckoo that lays its egg randomly, leaving the survival of the eggs to chance, would produce fewer offspring than one that selects hosts whose eggs match her own. Although the control of egg coloration is probably genetically determined, the choice of correct hosts is believed to be the result of a learning process that takes place when the female cuckoo is a nestling and learns to recognize its own foster parents.

Egg matching by cuckoos

Parasitic weaverbirds. Brood parasitism is also found in African whydahs, or widow birds, the subfamily Viduinae of the weaverbird family, Ploceidae. Each species of whydah parasitizes a single species of estrildid finch (Estrildidae). In this case, egg colour does not seem to be a factor in acceptance of the parasite's egg, because both groups have pure white eggs. It has been argued that the whydah, many relatives of which have spotted eggs, have evolved white eggs to match those of their estrildid hosts. More significant, however, is the highly specific pattern of spots and protuberances at the corners of the mouth (gape) and on the palate, tongue, and lower mandible of the nestling. This pattern, which varies from one species of finch to another, serves as a releaser for feeding behaviour on the part of the parents, which ignore any nestling that does not display the proper pattern for the particular

Mimicry of mouth patterns

species. In every species of parasitic weaver studied, the nestling has been found to match perfectly the mouth pattern of the estrildid host. In addition to mimicking the mouth patterns of their hosts, whydah nestlings also duplicate the specific begging calls and peculiar head movements of their hosts. The coloration of the juvenile plumage of the young whydah is identical to that of the host species, ensuring that the whydah will be fed after fledging. The digestive system of the young whydah is closely adapted to the particular type of food utilized by its host species, unlike that of the young cuckoo, which seems to be able to accept a variety of foods, from insects to mouse meat.

With each species of parasitic weaver closely committed to a single species of estrildid finch, it is obviously important that whydah species not hybridize, for the hybrid offspring would certainly not match either possible host in all of the important features. It is surprising to find seven forms of the paradise whydah (*Steganura*) so similar in appearance that they were once considered races of one species. Each of the seven, however, has its own estrildid host species, indicating that seven species of paradise whydah are represented. As is frequently the case with closely similar bird species, hybridization is effectively prevented through the use of species-specific vocalizations by the males. An unusual feature of this situation is that each widow bird species uses the same vocal pattern as its estrildine host. The young parasitic weavers apparently learn the songs of their host species during the critical learning period common to songbirds generally. As adults, the male whydahs use these estrildid vocalizations and gain response only from females that have been reared by the same foster species. This example is the only known one of a species-isolating mechanism consisting of vocalizations learned from another species.

Mimicry to effect pollination and dispersal. In some instances, plants have been found to rely on mimicry to attract insects as aids in pollination or in the dissemination of seeds or spores.

Orchids. Many flowering plants lure insects through the use of bright colours that indicate the presence of nectar. Some orchids mimic other flowering plants without offering any nectar, relying on those that do provide nectar to reward the nectar seekers.

Flowers resembling female insects

A group of orchids, often known by descriptive names such as fly orchid, bee orchid, and spider orchid, carry the deception further, actually mimicking the insects themselves. The best known orchids of this type are members of the genus *Ophrys*, the flowers of which have interested man for centuries. The labellum of the *Ophrys* flower is a specialized median petal that acts as a dummy female of a species of bee or wasp (depending on the species of *Ophrys*), the resemblance being so close that males visit the flower in an attempt to copulate with the dummy female. In the course of precopulatory and copulatory movements, the visiting insect acquires the pollen sacs (pollinia) of the orchid and subsequently transmits them to other blossoms. A similar situation occurs in an Australian orchid, *Cryptostylis leptochila*, which bears a sufficient resemblance to the female of the ichneumon wasp *Lissopimpla semipunctata* to induce copulation by the male wasp.

An important feature in the mimicry of female insects by orchids is flower size. Too small flowers do not provide adequate stimulus for the copulatory attempts necessary to plant the pollinia on the insect. If the flower is too large for the insect, the insect's head does not reach the stigma of the flower (the female receptor site for pollen), and the pollinia are not deposited.

The colour of the labellum is also important in attracting and properly orienting the male insect. The males are more attracted to dark than to light colours and to a contrasting dark spot on a light background. A velvety surface is a more effective attractant than a shiny one. The centre of the labellum of *Ophrys insectifera* bears a dark red spot, almost black, whereas the lateral lobes of the labellum are a somewhat lighter purple-red. Female wasps of the genus *Gorytes*, males of which are highly attentive to this orchid, are black dorsally with dark purple wings. Where the folded wings overlap over the female's body, there is a glistening area closely resembling the shiny central spot on the flower. The overall effect is further enhanced by movements of the flower in the wind.

Odour plays a particularly strong role in attracting male insects to the mimicking orchids. Some female Hymenoptera secrete odoriferous substances (pheromones) that initiate search behaviour in the males and guide them to the stationary females. Such attractant chemicals are usually limited in effectiveness to the producing species or a few close relatives. *Ophrys* flowers give off odours similar to, if not identical with, those produced by females of their associated insects. In a few cases the odour of the flower is a more potent attractant for the male than that of the appropriate female; the orchid not only mimics the sex attractant, but improves upon it.

Carrion flowers, stinkhorn mushrooms, and mosses. A group of flowers are able to attract dung beetles (certain of the Scarabaeidae) and carrion flies (Calliphoridae) by mimicking the odours of dung or rotting flesh used by these insects as guides to sites for egg deposition. The deception is so complete in some carrion flowers (*e.g.*, *Stapelia*) that blow flies actually lay their eggs in the flowers. The cuckoopint (*Arum maculatum*), by a metabolic level unequalled among plants, spreads its odour over a wide area by an elevation of temperature that increases the vaporization rate of the volatile odour substance. An elaborate mechanism in the cuckoopint ensures that a pollen-laden visitor remains long enough to deposit the pollen. The sheath of the floral structure, upon which the insect lands, is made slippery by oil droplets with the result that the insect slides down into a cup equipped with a ring of spines to prevent escape. In trying to climb out, the insect deposits pollen on the tiny female flowers, from which the insect receives nectar. During the night, the male flowers mature and cover the resting insect with pollen. Then the spines shrink, and the insect is released. The production of the attractant odour occurs at midday, when many carrion-seeking insects are active, on the day before the male flowers mature; the timing of odour production is controlled by a substance produced by the male flowers six to 18 hours before maturity.

The insect trap of the cuckoopint

A similar situation is found in stinkhorn mushrooms of the genus *Phallus*, found in woodlands and meadows of the Northern Hemisphere. The cap of the young stinkhorn is covered with a thick, greenish-black, shiny layer of gelatinous spore slime (gleba), which is eaten by blowflies and other insects attracted by the carrion-like odour. The spores pass through the digestive tracts of the insects and are voided with the feces, thus ensuring the dispersal of the mushroom.

Some mosses (*e.g.*, some members of the genus *Splanchnum*) have flowerlike structures that attract flies to aid in spore dispersal. Insects are attracted by the mimic of a nectar-bearing true flower and by a carrion-like odour.

Mimicry within species. The three essential participants in mimicry, model, mimic, and receiver, need not always be members of different species. In mimicry of the host by a parasite, for instance, the host species provides both model and receiver. In another type of mimicry the mimic and receiver are members of the same species. An example of this type is found in the small South American characoid fish *Corynopoma riisei*, in which the gill cover of the male is elongated into a thin whitish stalk that terminates in a small blackish plate. During courtship, the male raises the stalk and waves it jerkily in view of the female, who mistakes the tip for an edible object, such as a tiny crustacean. As she nears the male in her attempts to grasp this supposed prey, mating takes place.

Egg mimicry in cichlid fishes

Another remarkable form of mimicry within the same species occurs in the African mouthbrooding cichlid fish of the genus *Haplochromis*. The female takes the eggs into her mouth immediately after they are laid, even before the male can fertilize them. The male, however, carries conspicuous yellow or orange spots near the base of the anal fin, which closely resemble the eggs of the particular species. Although the female is inhibited from eating while carrying eggs, she is strongly motivated to

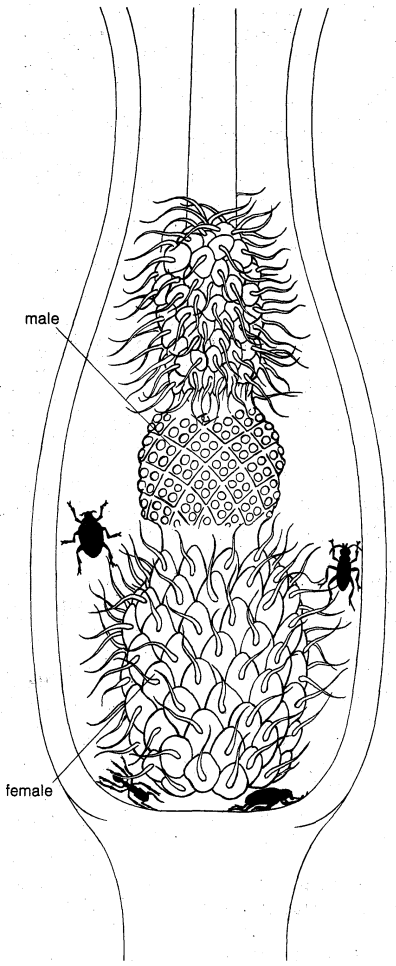


Figure 3: Flowers of the cuckoopint (*Arum maculatum*). The odour emitted by the flowers attracts insects that deposit pollen (see text).
From W. Wickler, *Mimicry in Plants and Animals* (© 1968); used by permission of McGraw-Hill Book Co., Inc.

pick up loose eggs in her mouth. The male displays the fin spots to the female while releasing sperm; the female, as she attempts to pick up the false eggs, takes in sperm that fertilize the eggs in her mouth. In this case the model (real eggs), mimic (false eggs), and receiver (adult female) are all of the same species.

THE EVOLUTION OF MIMICRY

The effectiveness of warning systems. There is considerable experimental evidence to illustrate how effectively predators learn to avoid certain adverse stimuli. Chickens conditioned by electric shock to avoid drinking dark-green water drank progressively more from paler solutions in proportion to the intensity of the colour. This experiment suggests that even an incomplete warning system provides a modicum of protection. The degree of protection provided is also affected by the strength of the punishment; after strong shocks the chickens drank only from very light coloured solutions. In the presence of severe punishment, an improved warning system made little additional effect once a threshold level was reached.

In other experiments, starlings (*Sturnus vulgaris*) were fed normal meal worms, two segments of which had been painted orange. To provide aposematic "models" the experimenter made other meal worms distasteful and painted the same segments green. "Mimics" were marked with green but not rendered unpalatable. There is no known instance in nature in which animals employ green for warning; there was therefore no possibility that the birds had already learned to avoid the experimental colour pattern. Before long the green marked worms were completely avoided, regardless of palatability, even when the ratio of edible to distasteful was 60:40. This indicates that the number of mimics can exceed that of the model, when the resemblance is close, without loss of protection. When the ratio was increased to 90:10, 17 percent of the mimics were avoided, probably sufficient to a selective advantage in nature. Although a test bird would occasionally peck at a model, then reject it, the same action was sometimes shown to a mimic that it had picked up, suggesting that a premature response had been altered by subsequent consideration.

The reconstruction of evolutionary pathways. Analysis and understanding of a given mimicry system require a rather comprehensive knowledge of morphology, behaviour, ecology, and mutual relationships of animals usually in different classes—for example, wasps (Hymenoptera), flies (Diptera), insect-eating amphibians, reptiles, birds, and small mammals. Tracing the evolution of such a complicated system requires a detailed acquaintance with a large group of forms related to each of the animals involved. Such data, in fact, are seldom available.

Reconstructing the evolution of a case of mimicry within the same species, however, is relatively simple, requiring detailed knowledge of but one rather narrow taxonomic unit. Such a reconstruction is valuable, because mimicry is an indispensable tool in the study of the evolution of animal communication, and usually starts from conspicuously elaborated signals, which postulate a signal receiver interested in them. The receiver practically always has undergone a special molding toward optimal receiving of the signal. The mutual adaptations of the sender and the receiver must be examined separately. This examination is easily made, so far as the evolution of a reaction or of a receiving mechanism is concerned, in all predators trying to find their prey and in all prey animals attemping to escape an approaching predator. The suppression of signals may be studied in predators trying to sneak up on a prey unnoticed. The elaboration of a signal, which must, of course, be important to the receiver, can only be studied after consideration of compensatory adaptations in the receiver and in situations where the sender has a one-sided interest in the signal. The deceiving signal can be derived only from one of two types: a signal developed by the receiver and another signal sender in their common interest or a signal emitted by another signal sender and only made use of by the receiver in its own interest. Both cases, by the definition given above, are called mimicry. An additional advantage is that the model is known to be the final stage toward which the mimic will evolve (so far as the signal characters are concerned), thus indicating a trend in evolution that is still operating and that probably will further elaborate the mimetic signals.

Elaboration of the warning system

If a *Haplochromis* female fish were to discriminate between real eggs and the egg dummies of the male and were to stop reacting toward the latter, her eggs would remain unfertilized. In such cases of deceptive signals developed within the same species natural selection operates against better signal discrimination on the part of the signal receiver.

The importance of the signal receiver. Fundamental characteristics of mimicry are determined mainly by behavioral properties of the signal receiver. A precise knowledge of the identity of the receiver and a thorough study of its behaviour are therefore indispensable for the understanding of mimicry. Moreover, mimicry is not sharply distinct from, but gradually merges into, other sender-receiver communication systems. Palatability is a matter of degree; whole ranges of increasing distastefulness therefore exist, even in the mimics, model and mimic in the extreme case being equally unpalatable and sharing the same warning coloration. This system, called Müllerian mimicry, could be considered not to be true mimicry, after all, because no one is deceived, and it is impossible to designate one as model and the other as mimic.

Although all individuals of a given wasp species look alike and are all equally protected, this phenomenon is not usually called Müllerian mimicry, simply because the signals were not independently, convergently evolved. Because, however, the male wasps have no protective properties but retain their group-specific warning color-

ation, this is Batesian mimicry, although model and mimic are of the same species and their signals homologous (evolved from the same source). Convergence of the signal characters, therefore, is essential only for the so-called Müllerian mimicry, and thus Müllerian mimicry is distinguished from other cases of signal standardization. The typical (Batesian) mimicry merges into Müllerian mimicry if the difference between the consequences for the receiver of reacting similarity to model and mimic diminishes; and by homology of the signal characters it further merges into general signal standardization.

Insects may be protectively coloured to resemble a wasp or a twig, the first case being called mimicry, the latter mimesis, or protective coloration. The difference again lies within the signal receiver. If the mimetic signal does not release any reaction in the receiver, it falls under mimesis. This distinction is illustrated by the Dutch biologist L. de Ruiter's experiments with stick caterpillars, which, by virtue of their close resemblance to twigs, are protected against insect-eating birds. As soon as the number of caterpillars becomes too large, however, the bird develops an interest in them, attacks real twigs, and also finds some caterpillars. If one positive experience with the caterpillar has the same weight as a negative one with the twig (the signal remaining unchanged), the relative abundance of caterpillars and twigs determines whether all twigs are mistakenly exterminated or whether the feeding reaction toward twiglike objects disappears, thus protecting the caterpillars.

This study again illustrates the importance of the bird's ability to decide correctly which is the model and further shows how easily an object (the twig) may quite involuntarily become a "mimic." Another example illustrating the importance of a correct model is found in the common farming relationship between ants and aphids. The protuberances, called the siphones and cauda, on the abdomens of aphids resemble respectively the bases of the antennae and labium of the ant's head. The aphid's abdomen is thus mistaken by the ant for the head of a fellow ant, thereby eliciting the food-begging response, which is identical with milking. Saturated ants in turn even try to feed the abdomens of the aphids. Aphid species with reduced abdominal siphones use their hind legs as antennae dummies, the movements elicited being originally defensive movements. This situation is exactly the way in which mimicry arises. Mimetic characters need not have evolved under the selection pressure of mimicking; in fact, their earliest evolutionary stages could not even have been brought about in this way. All cases studied thus far can be traced back to an incipient stage of deceptive resemblance, initiated as a preadaptive, nondirected by-product of pre-existing species-specific features, thus providing a point of attack for new selective pressure.

The effects of selective pressure. The selective consequences for the signal receiver of responding to the model are always positive (the reaction would disappear if, on balance, it were unfavourable to the receiver). The mimic always has a selective advantage in releasing the reaction from the receiver. An unfavourable signal by the mimic would also disappear by natural selection.

The selective consequences of signalling by the model

The selective consequence for the model eliciting and obtaining the reaction from the receiver may be of several types. Consequences may be absent, if the model is an inanimate object, on which natural selection does not act. They may be negative, if the model is nonaposematic (nonwarning), such as the tiny crustacean, usually eaten by the signal receiver and mimicked by the male *Corynopoma* characin in order to attract the female. They may instead be positive, as in the wasp, which remains alive if it is avoided by the predator; in the cleaner, which feeds on parasites harmful to other fish; or in those hymenopteran females whose male-attracting signals are mimicked by certain orchids. Mutual interest is present between model and receiver in cases of aggressive mimicry where both belong to the same species and in typical Batesian mimicry.

Constant learning by the signal receiver results in a strong selective pressure on the mimic against detectable differences from the model, but it also exerts a strong pressure on the model to develop new differences from the mimic. Typically, it is the group of songbirds parasitized by cuckoos that has developed the most divergent egg colour patterns; the group of estrildine finches, parasitized by widow birds, that has developed particular gape patterns; and among the cleaner wrasses the species *Labroides dimidiatus*, mimicked by the blenny *Aspidontus*, that develops into many different local races.

There is a boomerang effect, characteristic of parasite-host relationship—the more successfully a bird rears young cuckoos, the more certain it is that it will lose its own young, because they are killed off by the young cuckoo. Parasites that are too successful, therefore, harm themselves, for each female cuckoo needs several nests of the same host species for her eggs. In an area containing particularly successful cuckoos, the number of reed warbler nests has been found to decrease from year to year, while the percentage of nests parasitized by cuckoos increases from year to year. This ratio means that a cuckoo too well adapted reduces the availability of its own hosts, while one insufficiently adapted kills off its own offspring. Presumably, selection in both directions produces a continual oscillation in the densities of hosts and cuckoos. A similar dilemma is inflicted on man, who acts as a predator against weeds in crop fields and by winnowing selects the wanted seeds from the usually smaller weed seeds. The flax dodder (*Cuscuta epilinum*), growing as a creeper around flax and linseed plants and damaging them, originally had small seeds that could be easily separated from the larger flax seeds. A special mutant of the flax dodder is now cultivated and spread by man, despite being against his interests. By a mutation that produced twin seeds, the dodder has evolved the capability of being separated out and planted with the desirable flax seeds. In this case, the parasite mimics the protected plant, receiving the same protection.

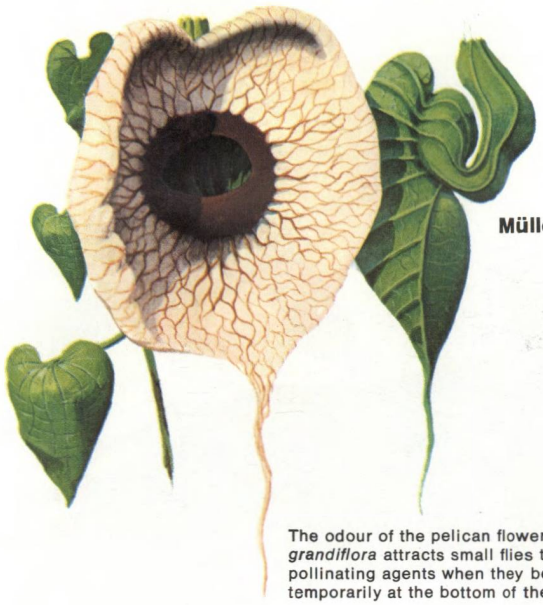
The effects of geographical distribution and population density. It has been postulated that the model and mimic should occur in the same area—*i.e.*, be sympatric. They need not always be sympatric, however, but must only have a signal receiver in common: a model might be in Africa and its mimic in Europe (or vice versa), functionally connected by a migratory bird. Another postulate, that mimics must be less numerous than their models, means, correctly stated, that the receiver has to meet the mimic less often than the model; this postulate is based on the assumption that one experience with the model has the same aftereffect, the same weight, as has one with the mimic. This assumption, however, has been proven not always to be so; in fact, the negative experience seems usually to be the stronger one. This negative experience may result from an encounter with the model (a wasp) or with the mimic (for example, the *Aspidontus*). There might very well be more wasp mimics than wasps, but, in the case of the cleaner mimic, it is probable that the mimic in fact has to be less numerous than the model. The protective power of the model, of course, which provides the advantage of being its mimic, is reduced with an increasing number of mimics, because the predator will eat larger numbers of them before his first encounter with the model.

Relative populations of model and mimic

The importance of mimicry to evolutionary theory. The mimicry hypothesis emerged in the middle of the Darwinian controversy and provided an ideal test case for the views of Darwin and his contemporary Alfred Russel Wallace on the operation of natural selection in the evolutionary change of living organisms. It is now quite evident that the basic theory of natural selection is correct, and that the theory is strengthened by many detailed studies of the process by which a mimetic resemblance is brought about and selected for. Investigating suitable cases of mimicry provides important insight into the evolution of signals and the process by which signals get their meaning.

BIBLIOGRAPHY. L.P. BROWER, "Ecological Chemistry," *Scient. Am.*, 220:22–29 (1969), a discussion of relative palatability, experiments on automimicry; C.A. CLARKE, P.M. SHEPPARD, and I.W.B. THORNTON, "The Genetics of the Mimetic Butterfly *Papilio memnon L.*," *Phil. Trans. R. Soc.*,

Müllerian and other forms of mimicry



The odour of the pelican flower *Aristolochia grandiflora* attracts small flies that act as pollinating agents when they become trapped temporarily at the bottom of the flower.

The labellum (lip) of the flower of the orchid *Ophrys apifera* imitates a female bee; males attempting to copulate with the flower act as pollinators.

Mimicry in coral snakes and similar forms. (Left) venomous Eastern coral snake *Micrurus fulvius*; (centre) harmless king snake *Lampropeltis polyzona*; (right) moderately venomous rear-fanged false coral snake *Erythrolamprus aesculapii*.

Müllerian mimicry in unpalatable New World tropical butterflies.

Melinaea ethra (family Ithomiidae)

Mechanitis lysimnia (family Ithomiidae)

Lycorea halia (family Danaidae)

Heliconius narcaea (family Heliconiidae)

Painting by C. Olsen

The zone-tailed hawk is able to surprise its prey because of its resemblance in appearance and flight pattern to the carrion-feeding turkey vulture.

Aggressive mimicry in birds

Superficial resemblances of cuckoos to the sparrowhawk and shikra frighten other birds from their nests, allowing the cuckoos to deposit their eggs in them.

The parasitic whydah deposits its eggs in the host finch's nest. The young whydah mimics the mouth pattern of the young estrildid finch, thereby triggering feeding behaviour by the estrildid host.

Painting by C. Olsen based on (*Steganura paradisaea*, *Pytilia melba*, nestlings) W. Wickler, *Mimicry in Plants and Animals* (© 1968); used with permission of McGraw-Hill Book Company.

Series B, 254:37–89 (1968), an analysis of genetic polymorphism in a Batesian mimic; H.B. COTT, *Adaptive Coloration in Animals* (1957), a classic work on all aspects of animal coloration; C.J. DUNCAN and P.M. SHEPPARD, "Sensory Discrimination and Its Role in the Evolution of Batesian Mimicry," *Behaviour*, 24:269–282 (1964); D.C. DUNNING, "Warning Sounds of Moths," *Z. Tierpsychol.*, 25:129–138 (1968), concerned with aposematic sounds of model and mimic; T. EISNER and J. MEINWALD, "Defensive Secretions of Arthropods," *Science*, 153:1341–1350 (1966), a study of chemical weapons; B. KULLENBERG, "Studies in *Ophrys* Pollination," *Zool. Bidr. Upps.*, vol. 34 (1961), a monograph on orchid mimicry; B.J.D. MEEUSE, *The Story of Pollination* (1961), a textbook on insect-flower-relationship; L. DE RUITER, "Some Remarks on Problems of the Ecology and Evolution of Mimicry," *Archs. Néerl. Zool.*, suppl. 1, 13:351–368 (1958); J.R.G. TURNER, "Evolution of Complex Polymorphism and Mimicry in Distasteful South American Butterflies," *Proc Int. 12th Congr. Ent.*, sect. 4, p. 297 (1965), concerned with Müllerian mimicry in butterflies; W. WICKLER, *Mimikry: Nachahmung und Täuschung in der Natur* (1968; Eng. trans., *Mimicry in Plants and Animals*, 1968), a monograph on modern views of all aspects of mimicry.

(W.J.H.W.)

Minamoto Yoritomo

Minamoto Yoritomo was the founder, in the 12th century AD, of the shogunate, or *bakufu* (military administration), and of the line of feudal lords, which, through the shogunate, ruled Japan for 700 years. He was of noble and, as a descendant of the emperor Seiwa (reigned 858–876 AD), even royal lineage. His family name, Minamoto, is Genji in Chinese (Gen being the Chinese reading of the symbol for Minamoto), and it is immortalized as the embodiment of ancient courtly ways in *The Tale of Genji* (*Genji monogatari*), one of the world's earliest and greatest novels. But the family's immediate past was military as well as aristocratic, and Yoritomo's personality was impatient with the court's cultured and precious subtleties. He wanted power and was jealous, suspicious, and cold-hearted, even in his own circle. He went as far, in fact, as to liquidate several near relations. But once in power, he proved an excellent administrator.

Aristocratic and military background

By courtesy of the International Society for Educational Information, Tokyo, Inc.

Minamoto Yoritomo, colour on silk cloth attributed to Fujiwara Takanobu (1142–1205). In the collection of Jingō-ji, Kyōto.

Yoritomo was born in 1147, the third son of Minamoto Yoshitomo, who, in 1159, attempted to destroy Taira Kiyomori (scion of another dominant military family, the Taira clan) in the Heiji War, in Kyōto Province. He was defeated, however, and his son Yoritomo was captured and banished to Izu Province (a peninsula southwest of Tokyo), where he lived for 20 years under Taira surveillance.

Yoritomo enlivened his rustication by seducing the daughter of his jailer, Itō Sukechika. The latter's rage forced Yoritomo's flight to the protection of Itō's superior and neighbour, Hōjō Tokimasa, a Taira vassal whose hostile attitude to the Taira clan typified the contemporary split between court and country. Hōjō's daughter also succumbed to Yorimoto's blandishment but had to postpone marriage until 1180, when her official fiancé, the pro-Taira acting governor, had been eliminated. The essential feature, however—an understanding between Tokimasa and Yoritomo—was swiftly completed; and Yoritomo's political pretensions now enjoyed support.

Meanwhile Taira Kiyomori, the head of the Taira clan, exercised his power over the Imperial court, thus alienating Shirakawa II, the retired Emperor. (At this period of Japanese history the emperor often lived in "retirement" away from court so he could rule without the hindrance of the highly detailed court ceremonials. This practice was known as *insei*.) Most of the aristocracy and the heads of the great temples and shrines were also resentful of the Taira clan's hold over the Emperor.

In 1180 Minamoto Yorimasa, another member of the Minamoto clan, joined in a rebellion with an Imperial prince, Mochihito-ō, who summoned the Minamoto clan to arms in various provinces. Yoritomo now used this princely mandate as a justification for his own uprising. Despite Mochihito-ō's death, which occurred shortly before Yoritomo's men were led into battle, he succeeded in gaining much support from the feudal lords in the eastern provinces. Many members of the Taira family also enrolled under Yoritomo's banner, for they were disappointed with their meagre rewards from cousins at court. Yoritomo immediately advanced to Kamakura (about ten miles south of Tokyo) and established his headquarters there. As well as consolidating a hold over his own vassals in the Kantō area (west of Tokyo), Yoritomo tried to organize the Minamoto followers under his direct control. He was loath to relinquish control to any of his various relatives, and to this end he established the Samurai-dokoro (Board of Retainers).

In 1183 Minamoto Yoshinaka, a cousin of Yoritomo, occupied the Hokuriku district and invaded Kyōto, the seat of the court. The emperor Shirakawa II, who always hoped to play off supporters, as well as enemies, against each other to regain some of the substance of Imperial power, invited Yoritomo to put an end to Yoshinaka's dangerously successful career; and Yoritomo accordingly crushed Yoshinaka at Kyōto. Yoritomo now established the Kumonjo (Board of Public Papers) and Monchūjo (Board of Questioning), setting up not only a military but also an independent political government in the east, yet one that was recognized by the central Imperial court in Kyōto. In 1184 Yoritomo's considerable armies, commanded by his two younger brothers, Minamoto and Yoshitsune, a brilliant commander of whom Yoritomo was jealous, were ranged against the Taira forces for what was hoped would be a climactic campaign, but decisive victory was not gained until the following year. After Minamoto's next victory, the Emperor supported Yoshitsune in efforts to restrain Yoritomo's power. But Yoritomo immediately expelled Yoshitsune and imposed on the Emperor the establishment of *shugo* ("constables" or "guardians") and *jitō* ("district stewards") throughout Japan, avowedly to capture Yoshitsune, though such arrangements were instrumental in making Yoritomo's ascendancy nationwide. Soon after, Yoritomo succeeded in having his brother Yoshitsune put to death.

Yoritomo gave his *shugo*, each placed in a province, the function of administering and policing the Minamoto vassals locally. The *shugo* also administered the judicial proceedings in cases of rebellion and murder, and they thus acquired something of a military hold over each province. To supervise individual estates, the more pacific office of *jitō* was created, which levied taxes and undertook the management of the estates. And, not surprisingly, both the *shugo* and the *jitō* became feudal lords. Through these institutions, Yoritomo was thus able to undermine the central government's local administrative power, and subsequently he even made efforts to rule

remote districts, such as Kyushu, Japan's southernmost island.

Supremacist ambitions

In 1185 he destroyed Fujiwara Yasuhira, an independent noble of the Tohoku area, demonstrating his ambition to create a power structure independent of the capital, at Kyōto. In 1192, a few months after his old rival Shirakawa II's death, Yoritomo, now with no one to balk him of his ultimate ambition, titled himself Sei-i Taishogun (meaning "Barbarian-Quelling Generalissimo"), becoming the supreme commander over the feudal lords. The Kamakura shogunate was now formally complete.

After 1192 Yoritomo's policies were designed to relieve the strain between the military lords and the court aristocrats, and the powerful temples and shrines. Thanks to the institutions of the *shugo* and *jitō*, relations between the court of Kyōto and Yoritomo's government at Kamakura were fairly stable. Yoritomo died in 1199.

Yoritomo is often charged with cruelty, particularly in the killing of his cousin and brothers, but political circumstances of his time were difficult. He had above all to prevent discord among his vassals and the whole military class if his work was to attain the permanence that it did. (K.N.)

Minas Gerais

Minas Gerais, an inland state (*estado*) of Brazil, is the storehouse of the mineral riches of that country, as is indicated by its name, which means "general mines." The state is bounded on the north by the states of Goiás and Bahia; on the east by Bahia, Espírito Santo, and Rio de Janeiro; on the south by Rio de Janeiro and São Paulo; and on the west by São Paulo, Mato Grosso, and Goiás. Its area is 226,708 square miles (587,172 square kilometres). Its population (1970) was 11,600,000, being exceeded among Brazilian states only by that of São Paulo. The state capital is Belo Horizonte.

History. Minas Gerais was first explored by Fernão Dias Pais Leme between 1664 and 1677, though he was not the first European to penetrate it. The discovery of gold in 1692 to 1695 by bands of adventurers from the São Paulo settlements led to every occupation and profession being abandoned in the mad rush for the new mines. Minas Gerais at first formed part of the captaincy of São Paulo, but in 1720 it was established as a separate captaincy general and was brought more directly under the Portuguese crown. The arbitrary restrictions imposed upon the colonists aroused dissatisfaction among them and eventually led to a conspiracy in 1789, inspired by a fear that the Portuguese government was about to enforce the collection of its "royal fifth" of the mining output, which had largely fallen into arrears. Among the conspirators was one José Alves Maciel, who had just returned from France, where he had met Thomas Jefferson and had become infected with French revolutionary ideas. A number of residents became involved, and reckless talk in public places led to the arrest of the conspirators. Only one was executed, a poor, uneducated subaltern militia officer, Joaquim José da Silva Xavier, nicknamed O Tiradentes (the Tooth Puller), the others being imprisoned or banished. Tiradentes has since been glorified as a national hero of Brazilian independence. Minas Gerais has traditionally played an important role in Brazilian politics, in part because of its central position but also because of the dense population in its southeast. In 1822 it became a province in the Brazilian Empire, and in 1891 a state in the federal republic. In 1897 the old colonial capital of Ouro Prêto was deprived of its administrative function. The new capital was placed at Belo Horizonte, where a new city, laid out on previously almost unoccupied land, was planned and built. Ouro Prêto became a national monument, and Belo Horizonte has become a city with a population of more than 1,100,000.

The Minas Gerais gold rush

The natural environment. *Relief and drainage.* Minas Gerais lies almost wholly within the geologically ancient, mineral-rich uplands known as the Brazilian Highlands. Only small sections of the southern and eastern parts are within the zone of the Great Escarpment which forms the eastern margin of the highlands. In the southwest, along the border of São Paulo, is the commanding range of mountains known as the Serra da Mantiqueira, whose highest summit, Pico das Agulhas Negras (also known as Pico do Itatiaia), reaches 9,144 feet (2,787 metres). This range ends near the easternmost border of São Paulo. The other principal mountain range is the Serra do Espinhaço, which begins a little south of Belo Horizonte and extends northward into Bahia. It forms the divide between the rivers draining to the São Francisco River and those draining eastward directly to the Atlantic (the Doce, Mucuri, Jequitinhonha). The greater part of the state consists of a much dissected hilly upland reaching an elevation of about 2,600 feet (790 metres) above sea level. In the west, on the border of Goiás, Minas Gerais consists of a plateau, part of an ancient erosion surface still untouched by stream erosion over wide areas and standing about 4,000 feet (1,200 metres) above sea level. The watershed divides, often erroneously described as mountain ranges, are in fact remnants of this old erosion surface. Minas Gerais is drained to the east and south of the Serra do Espinhaço by the smaller streams named above and by tributaries of the Paraíba. In the southwest it is drained by the Rio Grande, one of the major headwaters of the Paraná River, which reaches the sea at Buenos Aires. The whole northwestern part of the state is drained by the São Francisco, which flows northward through a rift valley to the west of the Serra do Espinhaço. Along the western border is a triangular projection (Triângulo Mineiro) lying between the Rio Grande on the south (border of São Paulo) and the Paranaíba on the north (border of Goiás).

Drainage

Climate. The climate is characteristic of an upland region in the low latitudes. Temperatures are not so high as they are along the coast. At Belo Horizonte, 2,800 feet (852 metres) above sea level, the average of the warmest month (February) is 72° F (22° C); and the average of the coldest month (July) is 62° F (about 17° C). The rainfall throughout most parts of the state averages between 40 and 60 inches (1,000 and 1,500 millimetres) but is concentrated in a rainy season between November and March. The period between May and September is very dry, and lack of moisture is a major factor in agriculture along the stream divides.

Natural vegetation. Reflecting these climatic conditions, the southern and eastern part of the state was originally covered with a dense tropical forest, consisting of many species of trees that retained their leaves throughout the year but with a few that shed their leaves in the dry season. Semideciduous forest was called *mata de primeira classe* ("first-class forest") by Brazilians.

On the upper slopes of the Serra da Mantiqueira and of the Serra do Espinhaço, the forest gives way to upland meadows or to a scrubby growth of deciduous woodland. All of southwestern and western Minas Gerais is covered by a woodland savanna, consisting of stretches of grassland mingled in complex fashion with stands of woodland. The latter (*campo cerrado*, literally "closed grassland") is made up of low broadleaf trees with gnarled trunks. Even in the midst of the woodland savanna, however, semideciduous forests extend along the deeply cut river valleys.

Population. The population of Minas Gerais is chiefly of Portuguese origin, with a considerable admixture from other European countries. Persons of African and mixed (African and European or European and Indian) ancestry constitute other important elements, although their numbers are proportionately less than in some of the other states. As in the nation as a whole, the population is young, with a high proportion in the younger age groups and a very small proportion in the group aged 65 and over. The language is Portuguese, and the predominant religion is Roman Catholic.

The cities and towns

Belo Horizonte, not only the state capital but also a leading industrial centre, is by far the largest city. The second city is Juiz de Fora (population 220,000) in the south, noted for its textiles. Other important cities and towns are Barbacena (58,000), Governador Valadares (125,000), Venda Nova (120,000), Uberlândia (110,000), and Uberaba (108,000). Outside of the cities there

is a dense rural population in the south and east, but the whole northern and western part of the state is still very thinly occupied so that the population density as a whole is only about 51 persons per square mile.

Administration and social conditions. The state government and its principal agencies and offices are located at Belo Horizonte, the capital. The state has its own constitution and laws, in conformance with the national constitution. The governor and the members of the legislature are elected. Minas Gerais has been one of the few states of Brazil able to sustain an economically viable state government, most of the others having to rely on aid from the central government.

Living conditions in general are higher in the urban areas, and health services are more accessible in the cities. Primary, or elementary, education through the first four years is free and compulsory, as in the rest of Brazil. Although shortages of schoolrooms and teachers throughout the nation have prevented the achievement of universal literacy, Minas Gerais has been one of the leading states in providing for education. About half of all Brazilian primary pupils and even higher proportions of secondary and university students live in Minas Gerais and three other states, Guanabara, São Paulo, and Rio Grande do Sul. The Universidade Federal de Minas Gerais, one of the country's 24 official universities, was founded in 1927, and the Universidade Católica de Minas Gerais was established in 1958; both are in Belo Horizonte. The Escola Nacional de Minas e Metalurgia (founded 1876) is at Ouro Prêto.

Economy. The largest source of wealth in Minas Gerais is mining. Of major importance are iron ore and ferroalloys. Large steel plants make use of the ores. But the largest source of employment is agriculture, and the greater part of the area is used for pasture of cattle.

Minerals. Minas Gerais has long been known for its wealth of minerals and its mining activities, supplying half the mineral production of Brazil. Gold was discovered in 1698. At first the many prospectors who rushed into Minas Gerais only looked for gold in the stream gravels, and there are many marks of abandoned placer mines at the valley bottoms and on hillsides. Gold was found in various streams draining away from the Serra do Espinhaço.

In 1729 diamonds were discovered, also in stream gravels, near Tejuco (now called Diamantina). These precious commodities created great wealth in the 18th century, and Ouro Prêto, the administrative centre of Minas Gerais, became a city of great splendour, with numerous beautifully decorated churches. But by 1800 the richest gold sources had been used up, and the miners turned their attention elsewhere. It is interesting, however, that at Morro Velho, a few miles south of Belo Horizonte, a British mining company has been mining gold continuously since 1834. This mine is now one of the world's deepest. Minas Gerais is also widely known for its semiprecious gems, such as aquamarine, amethyst, tourmaline, and topaz, and it is an important source of industrial diamonds. From the high country in the western part of the state come large quartz crystals, used in radios and other electronic communications equipment.

Iron and manganese deposits

Of greater importance in modern times, however, are the vast deposits of high-grade iron ore and manganese. In the southern part of the Serra do Espinhaço there is a large reserve of iron ore, containing few impurities and so located that it can be mined with mechanical shovels. By the late 1960s there were three large-scale steel plants using this ore. Iron and manganese from Conselheiro Lafaiete, near the southern end of the Serra do Espinhaço, are used to supply the steel plant at Volta Redonda and also a newer plant located near Santos. Ores from Itabira, on the headwaters of the Rio Doce, supply a new plant nearby. These ores are also used in several large plants and hundreds of smaller ones that use charcoal as fuel. Moreover, Brazil now exports its iron ore. Brazil has contracted to send some 50,000,000 tons (45,000,000 metric tons) of iron ore to Japan before 1977. This ore moves down the Rio Doce to the new ore port of Tubarão, just north of Vitória in Espírito Santo.

Another iron-ore shipping arrangement was made with a combined Brazilian–U.S. firm to send ores from Minas Gerais to a new ore port on the Baía de Sepetiba near Angra dos Reis, west of Rio de Janeiro.

In addition to iron and manganese, industrial diamonds, and gem stones, Minas Gerais supplies bauxite, zirconium, iron pyrite, rutile, graphite, chromite, molybdenum, nickel, tungsten, titanium, beryl, and mica.

Manufacturing. Manufacturing and processing industries have developed in and around the cities, especially around Belo Horizonte, where industrial air pollution has become a problem. In addition to iron and steel, principal industries include food processing, textiles, chemicals, and furniture and other consumer and capital goods.

Agriculture. The agricultural area of Minas Gerais is almost exclusively a region once occupied by semideciduous forest, which has been cleared by centuries of slash-and-burn agriculture, or land rotation. In this system of agriculture, a tenant farmer clears the forest by cutting and burning it at the end of the dry season. He then plants maize (corn), rice, beans, and cassava (manioc) and is able to reap fair harvests for a period of two or three years before the essential infertility of the soil becomes apparent. He then plants pasture grass, chiefly for cattle grazing, and moves to a new part of the forest. This area is the chief source of beef for Rio de Janeiro, São Paulo, and other cities. Dairy farming is important in southern Minas Gerais. In only a few spots is the land used for sugarcane or coffee, which were formerly of greater importance. On the hillsides, used for poor pasture, are traces of coffee plantations. The state is still a leading producer of rice, beans, maize, and cheese.

Tourism. Another aspect of the economy of Minas Gerais is tourism. In the Serra da Mantiqueira and in the southwestern uplands of the state, to the north of these mountains, are a number of resort towns, most of them possessing various hot and curative baths. The mineral waters are bottled and widely dispensed in Brazil. The more important centres are Araxá, Caxambu, Lambari, Poços de Caldas, and São Laurenço. Health resorts have also been established at the higher elevations.

Roads and railroads

Transportation. Minas Gerais is served by railroads, all-weather highways, airports, and by river steamers that descend the São Francisco from Pirapora to Juàzeiro in Bahia. The main railroad, Estrada de Ferro Central do Brasil, runs from Rio de Janeiro to Belo Horizonte, with a narrow-gauge extension to Pirapora. To the west of this line is the narrow-gauge Viação Ferrea Centro-Oeste; east of it is the Estrada de Ferro Leopoldina, also narrow gauge. A railroad also ascends the valley of the Rio Doce from Vitória to Itabira and continues to Belo Horizonte.

Since World War II all-weather motor highways provide easy access to Rio de Janeiro, São Paulo, and the capital, Brasília. Most of these roads parallel the earlier railroads rather than providing feeder service. Away from the main highways and the railroads the roads are passable for motor traffic only in dry weather.

Cultural life. As is to be expected, the cultural life of the state is concentrated in its cities and especially in the capital, Belo Horizonte. The city plan was based on that of Washington, D.C., with broad, tree-lined avenues and a main concourse, Avenida Afonso Pena, leading from a large central square, Praça da Liberdade. There are many impressive government and other buildings. Its university library is one of the nation's 10 regional centres of bibliography and documentation for scientific information. The city's plan included the large recreational area at the artificial lake Pampulha. The half-dozen or more newspapers in the state are published in the capital.

The state's chief historical attraction is Ouro Prêto, the old colonial capital. It was decreed a national monument in 1933 and the surrounding area a biological preserve in 1967. Public buildings, churches, and private homes, mostly from colonial times, have been preserved or restored. The Escola Nacional de Minas e Metalurgia and a museum of minerals are located in the colonial governor's palace and a historical museum in the old penitentiary. A national park system was established in 1937.

(P.E.J./Ed.)

Mind, Philosophy of

One thing that sharply distinguishes man from the rest of nature is his highly developed capacity for thought, feeling, and deliberate action. Here and there in other animals, rudiments, approximations, and limited elements of this capacity may occasionally be found; but the full-blown development that is called a mind is unmatched elsewhere in nature. The existence of a mind in some sense is reflected in an extremely diverse class of occurrences: sensations such as tickles and aches, mental images, sense perceptions, emotions, memory and expectations, desires, believing, inferring and other types of reasoning, dreams, motives, choices, actions, traits of character and personality, and the unconscious.

PHILOSOPHY OF MIND AS A DISCIPLINE

Task and scope of the field

The task of the philosophy of mind is to examine and analyze those concepts that involve the mind (including the very concept of the mind itself) in an attempt to discover the nature of each of these concepts, the relations between them, how they are to be classified, and how they are to be related to certain other concepts—especially to the concepts of matter and energy, the human body, and, in particular, the central nervous system.

It should be clear that the range of topics in the philosophy of mind goes far beyond what is intended in everyday discourse by "mind." When, for example, the layman speaks of someone as having "a good mind" or as pursuing "the pleasures of the mind," he is thinking of those particular activities that have to do with abstract reasoning, intellectual pursuits, and the exercise of intelligence. The "mind," as the term is used more technically in this article and in the philosophy of mind in general today, encompasses sense perception, feeling and emotion, traits of character and personality, and the volitional aspects of human life, as well as the more narrowly intellectual phenomena.

In this article, the following topics are examined: the relation of the philosophy of mind to other areas of investigation, both inside and outside philosophy; various philosophical theories of what is peculiar to the mind and mental phenomena; various philosophical theories of the nature of the mind and its relation to the nonmental; philosophical issues concerning particular mental phenomena; and some implications of the philosophy of mind for metaphysics and epistemology.

Relation to other disciplines

In distinguishing the field of philosophy of mind from other sorts of investigation, one immediately obvious feature is its subject matter, the nature of mind and its various manifestations. This serves to distinguish it from empirical sciences like astronomy and physics, which study matter in motion; from formal disciplines like geometry and algebra, which study mathematical relationships; and from other fields of philosophy like the philosophy of art and the philosophy of law. But subject matter alone does not serve to distinguish the philosophy of mind, since the mind is the subject of investigation of other disciplines as well—especially of psychology and of certain phases of biology, physiology, sociology, and anthropology. In comparison with these fields, it is by its method that the philosophy of mind is to be distinguished; for it proceeds not by the methods of empirical investigation—detailed sense observation, the formulation of predictions, the construction of experiments, inductive confirmation, the inventing and testing of contingent generalizations, theories, and laws—but by the method of philosophical reflection. That method consists of the examination of meanings, the analysis and clarification of concepts, the search for necessary truths, the use of deductive inference, *reductio ad absurdum*, and arguments with infinitely repeating terms and other forms of a priori reasoning, and the attempt to arrive at and evaluate the fundamental principles that underlie and justify the basic forms of human thought and endeavour.

Although the philosophy of mind is a distinct field of investigation, it has many important relations with other fields. First, its methods, being those of philosophy in general, are to be tested by the fruits that they have yielded in other areas: if a method has been successful in other areas, it is reasonable to try it here; if unsuccessful in other areas, it is suspect here. Second, the conclusions achieved in such fields as theory of knowledge, metaphysics, logic, ethics, and the philosophy of religion are quite relevant to the philosophy of mind; and its conclusions, in turn, have important implications for those fields. Moreover, this reciprocity applies as well to its relations to such empirical disciplines as neurology, psychology, sociology, and history. Thus, the philosopher of mind must keep himself informed of developments in all related fields of investigation.

THE SEARCH FOR A CRITERION OF THE MENTAL

The bewildering variety of phenomena that fall under the heading of the mind or the mental was listed earlier. The question arises, however, whether there is something that all these mental phenomena have in common, something that characterizes them or can serve as a criterion of the mental. More specifically, are there features that are necessary or sufficient for mental phenomena?

Family resemblances among phenomena

It should be pointed out at the very start that, according to some contemporary philosophers, it is a basic mistake to lump all these phenomena together under a single heading. An Analytical philosopher, Gilbert Ryle, in his *Concept of Mind* (1949), maintained that it is only through a series of confusions that one comes to think that there exists here a natural class of kindred phenomena. Even if, contrary to Ryle, it is held that it is reasonable to group these phenomena together, it still might well be that they have nothing in common; and even if they have, that something could turn out to be either peripheral and inessential or comparatively minor and thus might not go very far toward explaining why they all had come under that heading. The remarkable pioneer of 20th-century Analytical philosophy Ludwig Wittgenstein showed that a number of things could come under a particular heading not by virtue of having certain features in common but by virtue of having a "family resemblance," just as members of a family might have a typical appearance even though there was no set of features possessed by all of the members; and the possibility must be left open that this is the case with the various instances of mental phenomena.

Many philosophers, nevertheless, have searched for something that could serve as the criterion of the mental, something that distinguishes what is mental from what is not. Some of the more promising candidates are examined below.

Purposeful behaviour. Whenever a man watches a hungry animal using stealth and cunning in searching out, attacking, and killing its prey, he cannot but believe that the animal has a purpose and uses intelligence in achieving his goal. Whether it be a team of scientists designing a way of launching a man to the moon, an ape figuring out how to screw two pieces of pipe together to get a banana that is out of reach, or—at a much lower level—a lobster trying to get out of a pot of boiling water, there seems to be a mind at work. Somewhere, as the phenomenon is traced farther down the ladder in the scheme of things toward inanimate matter, a line must be drawn; but there is no agreement about where it should be drawn. It would be widely agreed that an ovum that has just been fertilized does not have a mind and that a normal adult does—but it is impossible to say exactly where in human development the change occurs. On the other hand, it would be just as wrong to conclude that because there is no sharp dividing line there is no change as it would be to conclude that red and orange are the same colour because no sharp line divides them. In both instances there will be a range, as with all vague terms, within which the picking out of a dividing line will be arbitrary; but at either end beyond that range a clear and definite difference will be evident.

The question may then be raised of how adequate purposeful activity is as a criterion of the mental. A major issue arises from the fact that a mechanical device can be built that exhibits the kind of activity ordinarily called purposeful—a surface-to-air missile, for example, so designed that it will hunt for a jet aircraft by searching for,

Teleological mechanisms versus mental activity

and zeroing in on, the heat exhaust, thereby finding and destroying the aircraft. Such devices, known as "servomechanisms" (of which a thermostat is a simple example), achieve an end state by systematically diminishing any deviation from that predetermined end state. There are, in addition, the modern computers, capable of receiving, storing, and retrieving information, of making inferences, and of communicating information. Near the height of modern sophistication in machine technology are those machines, combining servomechanisms and computers, designed to roam the surface of the moon, gathering, processing, and sending back data and thus providing what would seem to be paradigms of purposeful behaviour and raising the question whether such systems are examples of minds at work. There are three possibilities worth considering: (1) one might continue to hold that purposeful behaviour is the criterion of the mental but argue that these devices do not really meet the criterion of exhibiting purposeful behaviour; (2) one might hold that the more complex of these devices do indeed have minds—simple and rudimentary, to be sure, but still minds—"artificial intelligences" that in a very literal sense have beliefs, think thoughts, solve problems, and achieve goals; or (3) one might give up the criterion of purposeful behaviour, or at least give it up as a *sufficient* condition of the mental.

Someone defending possibility (1) would attempt to find some feature of purposeful behaviour that could not be found in any mechanical device. He might hold, for instance, that the alleged purposes of mechanical devices are built in by their designers and that it is not up to the devices to choose their purposes. Against this line of argument, however, others might assert that for a large number of organisms the basic purposes are also built in through the genetic mechanisms of heredity: basic biological drives for food, sex, and safety govern most of their purposeful behaviour. People who accept this point of view argue that, in most cases, the organism is not really free to choose its purposes. On this argument, even human beings seem to display such built-in purposes; and it would be specious, at least in this instance, to deny them minds on this account.

Machines, man, and inventiveness

Conceding that organisms have some built-in controls, the defender of the first possibility might then argue that the higher organisms have a flexibility that allows for the development of purposive methods that are novel. Machines, he would allege, solve problems only by the methods designed into them, whereas creatures with minds can invent new methods. Against this line of argument others might argue that machines have been developed to use trial and error, analyze the outcomes of trends, and come up with new approaches that are more successful than earlier ones, and that, in an important sense, such machines might be called creative, since they "learn from experience" and use "ingenuity." Many observers assume that this trend in machine technology will continue and that machines will be developed with much, if not all, of the flexibility of many of the creatures that would unhesitatingly be acknowledged to have minds.

The basic issue, then, is whether a philosopher would want to say that a machine has a mind if it exhibits the flexibility in purposeful activity of, say, a normal seven-year-old child, who undoubtedly has a mind. In accepting possibility (2), he would maintain that such a machine does have a mind and exhibits a wide range of the phenomena called mental. Many contemporary philosophers, however, would reject that thesis, considering it a needless flaunting of common sense and an affront to ordinary language to speak of a heat-seeking missile, or even a lunar robot, as having a mind. Because something essential still appears to be omitted, further attempts are made below to get at the essence of the mind.

Intentionality. Another characteristic of the mental is sometimes thought to be found in certain ways in which a man may be said to have something as his object. Thus, thinking, believing, desiring, and other such attitudes are thought to resemble one another in that they may be said to take an object, or to be directed upon an object, in a way quite unlike anything to be found in what is purely physical. "Intentionality" is the term for this way of being directed upon an object. The concept had been emphasized by some of the Scholastics and was introduced to modern philosophy in 1874 by a German philosopher and psychologist, Franz Brentano, and more recently has been clarified and defended by a United States philosopher, Roderick M. Chisholm.

The nature of intentionality. The idea of intentionality can be explicated in the following way: If one imagines three objects arranged as in Figure 1 and then supposes

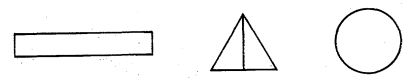


Figure 1: Three arbitrary physical objects in meaningless configuration (see text).

that the wind blows them so that they are arranged as in Figure 2, there results, from the physical point of view, simply a new arrangement. From a psychological point

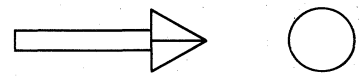


Figure 2: Physical objects with intentionality.

of view, however, something radically new also has appeared: one object is now pointing to another—aimed at or directed toward it. It would seem that such pointing cannot be accounted for if the observer confines himself to the purely physical facts of the new configuration and that his mind has to be brought in to account for the feature of pointing to. Thus, intentionality is a prima facie reasonable candidate for the criterion of the mental.

Added characteristic of aim or direction

Intentionality is exhibited in a variety of phenomena. Thus, if a person has an emotion toward an object—*e.g.*, loves, fears, pities, envies, or reveres it—he has an intentional attitude toward it. Other examples of intentional attitudes toward an object are: looking for or expecting, believing in, doubting or conjecturing about, daydreaming, reminiscing, imagining, favouring or disapproving—a list that seems to go on endlessly. Because it clearly comprises so many of the things that one thinks of as typically mental, intentionality, being of broader scope than purposefulness, would seem to be a better choice than purposeful behaviour for the criterion of the mental.

One of the characteristics of intentionality is what the Scholastics called "inexistence": a man may be intentionally related to an object that does not exist or to an event that does not occur. Thus, what a man looks for may not exist, and an event that he believes to occur may not occur at all. In contrast with such a nonintentional phenomenon as bumping into something, in which the object bumped into must be real, looking for something (an intentional act) does not necessarily imply that the object looked for exists. Similarly, in contrast with an explosion's resulting in the fact that many were hurt, a witness's believing that many were hurt (again an intentional act) does not imply the fact that many were hurt. Thus, existence and truth are irrelevant to intentionality.

Though this possible relationship to nonexistent objects as opposed to existences is a necessary feature of intentionality, it is not sufficient to define it, for there are phenomena that are equally concerned with nonexistent essences that are nonetheless nonintentional. "That lady resembles a mermaid," to use an example of Chisholm's, may be a true sentence even if mermaids, though having an essence, do not exist. Similarly, "That metal will ignite at temperatures above 1,000,000°" may be true whether or not such temperatures exist. Thus, intentionality requires further characterization if its scope is to be narrowed to exclude such examples as these.

Another characteristic of the intentional state is that not every description of its object will be appropriate. Assuming that his pen is the millionth pen produced this year, for example, a man may be in the intentional state of searching for it as his pen but not in a state of searching for the millionth pen produced this year; similarly, he may believe this is his pen and yet not believe that this is

the millionth pen produced this year. This second feature of intentionality, often called "referential opacity," is such that a true sentence asserting an intentional state will become false when some alternative description of the object of that state is substituted for it (it is false that he is searching for the millionth pen).

There is no general agreement on the best way of conceiving of intentional phenomena. Brentano, at one point, thought of intentionality as being a relation between a subject and an entity, in which the entity is something that might or might not exist; but, grave difficulties arise in the effort to characterize the ontological status of such an entity—*i.e.*, its kind of reality. More popular today are certain linguistic approaches: intentionality may be viewed, for example, as it was by Rudolf Carnap, a philosopher of science, and others, as a relation between a subject and a piece of language or, as it has been explained by others, as a relation between a subject and a linguistic practice or linguistic role. Under this view, the intentionality consists not in the relation of a subject to an *essence* (that of a millionth pen) but in its relation to a *sentence* ("This is the millionth pen . . .") that has that alleged essence as its meaning. It remains to be seen whether such approaches will succeed in dealing with all intentional phenomena. It would seem to pose particular difficulty in those cases in which the intentional state is overtly directed toward some existent entity, with language playing a minor or null role, as in situations when one feels anger, pity, or love toward someone; or when an animal is stalking its prey (which involves an intentional state). In such instances, an analysis in terms of linguistic attitudes would seem wide of the mark.

Application of intentionality to nonliving systems

The scope of application of intentionality. The question of whether intentionality can apply to nonliving physical systems has become a controversial issue. If it can apply, then either intentionality would have to be given up as a criterion of the mental or else one would have to say that such systems have some mental characteristics.

It will be useful to consider a system designed for some purpose, taking the example again of the surface-to-air missile that searches for jet aircraft, and to ask whether it has intentionality. It does satisfy the first characteristic mentioned above: that it can be truly said to be a jet searcher regardless of whether there are or have been any jet aircraft (one could similarly design a unicorn searcher built to detect unicorns by their special horn). The question next arises whether some descriptions of the object are inappropriate. On the supposition that all and only jet aircraft have a component made of compound *X*, one can ask whether it would be true of the missile that it searches for things with a component made of compound *X*. It would seem that—unless this compound chances to be what the search system was sensitized to—the foregoing statement is a false, and thus an inappropriate, description of the device; and if so, it is plausible to regard the missile as a physical system with intentionality as that notion has been here characterized. An intentional physical system, however, would have to be of considerable complexity. One would not want to say that the left-hand complex in Figure 2 was pointed to the right-hand figure unless he had in mind that, if the right-hand figure were to shift its location, the left-hand complex would shift appropriately. If it did keep shifting appropriately, however, it would seem proper to say that it points to the right-hand figure, for a compass clearly points to magnetic north. The question remains, however, whether *all* intentional phenomena are capable of appearing as instances in nonliving physical systems—whether such a physical system could have, for example, an emotion toward something, daydream about something, or be amused by something. Here it is very difficult to cite a plausible case; it would appear that one would have to strain such concepts considerably to apply them to nonliving systems.

The thesis that intentional phenomena are the essence of the mental thus seems problematic. Its suggestion that the jet-searching missile has a mind or partakes of the mental, as appeared in the discussion of the criterion of purposeful behaviour, would, to many scholars, appear to be quite implausible. Nor does it seem that the trouble lies in the limited number of intentional phenomena found in the missile. Even the lunar-exploration machines, with all of their flexibility and multiplicity of functions, would not be said by most analysts to have minds.

It might be possible to save intentionality as the criterion of the mental by insisting on the presence of such highly sophisticated intentional phenomena as emotions, daydreams, or amusement, but then one would have to deny minds to those human beings who lack a sense of humour, never daydream, or are coldbloodedly unemotional, which does not seem correct either. Some progress can be made here if the question is asked whether intentionality is a necessary condition of all mental phenomena —whether there are any phenomena that are mental but nonintentional. Examples of a mental phenomenon that can most plausibly be said to be nonintentional are sensations, such as feeling pain, which lack both of the aforementioned characteristics of intentionality—inexistence and referential opacity. A man cannot feel a pain that, unbeknown to him, does not exist; if he feels pain, there must be something he feels. Moreover, if a man feels pain, and the pain is identified with the effect of a tumour, then he does feel the effect of the tumour.

Sensations, which thus lack both of the characteristics of intentional phenomena, are not just an odd counterexample; not only do sensations comprise a large and central group of mental phenomena, but they also call attention to an important aspect of many other mental phenomena, viz., subjective experience. The arrow in Figure 2 may point to the circle, but it does not have the subjective experience of pointing, it does not feel itself pointing; and the jet-searching missile does not experience how it feels to be searching for jets. But when a person is in an intentional state, directed toward an object, he—at least sometimes—experiences, feels, or is aware of that directedness.

Clearly, this usage of "intentionality" differs somewhat from that of medieval philosophy; and there are other features of the concept, not covered here, that are stressed by Phenomenologists and Existentialists.

Subjective experience. It is often maintained that the essence of the mental consists of states of consciousness taken as subjective experiences. When a person wakes up or regains consciousness after a general anesthetic, a host of experiences of colour and light, sounds, feelings, thoughts, and memories flood in on him. As far as his objective, observable behaviour is concerned, he may be lying unmoved and unmoving; but as far as his state of consciousness is concerned, he may be undergoing a series of subjective experiences.

Consciousness, "raw feels", discrimination

To take an example, when a person sees a scarlet patch, he experiences the homogeneous, spread-out, distinctive scarletness present before him. A blind or colour-blind person who has never experienced scarlet would not have the awareness of scarlet that the normally-sighted person has. He might have some vague idea, as did the hero of the story told by John Locke, a 17th-century British Empiricist:

> A studious blind man, who had mightily beat his head about visible objects, and made use of the explication of his books and friends, to understand those names of light and colours which often came in his way, bragged one day, that he now understood what scarlet signified. Upon which his friend demanding, what scarlet was? the blind man answered, It was like the sound of a trumpet.

This reply is not totally wide of the mark; but any sighted person will have a far more precise idea of scarlet than that. What he has and what the blind person lacks is something that philosophers have called the "raw feel" of scarlet, that peculiar and special way scarlet looks.

The subjective experience of scarlet is to be contrasted with the discrimination of scarlet things. One could imagine a blind person who was able to discriminate scarlet from other colours by the use of optical instruments (*e.g.*, spectroscopes with Braille printouts). But he would lack the subjective experience of scarlet; he would not know the look of scarlet. A blind person might even be

imagined with, say, a totally severed optic nerve, who was so equipped with instruments that his behaviour was indistinguishable from that of a sighted person; but he would still lack the subjective experiences of sight.

Defenders of this view would claim that there is a great variety of subjective experiences and that the experience of colours is only one of them. Sensations (*e.g.*, the experiences of pain, tickles, throbbings, pangs, nausea, and tiredness) provide another such example. Still other subjective experiences include: the experiencing of images (afterimages, memory images, and others); feelings of exultation, depression, pride, anger, fear, and love; and thoughts (imaginings, surmisings, doubtings, and recollectings). All of these are episodes, occurring at a particular time and place, in which the subject is in a state of awareness that has a particular content.

Adequacy as a criterion of the mental. The question now arises of how adequate subjective experience is as a criterion of the mental—whether, though it is obviously a sufficient condition for something to be mental, it is a sufficient condition for something to have a mind. The Scotsman David Hume, an 18th-century philosophical Skeptic and historian, once asked whether a creature that had but one state of consciousness could be said to have a mind and concluded that it could not. In his view, it takes, at the very least, a number of states of consciousness linked by memory before one would say that the creature has a mind; and it may be that there has to be a certain level of complexity in the nature and relation of the conscious states for there to be a mind.

The unconscious, dispositional, behavioral

It is doubtful, however, whether consciousness is a necessary condition for the mental. Before Sigmund Freud, it would have been widely agreed that the notion of unconscious mental phenomena was logically impossible—a contradiction in the very terms. That view had one important exception, however: Gottfried Wilhelm Leibniz, a 17th-century Rationalist and mathematician, held that there are *petites perceptions* of which the subject is unconscious. They are so slight, so similar to others, so familiar, or in such a crowd of other perceptions, that the subject is unaware of them at the time. One of the examples that Leibniz cited is the person who is unaware of the roar of the waterfall or the rumble of the mill if he has lived nearby for some time. Leibniz seemed to have had in mind what modern psychologists call "subliminal" perceptions, viz., those below the threshold of awareness but still capable of leaving some effects on the mind. But Leibniz confined unconscious states to perceptions; he would not have allowed unconscious beliefs, desires, emotions, or judgments.

It was Freud's great contribution to have discovered a range of phenomena of which the patient was unconscious but which was very much like typically mental phenomena, especially in their behavioral manifestations. In the light of such similarities, it was plausible to extend the concept of the mental to include these unconscious phenomena—especially since they were such that the patient could become conscious of them through hypnosis or psychotherapy. Freud postulated a mechanism that he called "repression" to explain why the patient is unconscious of them.

In addition to the subliminal and the unconscious, there are more familiar characteristically mental phenomena that do not consist of states of consciousness. When a man falls into a dreamless sleep, he does not lose all his beliefs or abandon all his goals, he does not cease wanting a better world or being artistic or imaginative or lazy, nor does he forget how to do arithmetic or speak French. A person is not jealous of someone only when thinking of him, nor does a businessman have confidence in the dollar only when concentrating on business. Obviously, these mentalistic characteristics can apply in a dispositional way to people who are not at that moment expressing or exhibiting the disposition.

Furthermore, as Ryle has pointed out in great detail, a person may use his mind on many occasions without the feeling of subjective experiences. As he says,

> When we describe people as exercising qualities of mind, we are not referring to occult episodes of which their overt acts and utterances are effects; we are referring to those overt acts and utterances themselves. (This and the following quotations attributed to Ryle are from *The Concept of Mind* by Gilbert Ryle. Copyright © 1949 by Gilbert Ryle. Reprinted by permission of Barnes & Noble, Publishers, New York, Hutchinson Publishing Group Ltd., London.)

To be responsive to one's surroundings, to act intelligently, deliberately, with wit or good grace, to do arithmetic or logic, to be sympathetic or coldhearted, to drive alertly or absentmindedly—none of these requires the occurrence of subjective experiences or inner states of consciousness, the immediacy of feelings or sensations. In such activity, there may be nothing going on except performances of a particular kind, and there may be nothing more required except that under further circumstances other performances of a particular kind will be forthcoming. It is, thus, reasonable to conclude that subjective experience is not a necessary condition for the mental.

Ryle's charge of a category-mistake

Those who have put the private events of subjective experience at the center of the mental have committed what Ryle calls a "category-mistake . . . represent[ing] the facts of mental life as if they belonged to one logical type or category, . . . when they actually belong to another." The mistake consists of taking talk about a person's mind as talk of events in a world parallel to the ordinary world but occult and mysterious. The truth, according to Ryle, is that

> to talk of a person's mind . . . is to talk of the person's abilities, liabilities and inclinations to do and undergo certain sorts of things, and of the doing and undergoing of these things in the ordinary world.

It would be rash, however, to draw the further conclusion that subjective experiences are in no way involved in whatever is mental. Returning to the case of Leibniz' *petites perceptions* that are not experienced, a person can be conscious of them in various ways, either before getting used to them or when they are alone or when their intensity or his own sensitivity is increased; and Freud's unconscious phenomena can become conscious phenomena under favourable conditions. The beliefs that a man is not aware of in sleep are sometimes the objects of his consciousness, as are his moments of laziness and imaginativeness, his knowledge of arithmetic, and his goals. It is dubious that something that has no connection with states of consciousness could qualify as mental.

Core characteristics of subjectivity. Philosophers are still in deep disagreement on how to characterize what is peculiar to subjective experiences. Some hold that the existence of subjective experience indicates that there are peculiar events that do not occur in the public space–time world that everyone shares and has equal access to but occur only in a private world that each person has exclusively to himself, which he cannot share with others, and to which no one else has access. One man's sensations, for example, are private events that occur in his private world, and another man's sensations occur in a second private world, so that each person has a private world with its own private events.

The role of privacy

Ryle has called this view, with what he admits to be "deliberate abusiveness," "the dogma of the Ghost in the Machine." He characterizes the dogma as follows:

> Minds are not in space, nor are their operations subject to mechanical laws. The workings of one mind are not witnessable by other observers; its career is private. Only I can take direct cognisance of the states and processes of my own mind. A person therefore lives through two collateral histories, one consisting of what happens in and to his body, the other consisting of what happens in and to his mind. The first is public, the second private. The events in the first history are events in the physical world, those in the second are events in the mental world.

Are there private events? Even so adamant a critic of privacy as Ryle admits the existence of some private phenomena, chiefly dreams and daydreams, sensations, thoughts, and imaginings. He insists, however, that

> the sequence of your sensations and imaginings is not the sole field in which your wits and character are shown; perhaps only for lunatics is it more than a small corner of that field.

In Ryle's view, private events occupy a small and in-

essential place in the total range of mental phenomena; but they do occur.

The notion of privacy is really the conflation of two ideas: the metaphysical idea that mental events do not occur in space and the epistemological idea that mental events are objects of awareness solely to the person who is subject to them. Each of these may be considered in turn.

Nonspatiality of mental events

The Rationalist René Descartes, the earliest major philosopher of modern times, held that the essence of all that is nonmental consists in being extended in space. Turning this around and broadening it, one could say that the essence of the mental consists in the lack of spatiality; *i.e.*, the lack of shape, size, and, above all, location. If the philosopher confined himself to events, he would say that necessarily a physical event occurs in some place or other, but, necessarily, a mental event does not. It would be conceded that the person who experiences the mental event does typically have a location, and this leads to the question of why the event is not located where the person is located.

A defender of the nonspatiality criterion would argue that such ascriptions of location to mental events are very different from ascriptions of location for physical events. For a physical event, it is always possible to ask whether it occurred at some point, in some part, or throughout the location. Thus, if the temperature of a body of water rises, one can ask precisely where the rise occurred—at certain points, in certain parts, or throughout the volume. But if a thought occurs, it is senseless to ask whether it occurred throughout the area or only in some part of it. Furthermore, if the water undergoing the rise in temperature is in a box, it is reasonable to say that a rise in temperature occurred in the box; but if the person having a thought is in a box, it is senseless to say that a thought occurred in the box. So the sort of ascription of location is quite different for mental events, and the criterion can still be used to mark off the mental from the physical.

The question remains whether the sort of nonspatiality that is allegedly appropriate to the mental is peculiar to it. If such a physical event as recovering from an illness or changing shape is considered, it would appear also that it does not make sense to ask whether the event occurred throughout the whole volume, in some part, or at some point. Thus, it would seem that even this modified notion of spatiality does not uniquely distinguish the mental.

Privileged status of subjectivity. Other philosophers would interpret subjective experiences not as private events but as a special way of knowing certain events, specifically by introspection. This is called the "privileged access" view. John Locke, contrasting this way of knowing with sensation, called it "reflection," defining it as "that notice which the mind takes of its own operations and the manner of them." It is a way of being aware of one's own present states without the intervention or use of the senses. The emphasis here is on the nature of the way of knowing rather than on the nature of the events known. Someone who holds the privacy view will have to hold that there is some special way of knowing these special events, but someone who holds the privileged access view is not necessarily committed to holding that the events so known are in any way special. A person could hold that one and the same event can be known both by sense perception and in some other way. Being knowable by introspection would then be the characteristic that defines events as mental. Such an account, however, would not rule out that they may also be knowable by sense perception or by inference.

Objections and rejoinders

Some contemporary philosophers deny that there is any such special way of knowing. Ryle offers three objections: first, it would require that there be simultaneously multiple attentions—in the mental event itself and in the attending to that event; though it is not denied that such divisions of attention are possible, he suggests that they are more unusual and difficult to achieve than the proponent of introspection would have the reader believe; second, because there is obviously some upper limit to the number of simultaneous attendings that a person is capable of, there will have to be some mental acts of which a person is unaware, and if it is admitted that some mental events occur without being known in this special way, it is fair to ask whether one must assume that any of them are known that way. Finally, for many states of mind—*e.g.*, extreme panic or fury—the person is so involved that he is incapable of taking note of them; yet such states are not, in consequence, suspect—the person involved is as sure that they occur as he is of any so-called mental events. There is thus no need to postulate this special way of knowing to account for man's knowledge of any such events.

It is not clear how compelling Ryle's objections are. It is admitted that attention can be divided, though it may be contended that it is unusual and difficult to achieve this division. Others would reply that it is a lot more common and easier than might be thought, that it occurs whenever a man takes note of his mental states. And from the fact that he cannot take note at the same time of very many of his mental states, it hardly follows that he never does; each could still be introspectable even if it was not actually introspected on that occasion. As for Ryle's third objection, it might be that some states of mind cannot be introspected, but it does not follow that none can be introspected; they might still be private for all that. Ryle, for instance, while denying introspection, admits retrospection, a capacity to recall one's states just after they occur. It would seem, however, that there is no important difference between a concurrent "introspection" and a prompt "retrospection." One advantage of retrospection is that it would explain a man's self-knowledge of those events that are difficult to explain in terms of introspection; *e.g.*, extreme panic or fury.

Whether a person introspects or retrospects (the truth appears to be that sometimes he does the one, sometimes the other), he would still seem to have a kind of knowledge about his own present and recently past mental states that he does not have of the mental states of others and that others do not have of his. It is not possible either to introspect or retrospect the mental states of others; the knowledge that a man has of the present and immediately past mental states of others must be based upon perceptions or inferences from perceptions, whereas the knowledge that he has of his own present and immediately past mental states need not be, and usually is not, so based.

Introspection as defining the mental

It is possible that the notion of introspection can thus be used to define the mental. Such a definition would be of the form: a mental event is an inner event that can be introspected. The difficulty remains, however, of how "introspected" is to be defined. If it is defined merely as "known without inference or sense perception," then it would seem to apply equally to the knowledge of certain bodily events that no one would want to call mental. A man can know without sense perception, for example, that his heart is beating rapidly or that his fingers are crossed. To rule such cases out, one can include among the senses the kinesthetic sense that utilizes nerve endings within the body—those, for instance, that register the conditions of one's own muscles. But then it might appear that one must say that sensations are not mental phenomena, since the awareness of them typically involves such nerve endings. Such an admission, however, would be fatal for the privileged access view because sensations are precisely the sort of thing to which a person is supposed to have privileged access. If, on the other hand, the philosopher makes it a matter of definition that introspection applies only to the mental, then he cannot define the mental in terms of introspection. Thus, philosophers are at the present time faced with serious and unsolved difficulties in using the notion of introspection to define the mental.

Finally, it is significant here, as it was in the discussion of subjective experience, that to much of man's mental life and to many of the exercises of his mind—*e.g.*, employments of intelligence—he has no special privileged access; there are, in addition, the unconscious phenomena that are not introspectable. So introspection cannot be a necessary condition of the mental.

There is a clue to a more satisfactory criterion of the mental that can be found in the attack on introspection

cited above. There the difficulty was noted that there does not seem to be a way of distinguishing how a man knows his mental state from how he knows such inner physical states as the rapid beating of his heart. But it may be argued that there does seem to be a difference between these two ways of knowing: specifically, it is clear how a person could be shown that he was mistaken in believing that his heart was beating rapidly; but it is by no means clear what would show a person that he was mistaken in believing himself to be feeling a particular throbbing, pounding sensation. For mental events, the subject's own beliefs are peculiarly authoritative. This authority only holds, of course, for present mental events; it is clear that many things could show that a person's belief about a past mental event of his was mistaken. It is sometimes claimed that what distinguishes mental events is their so-called indubitability—the fact that a belief by the subject that the event is occurring cannot be false and is in no way subject to error. In recent times, however, the view that first-person, present-tense reports of mental events are indubitable has come under serious attack; to make such a report is to classify, and it is argued that it is always possible to err in classification.

Indubitability; incorrigibility; presumption

Instead of holding that such beliefs are indubitable, it is often more modestly maintained merely that such beliefs are "incorrigible," meaning that nothing will count as overthrowing (or correcting) such beliefs. A person who believes that he is experiencing a throbbing sensation may be mistaken; but there is nothing that will show an observer or him that he is mistaken, nothing that will entitle either of them to believe that he is mistaken. He may be experiencing a throbbing sensation even when no part of his body is actually throbbing, though the explanation for this curious fact might not be known.

It might be objected against the incorrigibility thesis that the same difficulty that arises for the alleged indubitability of first-person, present-tense beliefs about mental events also arises for their alleged incorrigibility. For if misclassification is possible, it would also seem possible to gather evidence that someone is misclassifying. If one could confuse a throbbing sensation with a different but somewhat similar sensation, it is reasonable to believe that this confusion could be known by others to have occurred. Perhaps the best that can be said for the incorrigibility thesis is that there is always a strong presumption that such beliefs are true, though this presumption can sometimes prove to be unwarranted. But then it is by no means clear that privileged authority is a unique criterion that distinguishes mental events, for such a presumption would also hold for many nonmental events as well (*e.g.*, the belief that one's heart is beating rapidly). Yet if the degree of presumptive force is taken into account, it is reasonable to say that it would be *comparatively* harder to overthrow beliefs about one's present mental events; and perhaps this is all that one needs to give the criterion force.

THE EXISTENCE AND STATUS OF THE MIND

The basic metaphysical issues in the philosophy of mind concern whether the mind exists and, if it does, what kind of existence it has and what its relation is to the rest of what exists. Materialists hold that only physical matter (and physical energy) exists. For those who hold, on the other hand, that the mind exists as an immaterial entity, there are dualistic theories for which both immaterial minds and material bodies exist, and immaterialistic theories for which only minds exist but not bodies. There are, finally, the so-called neutral monist theories for which the fundamental existents in nature are neither mental nor physical but some neutral stuff out of which both the mental and the material are formed.

History of Materialism

The mind as material. The basic contention of Materialism is that nothing exists but matter and its purely material properties, so that the concepts that are necessary and sufficient for describing and explaining matter will be necessary and sufficient for describing everything that exists. This view can be found in the early Greek philosophers. Thales of Miletus, who lived some 2,500 years ago (6th century BC) and who is generally regarded as the first philosopher in the Western tradition, is supposed to have held that all things are composed of water in some form or other; later thinkers added air, fire, and earth to the list of fundamental elements. The philosopher Anaxagoras of Clazomenae, born about 500 BC, introduced a new factor, *Nous* (Mind), which arranged all other things in their proper order, started them in motion, and continues to control them. There is still controversy as to how his concept of Mind is to be understood, but since he spoke of Mind as being "the finest and purest of all things" and as occupying space, it is likely that he did not think of it, as some later thinkers did, as nonmaterial stuff but rather as a very special kind of material stuff.

It was with the Atomists, Leucippus of Miletus and his disciple Democritus of Abdera, in the 5th and early 4th centuries BC, that Materialism was given its most developed statement. According to them, nature consists solely of an infinite number of indivisible particles, having shape, size, and impenetrability, and no further properties, and moving through an otherwise empty space. The shape, size, location, and movement of these particles make up literally all of the qualities, relations, and other features of the natural world. Such phenomena as sensations, images, sense perceptions, and thought—of particular interest to the philosophy of mind—are explicitly held to consist in the various qualities and relations of the particles.

Contemporary Materialists would, no doubt, wish to incorporate into their theory the latest findings of the physical sciences—the convertibility of matter and energy, the wave–particle duality, and the various subatomic particles and antiparticles with their peculiar properties, including the conservation of charge, direction of spin, and direction of time—but these would represent mere changes in detail. In broad outline, the theory would be the same.

Given the theory of Materialism in such broad outline, however, a serious question remains concerning the actual account to be given of such phenomena as sensations, images, perceptions, emotions, memories and expectations, desires, beliefs, thoughts, imaginings, and intentions. Among the possible views are those called eliminative Materialism, Behaviourism, and the central-state theory.

Eliminative Materialism. A philosopher might hold that there are no such things as sensations, images, perceptions, or emotions and that there never have been such. From this view, those who have believed in the existence of such things have simply been mistaken. A person might hold this view, called eliminative Materialism, on the grounds that all such talk is (1) meaningless noises, gibberish, or (2) part of a set of theories that are outmoded, scientifically fruitless, and to be discarded, like theories about witches or the Homeric gods. On either account, it is implied that all such terms should be eliminated from the philosopher's vocabulary. There is a further doctrine, however, to the effect that all such talk is (3) meaningful but nondescriptive and without truth value. On this account, the proper function of such language is not to state facts but might be to prescribe or evaluate.

Two kinds of reductive Materialism

Among Materialistic views, the alternative to eliminative Materialism is some kind of reductive Materialism. According to this view, there are indeed such things as sensations, images, perceptions, and emotions, but they are only complicated forms of matter in motion. The philosopher may thus continue to use terms referring to such things (in contrast with eliminative Materialism), but he should keep in mind that no extra entities or features are being postulated over and above the physical entities with their physical features.

Behaviourism. If one asks reductive Materialists what sensations, images, and the like are, he will find that two alternatives have been proposed. The first is Behaviourism, the view that all such terms refer to the behaviour or movements of certain bodies, particularly of the higher animals. Thus, the Behaviourist would claim that to feel pain is to groan, writhe, blanch, moan, and so on, or at

least to be disposed or tend toward such behaviour; to desire food is to engage in eating in the presence of food, in hunting in the absence if food, and so on, or at least to be disposed or tend toward such behaviour; and so also with all of the states and activities that one thinks of as mental.

Usually, Behaviourism is intended as a logical doctrine to the effect that the very meanings of the words referring to the mind, its mental states and activities, are to be analyzed in behavioral terms, that every mentalistic term is synonymous in meaning with some behavioral term. It is important to distinguish this view, logical Behaviourism, from the view of many psychologists that the most fruitful way to study psychological phenomena is to study human and animal behaviour. Such a view might be called methodological Behaviourism because it is actually the proposal that the science of psychology restrict itself to certain methods. It does not entail logical Behaviourism. Logical Behaviourism might also be distinguished from the view that psychological and behavioral terms, though not synonymous in meaning, have, as a matter of fact, the same denotation or reference, a view that might be called de facto Behaviourism.

Central-state theory. The second type of reductive Materialism is the central-state theory. In this view, mental states and activities are identical with states and activities within the body (hence this theory is sometimes called the identity thesis). In particular, they are identical with states and activities of the central nervous system or brain. Thus, to feel pain is for the brain to be in a particular state; to desire food is for the brain to be in another state.

Distinctions parallel to those for Behaviourism can be made for the central-state theory. A psychologist might hold that the only useful way of studying psychological phenomena is to study the central nervous system; this view might then be called methodological central-statism. A philosopher might hold that the very terms referring to the mind, its mental states and activities, are synonymous with neurological terms (or, more plausibly, that they *should* be taken to be synonymous—they obviously are not synonymous as language now stands). This position, which could be called logical central-statism, would differ from the eliminative Materialism mentioned above in that it would retain mentalistic terms rather than eliminate them, but would redefine them neurologically. If this came to pass, such terms might eventually disappear, a result that eliminative Materialism would strive to achieve more directly.

Plato's *psychē*; Descartes's substances

The mind as immaterial. Plato was the first important figure in the Western tradition explicitly to defend the doctrine that the mind is an entirely nonmaterial entity —without such defining material properties as size, shape, or impenetrability—separate and distinct from the human body and able to exist apart from it. Plato used the Greek word *psychē* (traditionally translated as "soul") to refer to this entity.

Plato held that the mind (*psychē*) was in charge of the body and directed its movements. In his dialogue the *Phaedrus*, Plato spoke of the mind as having both appetitive desires and the higher desires and as having, in addition, a rational capacity to control, direct, and adjudicate between these two types of desires. This rational capacity of mind is the most valuable aspect of man, the part most worthy to be nurtured and developed, and the aspect of man most likely to be immortal (much of the dialogue concerning the last hours of Socrates, the *Phaedo*, is about these topics).

Dualism. Plato was a dualist; he believed in the existence of both material entities and immaterial ones. The most explicit statement of dualism, however, is found in the writing of René Descartes, who argued that mind and matter are two separate and distinct sorts of substances, absolutely opposed in their natures, each capable of existing entirely independently of the other.

The dualist is faced with the question of how, if at all, mind and matter are related to each other. Most dualists would agree that in rocks, tables, and other material things, matter exists alone and unrelated to mind; and that at what is called death (since for Descartes the soul is immortal), immaterial minds exist unrelated to matter. In the case of a living human being, however, there are two substances: a mind and a body. Thus, the question arises of how the relation between them is to be conceived. Any dualistic theory would have to account for certain obvious facts about human beings. When people's bodies are affected in certain ways—when subjected, for example, to bright lights, loud noises, rises in temperature —people often experience colours, sounds, or sensations of warmth or of pain. Again, when people experience certain things, their bodies undergo certain changes—they shut their eyes, they put their hands to their ears, they perspire, or their faces become pale.

There are various ways that dualists have proposed to account for these facts. The most straightforward position is interactionism, the view (held by Descartes) that mind and body are capable of affecting each other causally, so that what happens in the body can produce effects in the mind and vice versa. Descartes decided that somewhere within the nerve tissues of the brain was the place where the interactions occurred and chose the pineal gland as the precise point because of its central location. (It is now known that the pineal gland cannot perform the functions that Descartes attributed to it, though its precise functions are still unknown.)

Interactionism; epiphenomenalism; parallelism

It is an implication of interactionism that there cannot be a complete explanation of brain functioning exclusively in terms of the laws of neurology because of the intervention at crucial moments of the influences of the mind. This limitation has struck many scholars as an important difficulty. One way around it is epiphenomenalism, the view that the body can affect the mind but that the mind cannot affect the body. Mental events are mere by-products of brain activity, like the exhaust from an engine or the shadows cast by moving figures. When the mind would appear to be affecting the body, as when the experience of pain seems to cause one to grimace, the epiphenomenalist hypothesizes that the very brain state that produces the experience of pain also produces the grimace.

Dualists of either the interactionist or epiphenomenalist persuasion are committed to the existence of causal relations between body and mind. Some philosophers, struck by how entirely different mind and matter are supposed to be on the dualist hypothesis, have held it to be impossible that they could affect each other. Psychophysical parallelism avoids this difficulty by postulating that mind and body are like two perfect clocks, each with its own mechanism but in constant and uniform correlation with the other. Unfortunately, the analogy does not hold very well, for the mind does not seem to have the kind of internal mechanism that would account for any precise sequence of its successive states, and without such a mechanism it would be implausible to expect a constant but noncausal correlation between those states and states of the body.

Immaterialism. Some philosophers have held the doctrine of immaterialism, so named by Bishop George Berkeley, one of the classic British Empiricists, in whose view everything that exists is mental, "of the stuff that dreams are made of," and there is no such thing as the material. There are two major alternatives here: that reality consists of one vast, all-encompassing mind, or that it consists of a plurality of minds. The former position is sometimes called absolute Idealism; the latter, which Berkeley himself held, is sometimes called subjective Idealism. (The term Idealism is often used with wider connotations to include any view that attributes greater value or importance to the mental than to the material aspect of things, or emphasizes the role of the mind in determining the nature of things.)

Berkeley's mentalism

The philosophy of Berkeley represents a highly developed and energetically defended statement of the position that reality consists wholly of minds, the divine Mind and the multiplicity of finite minds that includes all men. Whatever exists does so either because it is a mind or because it is dependent upon a mind; nothing material exists. Berkeley argued that the notion of the material should play no role in one's thinking, for its existence is

unverifiable, its postulation unnecessary, and, at bottom, the very notion is self-contradictory. How does Berkeley view the status of tables and chairs, rocks, the moon, and all of the other apparently material things that everyone accepts as existing? Berkeley agreed that they do indeed exist but only as collections of ideas that exist in the mind of God and that are often caused by God to exist in the minds of men as well.

There are well-known difficulties in Berkeley's view. His account of the nature of tables and other objects cannot be accepted as an account of the meanings of these terms because it is implausible to think that the concept of a divine Mind is somehow part of their meaning. Nor does it seem a plausible scientific theory about such objects because of its ad hoc character and its lack of predictive value. If the notion of God is dropped, however, the philosopher is left with the phenomenalistic theory that such objects are collections of appearances. But phenomenalism also has serious difficulties; in particular, it cannot in the end account for the difference between real objects and illusions because it cannot provide an account of the difference between circumstances in which perceptions are veridical and those in which they are not.

The other variety of immaterialism, called absolute Idealism, derives from certain doctrines of Immanuel Kant and of the classical German Idealists who followed him—Johann Fichte, Friedrich Schelling, and G.W.F. Hegel—concerning the fundamental dependence of reality on mind or spirit in general. Among the several philosophers who have defended this view, there was, at the turn of the 20th century, F.H. Bradley, whose *Appearance and Reality* (1893; 2nd edition 1897) comprises its most systematic exposition and defense. Bradley denied that a plurality of minds exists and insisted that there is only one infinite Mind, Idea, or Experience that comprehends all of existence within it.

Neutral theories. Another important view has been that neither the mental nor the physical is really fundamental; each is an aspect of some underlying reality that is neither mental nor physical but neutral between them. There are many variants of such a view. Spinoza, a 17th-century Rationalist, held that the underlying substance, which encompassed all of reality and which he called God or Nature, had both thinking (the mental) and extension (the material) as attributes. A modern version of this position is that of Peter Strawson, a leading philosopher of the Oxford "ordinary language" school, who differs from Spinoza in holding that there is a multiplicity of substances, some of which are purely material and some of which are persons (thus he is not really a monist). Strawson conceives of persons as substances whose nature is to have both mental and physical attributes. Thus, one and the same substance can have both qualities, and the difference between the mental and the physical is conceived as a basic difference between the qualities.

Spinoza's substance and attributes

A different approach was suggested in some of Hume's writings and diversely stated by the Pragmatist William James and by various Positivists (Ernst Mach, Rudolf Carnap, and A.J. Ayer). They postulate a number of particular entities, experiences, that go to make up minds when they are related in certain ways, as by the laws of association and memory, and that go to make up bodies when the entities are related in other ways, as by the laws of perspective. Thus, a person's mind is conceived to be just the collection of his experiences, whereas a physical object is conceived to be just the collection of experiences that people can have of it. Here the difference between the mental and physical consists in the different kinds of relations obtaining between the neutral particulars, experiences.

Recently, it has been suggested by certain Linguistic philosophers that the difference between mind and body lies in two different kinds of language or conceptual systems: the physicalistic-conceptual language, on the one hand, with its spatiotemporal terms, and person-talk, on the other, with its reference to norms for assessing the rationality, moral responsibility, and ethical value of human actions.

Existentialist and Phenomenological philosophers have expressed similar conclusions, supported not so much by linguistic considerations as by general observations of man's condition as a being in the world, with a body, which he experiences and which, by its nature, affects his experience. Man can be viewed as a spatiotemporal aggregate, an object for observation, study, and manipulation, an instance of the laws of nature. But man can also be viewed as a self-moved mover, a being who alters himself and the world through the decisions he makes, who determines values and invests things with those values, who can make his life and his world according to the values that he determines, and who, in the end, can negate his values and even terminate his life by choice. Here, on the mind–body problem, the philosopher finds surprising similarities between some Analytic philosophers of the English-speaking world and the more speculative continental philosophers.

Existentialist and Phenomenological views

THE ANALYSIS OF MENTAL PHENOMENA

When the specific phenomena that go to make up the mental are considered, one finds that they all raise philosophical issues, only some of which can be sketched here. Mental phenomena are traditionally divided into three areas: the cognitive, which is concerned with knowledge; the affective, with feeling; and the volitional, with action. It is no longer believed that this division reflects the three so-called basic faculties that comprise the mind; nevertheless, as a very rough classification, it provides a convenient approach to the variety of mental phenomena.

The cognitive. Many philosophers from Plato on have taken man's ability to know as the characteristic distinguishing him from all other animals. The very name of his species, *Homo sapiens*, means "man the knower."

If one asks what knowledge is, he has raised the central problem of a major field of philosophy, epistemology. But there is also a very important psychological aspect to knowledge, and that is where the philosophy of mind becomes relevant. It is often claimed, for example, that knowing that something is so entails believing that it is so; and the nature of belief lies clearly within the province of the philosophy of mind. Since a person does not lose a belief when he is not consciously attending to it (for example, when he is asleep), the approach to belief most in favour today is to treat it as a disposition, which, like all such, comes to open expression only sporadically. Other psychological phenomena falling within the area of the cognitive are attention, sense perception, understanding, memory, inferring, and doubting. The view that each of these requires a subjective experience has been effectively refuted in the writings of Ludwig Wittgenstein, one of the seminal thinkers of modern Linguistic Analysis. Remembering that the oven is still turned on may consist in nothing but getting up in the middle of a conversation, going over to the oven, and turning it off, all the while animatedly continuing the conversation. But exactly why this is called "remembering that the oven is still on" is not clear. Perhaps the best that can be said is that there are analogies between such instances of remembering and other, more self-conscious instances. It is the task of the philosophy of mind to examine, classify, and analyze the relations among such phenomena.

Psychological aspect and phenomena

The affective. Man has not only the capacity to know but also the capacity to respond emotionally to what he knows. A man may not only believe that some event will occur, but he may also dread it or welcome it. Concerning the things that a person knows, he may approve or disapprove, love or hate, pity or envy, enjoy or abhor them. Here, although the subjective experience often plays an important role, it clearly is not the whole story. To enjoy doing something, as has been pointed out, is not to do the thing *and also* undergo a series of experiences of enjoyment; it may simply be to do the thing when circumstances permit and make efforts to avoid its cessation or interruption. But a disposition-to-behave approach will be less successful for other affective phenomena. For example, people have feelings about the past—regret, nostalgia, pride—feelings in which future behaviour plays a relatively minor role.

All of the affective phenomena so far considered have

the property of intentionality, of being directed toward an object. It is clear, however, that this is not a sufficient condition for defining the affective, since it marks out too broad a scope—including, for example, believing, which, though intentional, falls not within the affective but within the cognitive. But neither does intentionality seem to be a necessary condition of the affective. Moods like depression, anxiety, or joviality may not have any specific object, though it is sometimes replied that such emotions take as objects anything the person happens to think of. Sensations also do not seem to be intentional, even though they are usually classified as affects. One view is that sensations are really cognitions—the awareness of some bodily disturbance. The difficulty in trying to decide whether a sensation is an affect or a cognition further illustrates the inadequacy of the classification of mental phenomena into the cognitive, affective, and volitional.

The volitional. Intellect and emotion often come to expression in volition and action, important topics in the philosophy of mind—topics that comprise such concepts as motive, desire and purpose, deliberation, deciding, intention, trying, and action, both voluntary and involuntary.

There is a rough distinction to be made between the things that happen to a person and the things he does or makes to happen. If a person slips on the ice, it is something that happens to him; if he walks on the ice, it is something that he does. "Henry slid on the ice" is ambiguous: it may report something that happened to Henry or something that Henry did, depending upon the meaning. In this example, the observable event may be the same: from a photograph of Henry sliding on the ice one may not be able to tell which it is. The problem of action is primarily to understand this distinction and its ramifications. Wittgenstein once put the question this way: "And the problem arises: what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?"

Wittgenstein's question on action

There are a number of different answers: (1) Actions are events produced by causes of certain sorts—volitions or acts of will according to some theories; beliefs and desires under other theories; and simply persons or agents in yet another theory. (2) Actions are events that are "caused" in a special sense; they have a teleological rather than an efficient or mechanical cause, or an immanent (or originating) cause rather than one that is merely a reaction to, or modification of, an action coming from some other source. (3) Actions are events that are properly characterized and assessed in terms of rules of conduct, or principles of rational and ethical behaviour, and for which the agent is held responsible, liable, accountable, to be praised or blamed, rewarded or punished.

Any theory of action is expected to throw light on the issue of free will, a matter of great importance for ethical theory. If the philosopher holds that free will is compatible with determinism, any of the views above will allow for free will. Even if he holds that an action is not free if it has causes that eventually lie outside the agent, his view will be compatible with the various views of action unless he holds the version of (1)—that an action is an event produced by volitions or beliefs and desires—and also holds the additional thesis (2)—that volitions or beliefs and desires themselves have causes that lie outside the agent. Only then will there be no freedom of the will.

SOME METAPHYSICAL AND EPISTEMOLOGICAL ISSUES

Personal identity through time. A person, as he goes through life, changes in many ways; but he remains the person that he was. He is that person who was born on a certain day, that person who graduated 23rd in a particular high school class, who married on a certain date in a certain place; he has a particular identity through time. It is difficult to state what exactly it is that makes a person one and the same self through time.

Bodily versus psychological criteria

An obvious starting point is the fact that throughout a person's natural life he has the same body and that this is what makes him one and the same person throughout a particular period of time. But there are difficulties in this view. First, since the body cells are constantly being replaced and in some instances whole organs are transplanted, it is not clear what makes a particular body identical with a body that existed, say, 20 years earlier. A second difficulty arises through the possibility of brain transplants; if two brains were interchanged, in all likelihood there would be a systematic interchange of memory, beliefs, personality and character traits, skills, and habits of thought and action. Such a transplant would incline one to say that not merely a small portion of each body had been interchanged but two people as well; for one also takes as a criterion of personal identity similarities through time of memory, personality, skills, and habits. After all, a man is often willing to say that this is the same person who did something in the past, not on the basis of knowing that it is the same body but on a quite different basis—that the person recounts the past situation with great accuracy, exhibits similar personal reactions, and displays the same skills.

Because two different kinds of criteria, bodily and psychological, are used for determining personal identity, it is possible to imagine instances of conflict in which the criterion of bodily identity would indicate that it is a different person but the psychological criteria would indicate that it is the same person and vice versa. The Bohemian novelist Franz Kafka, known for his nightmarish works, in his short story *Die Verwandlung* (1915; Eng. trans., *The Metamorphosis*, 1937) tells of a person who awoke one morning to find, to his horror, that he had the body of a large insect. Although his family accepted his conclusion that he was the same person even though he had an entirely different body, others would have disagreed. There is still, in fact, considerable disagreement among scholars on this whole issue—on how to state precisely the bodily criterion; on whether there is a psychological criterion as well and, if so, how it is to be formulated; on what is the basic criterion of personal identity; and on what to say about instances in which criteria conflict.

Personal immortality. Many people believe that when the human body ceases to be a living system, there is not total annihilation of the person but that in some respect the person continues to exist. The philosopher of mind can put aside the various watered-down versions of immortality in which the person continues to exist in the remote sense that people still remember him or his works or that his influence continues through history. The philosopher does look with interest, however, on the claim that there is an immortality in which a person, in his survival, meets the psychological criteria of personal identity, of inheritance of memory, beliefs, habits, and personality characteristics.

Problems of meaning and fact

It is clear that a person's view of immortality will be affected by opinions that he holds about the relation of mind to matter. Given the versions of Materialism that urge the elimination of mental terms or their definition in bodily terms or that take bodily identity to be the basic criterion of personal identity, the very notion of survival after death is completely unintelligible. Many philosophers, however, reason that, since the notion of survival is intelligible, such versions of Materialism cannot be accepted. Central-state or identity theorists would admit survival to be an intelligible notion but would view it, like lightning without electricity, as something that never happens; they would thus be in agreement with many dualists, in particular epiphenomenalists and psychophysical parallelists. Even an interactionist would be free to accept or reject survival, as would a neutral monist.

If a philosopher holds that survival is an intelligible notion, he is still left with the further question of whether it ever happens. In the past there have been various a priori arguments for survival after death. Arguments based upon the nature of the self, such as its indivisibility, can be found in Plato's *Phaedo*. Kant argued that man's moral principles require survival as a postulate. But among those today who hold that survival is intelligible, it is widely agreed that a priori arguments will not do. If there is survival, they say, it is a contingent fact and not a necessity; one must thus look to empirical data for guidance. A survey of the evidence shows that the case against survival, though strong, is by no means conclusive. One

thing is clear: if there is survival, the survivors can theoretically give firsthand testimony to it, whereas if there is no survival, there will be no one to give such testimony.

Knowledge of other minds. An important problem in the theory of knowledge has been the status of the belief in other minds, the belief that one's own consciousness is not the only consciousness in existence. Though few, if any, sane persons have seriously accepted solipsism (the view that one's own is the only consciousness), the grounds for rejecting it are not at all clear.

Again, as with the problems of personal identity and personal survival, one's view of the relation of mind to matter is relevant. On various Materialistic views, the problem reduces itself to that of justifying the belief in an external world that contains other bodies of the appropriate sort and with the appropriate behaviour. But for dualists and immaterialists, who hold that mental phenomena are something irreducibly different from the physical, there is the further question of whether that something is unique to oneself or whether there are other instances of it in the world.

Direct, transcendental, analogical approaches

Some scholars claim that men sometimes have direct awareness of the conscious states of others, either in telepathic experiences, moments of empathy, or even in everyday social intercourse. There is, moreover, a transcendental argument, found in Kant and recently defended by Strawson, holding that, unless a person could be confident of the existence of other minds, he could not be confident of the existence of his own mind. A quite different line of reasoning, the so-called argument from analogy, is based upon the similarities between one's own body and its behaviour, on the one hand, and other human bodies and their behaviour on the other. To pursue the argument, since a mind is known to be associated with one's own body, it is reasonable to conclude (by analogy) that another mind is also associated with the body of another person. Finally, there is the view that the best way to explain the complex behaviour of other bodies, especially their ability to behave rationally, and in particular to speak and communicate information, is to postulate other minds at work.

None of these arguments compels strong conviction, certainly not the degree of conviction that all men feel concerning the existence of other minds. Whether stronger arguments will be found, whether philosophers must admit that there is a considerable amount of faith required here, or whether they will reformulate their concepts of the mind in a more Materialistic way to bring them in closer accord with observable data remains to be seen.

BIBLIOGRAPHY. Although the philosophy of mind has only recently emerged as a special field, many of its topics have long been discussed. Various formulations of Materialism can be found in the writings of the pre-Socratic philosophers, especially LEUCIPPUS and DEMOCRITUS. PLATO'S dualism and his views on the nature of the soul appear in the *Phaedo, Timaeus, Phaedrus,* and books iv and x of the *Republic.* ARISTOTLE devoted a whole treatise, *De Anima* (*On the Soul*), to the subject, expressing a qualified Materialism. THOMAS HOBBES, *Leviathan* (1651), was radically Materialistic; whereas RENE DESCARTES, *Meditationes de prima philosophia,* 2nd ed. (1642), and other writings, present a classic interactionist dualism. Dualism of the parallelist variety can be found in the writings of GOTTFRIED LEIBNIZ. BENEDICT DE SPINOZA, in his *Ethica* (1677), rejected both Materialism and dualism, expounding a double aspect theory. GEORGE BERKELEY, *Treatise Concerning the Principles of Human Knowledge* (1710), provides the classic statement and defense of Idealism (in his word, Immaterialism).

The contemporary interest in the philosophy of mind is largely a result of GILBERT RYLE'S brilliant and polemical *Concept of Mind* (1949). Also of great importance was the wholly new approach to all of philosophy taken by LUDWIG WITTGENSTEIN in his *Philosophical Investigations* (1953), which contains many specific applications to the philosophy of mind. There followed a spate of journal articles, some of the best of which are in the following anthologies: ALAN ROSS ANDERSON (ed.), *Minds and Machines* (1964); MYLES BRAND (ed.), *The Nature of Human Action* (1970); VERE C. CHAPPELL (ed.), *The Philosophy of Mind* (1962); ANTONY FLEW (ed.), *Body, Mind, and Death* (1964); DONALD F. GUSTAFSON (ed.), *Essays in Philosophical Psychology* (1964); STUART HAMPSHIRE (ed.), *Philosophy of Mind* (1966); SIDNEY HOOK (ed.), *Dimensions of Mind* (1961); HAROLD MORICK (ed.), *Introduction to the Philosophy of Mind* (1970); GEORGE PITCHER (ed.), *Wittgenstein* (1966); G.N.A. VESEY (ed.), *Body and Mind* (1964); ALAN R. WHITE (ed.), *The Philosophy of Action* (1968); and OSCAR P. WOOD and GEORGE PITCHER (eds.), *Ryle* (1970). Selections written from Existentialist and Phenomenological perspectives are in STUART F. SPICKER (ed.), *The Philosophy of the Body* (1970). Introductory writings that include original contributions as well as balanced assessments are GERALD E. MYERS, *Self* (1969); ALAN R. WHITE, *The Philosophy of Mind* (1967); and JEROME A. SHAFFER, *Philosophy of Mind* (1968). For analyses of particular mental phenomena, see G.E.M. ANSCOMBE, *Intention* (1957); D.M. ARMSTRONG, *Bodily Sensations* (1962); P.T. GEACH, *Mental Acts* (1957); STUART HAMPSHIRE, *Thought and Action* (1959); ANTHONY KENNY, *Action, Emotion and Will* (1963); A.C. MACINTYRE, *The Unconscious* (1958); NORMAN MALCOLM, *Dreaming* (1959); A.I. MELDEN, *Free Action* (1961); R.S. PETERS, *The Concept of Motivation,* 2nd ed. (1969); H.H. PRICE, *Thinking and Experience,* 2nd ed. (1969); CHARLES TAYLOR, *The Explanation of Behaviour* (1964); RICHARD TAYLOR, *Action and Purpose* (1966); and A.R. WHITE, *Attention* (1964). Concerning some of the metaphysical and epistemological issues, A.J. AYER, *The Concept of a Person,* (1963); and SYDNEY SHOEMAKER, *Self-Knowledge and Self-Identity* (1963), deal with the problem of personal identity. Issues concerning the possibility of an afterlife are discussed in C.J. DUCASSE, *Nature, Mind and Death* (1951) and *The Belief in a Life After Death* (1961); and by C.D. BROAD in *Religion, Philosophy and Psychical Research* (1953) and *Lectures on Psychical Research* (1962). The problem of one's knowledge of other minds is explored in HAROLD MORICK (ed.), *Wittgenstein and the Problem of Other Minds* (1967); ALVIN PLANTINGA, *God and Other Minds* (1967); and JOHN WISDOM, *Other Minds* (1952).

(J.A.Sh.)

Minerals

A mineral is a solid formed by natural processes; generally it is crystalline and inorganic. Approximately 50 new species are added each year to the presently known 3,000 types. Minerals of recognizable shape, colour, and other properties are collected in museums and by private collectors; particularly beautiful rare minerals are prized as gems. These specimens are very atypical, however, and most minerals in nature occur inconspicuously as the components of rocks and soil.

The fundamental nature of minerals is revealed by complex chemical and physical techniques. The internal crystal structure and the chemical composition are characteristic properties of each mineral, and they permit definitive characterization and an understanding of the processes controlling the growth and subsequent history of minerals. The obvious surface properties of minerals such as shape and colour tend to depend on the fundamental properties but are often confused by accidental features, such as those produced by exposure to weathering, that hinder or prohibit identification. The classification of minerals historically began from consideration of surface properties, but it is now based on internal properties.

Synthetic equivalents of some minerals are produced industrially (*e.g.*, rubies and diamonds), and certain biological products also consist of minerals (*e.g.*, apatite in bones and teeth; phosphate and oxalate minerals in biological stones and certain tumours). Many industrial processes depend directly on mineralogical reactions (*e.g.*, crystallization of slag in blast furnaces, and of concrete and cement), and all metallic processes depend ultimately on minerals as a source. Mineralogy plays a role in all the geological sciences and in many of the other physical sciences. For example, agriculture depends on the maintenance of appropriate minerals in soils, and the distribution of ice is critical to meteorological phenomena. In fact the condensation of gases into minerals is proposed as a major accretion process in the universe—the ultimate origin of the Earth would be the condensation of solar matter.

This article provides a summary of the nature, properties, and occurrence of minerals and of mineral associations and origins. For more specific information on rock-forming minerals and other important mineral groups, see the separate articles: AMPHIBOLES; CARBONATE MIN-

ERALS; CLAY MINERALS; FELDSPARS; FELDSPATHOIDS; MICAS; NATIVE ELEMENTS; OLIVINES; PHOSPHORITES; PYROXENES; SILICA MINERALS; SILICATE MINERALS; SULFIDE MINERALS; and ZEOLITES. The crystal structure of minerals is presented in greater detail in CRYSTALLOGRAPHY, and the occurrence of minerals on and within the Earth is further treated in ORE DEPOSITS; IGNEOUS ROCKS; SEDIMENTARY ROCKS; METAMORPHIC ROCKS; and EARTH, STRUCTURE AND COMPOSITION OF. See GEMSTONES for additional information on precious minerals; for discussions of the conditions surrounding mineral formation in nature, see the articles GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES; and GEOCHEMICAL EQUILIBRIA AT LOW TEMPERATURES AND PRESSURES.

THE NATURE OF MINERALS

Crystal structure and chemistry. All minerals have regular arrangements (lattices) of atoms that have some geometrical symmetry. For each repeating unit cell, the smallest unit of a mineral that retains its basic structure and composition, there are sets of sites occupied by atoms at distances determined by the chemical bonding of the constituents involved. In quartz, for example, each silicon atom is surrounded by four oxygen atoms lying at the corners of a tetrahedron at a distance of 1.6 angstroms (one angstrom unit equals 10^{-8} centimetre). Each oxygen atom is close to two silicon atoms, and the whole atomic assemblage can be considered as an infinite framework of silicon–oxygen bonds. The crystal structure can be determined by X-ray-diffraction methods—*i.e.*, subjecting a crystal to radiation and employing Bragg's law, which relates the wavelength of radiation and the angle of diffraction of the radiation to the internal spacing of crystallographic planes. The cell dimensions, symmetry, and diffraction intensities provide unique identification.

From the geometrical distribution of atoms in the unit cell, and from the general principles of chemical bonding, the type of chemical bonding can be envisaged and used in the interpretation of mineralogic properties.

This topic is treated more fully in the article CHEMICAL BONDING, and the four principal types that are of mineralogical concern will be but briefly described here. The first of these is ionic bonding, the permanent transfer of electrons from atoms to form ions, which undergo electrical attraction and repulsion. A second type is covalent bonding, in which clouds (groups) of electrons are shared between atoms. Metallic bonding involves the formation of an electron "gas" throughout an entire crystal. Van der Waals bonding, the fourth type, is a weak attractive force resulting from temporarily induced distortions of electron clouds. Most bonds are a mixture of more than one type.

Copper and gold typify metallic bonding and have atoms lying at the faces and corners of the cubic unit cell, so that each atom has 12 neighbours. Arsenic is only semimetallic; its structure is similar to that of copper or gold but is distorted so that it has three nearest neighbours to which it is partly bonded covalently.

Halite (common salt) typifies ionic bonding. The sodium and chlorine atoms lie in interpenetrating lattices, and the symmetry is such that each atom has six near neighbours.

Diamond is a prime example of covalent bonding. Carbon atoms lie in a tetrahedral arrangement with covalent bonds linking all atoms into a single edifice. In graphite, however, which also consists solely of carbon, the carbon atoms lie in layers or planes of hexagonal symmetry; they are bonded by covalent forces within the layer but by weak van der Waals forces between the layers.

Irrespective of the type of bonding, the crystal structure is based on fascinating geometrical considerations. Even the most complex mineral structure has a simple underlying pattern that can be analyzed by using the principles of symmetry and topology (the study of geometric properties that are unchanged by continuous transformation). Complex structures are always distorted in response to subtle chemical forces not completely understood.

The condensed nature of crystalline minerals results in close packing of near-spherical units in most ionic and metallic forms. Covalent forces require special bond angles that are inconsistent with simple concepts of close packing. On Earth, oxygen (O^{2-}) is the dominant anion (negatively charged ion), and it is larger than most of the cations (positively charged ions). The outer part of the Earth may be said roughly to consist of close-packed oxygen ions with the interstices occupied by cations.

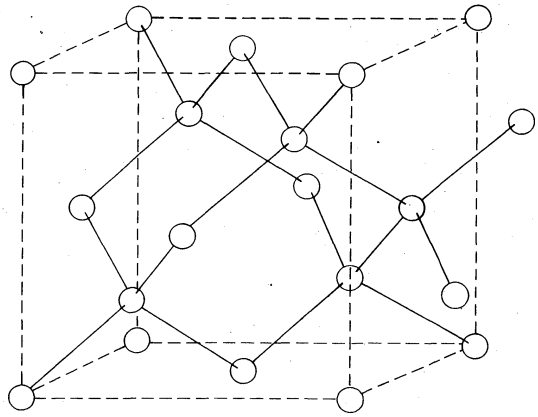


Figure 1: Structure of diamond. Positions of carbon atoms are shown by circles, direction of strong covalent bonds by the solid lines; dashed lines show the unit cell.

Close packing in mineral structures

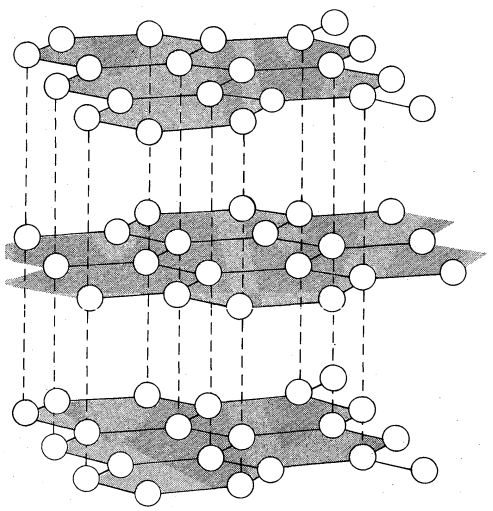


Figure 2: Structure of graphite. Each atom is bound triangularly to three neighbours, producing layers.

The two simplest close-packed arrangements are cubic close packing and hexagonal close packing. The former is illustrated by the structure of copper and gold. There are two types of interstices, one type between six atoms (*i.e.*, six-coordinated) and one type between four. The halite structure, for example, consists of small sodium cations (Na^+) that occupy the six-coordinated holes between the large chlorine anions (Cl^-). The substance magnesium oxide (MgO) may exist deep in the Earth with the

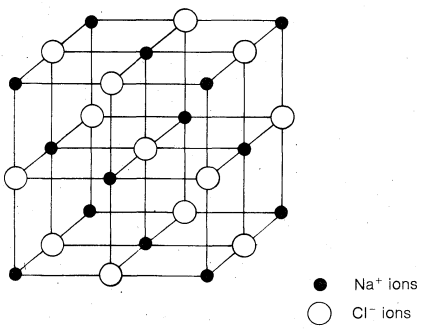


Figure 3: Structure of halite, NaCl.

halite-type structure composed of large oxygen anions (O^{2-}) and small magnesium cations (Mg^{2+}). The second type of interstice can be envisaged from the diamond structure. If every alternate atom grows while the other half diminishes in size, the large ones ultimately will come into contact, and the small ones will just fill half the holes, leaving the other half empty.

Effect of ionic size

Ions have different inherent sizes, and the actual crystal structures result from minimizing the electrical potential.

Ward's Natural Science Establishment, Inc.

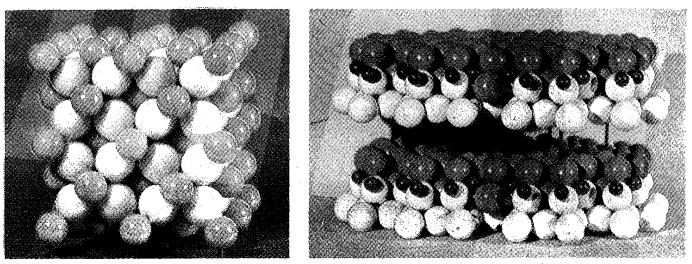


Figure 4: *Models of two types of crystal structures of minerals.* (Left) Crystal structure of fluorite, CaF_2, an example of strong ionic bonding. Red balls represent calcium and yellow balls fluorine ions. (Right) The structure of kaolinite, $Al_4Si_4O_{10}(OH)_8$; closely packed atoms in the composite slabs are held together by strong covalent and ionic forces. The slabs, however, are electrically neutral and only weakly held to one another by van der Waals bonding. Yellow balls represent oxygen atoms arranged in tetrahedra; silicon atoms occur within the tetrahedral pores and are not visible. Small brown balls are aluminum and blue are hydroxyl ions.

The complex structure of kaolinite, a clay mineral and aluminum silicate of composition $Al_2Si_2O_5(OH)_4$, serves as an example. Tiny silicon ions (Si^{4+}) lie between four large oxygen ions (O^{2-}) linked to form a hexagonal sheet. Large hydroxyl ions (OH^-) form another hexagonal sheet and also occur enmeshed with the oxygen sheet. Small aluminum ions (Al^{3+}) bond to four OH^- and two O^{2-}. The $Al_4Si_4O_{10}(OH)_8$ "sandwich" is electrically neutral and is bonded to adjacent sandwiches by weak and ionic van der Waals forces.

Many groups of minerals have the same topology in their crystal structure even though the actual cell dimensions are slightly different. This arises when the ratio of the ionic radii is within the tolerance for that particular structure type and when the ions have the appropriate charges. Examples of such isostructural minerals are calcite ($CaCO_3$), siderite ($FeCO_3$) and rhodochrosite ($MnCO_3$); halite (NaCl), sylvite (KCl) and periclase (MgO); diopside ($CaMgSi_2O_6$) and aegirine ($NaFeSi_2O_6$).

Polymorphic changes

Many chemical compositions yield more than one structural type, and these are known as polymorphs. Thus SiO_2 yields a host of polymorphs, including quartz, tridymite, cristobalite, coesite, and stishovite. The $CaCO_3$ composition occurs as calcite and aragonite. Under thermodynamic equilibrium, each polymorph would occur only for a specific range of temperature, pressure, and composition; however, a polymorph may grow metastably, or persist metastably—*i.e.*, resist change in response to small disturbances despite the fact that some other phase would be wholly stable under the same conditions —as a result of energy barriers to atomic movement. Other polymorphs are related merely by change of interatomic angles culminating in a symmetry change: the so-called high quartz changes to low quartz at 573° C (1,063° F), accompanied by a symmetry change from hexagonal to trigonal (see SILICA MINERALS). Other polymorphic changes involve the atomic distribution between sites. Sanidine and microcline, two of the feldspars *(q.v.)*, are polymorphs of potash feldspar ($KAlSi_3O_8$) and differ in that microcline has the Al and Si atoms segregated each into its own sites, whereas sanidine has a random distribution.

An ideal crystal structure would yield a chemical formula with exact ratios: $CaMgSi_2O_6$, for example. Packing tolerance exists in most structures, however, and extensive atomic substitution (often called solid solution) is permitted. Solid solution is essentially the compositional

Solid-solution series

variation of a single crystalline phase between finite limits without the appearance of a second phase. If the atomic substitution causes only minor structural distortions, which is the case for the common substitutions of silica (Si) and alumina (Al), then atomic substitution at temperatures above absolute zero is favoured on theoretical grounds. The extent of solid solution normally increases with temperature and may be complete. With falling temperature, dissolution or exsolution into two intergrown mineral phases may occur, or ordered structures may form.

Some minerals have unoccupied sites. Sulfides provide the best examples: pyrrhotite (iron sulfide) has a range of composition in which iron may be present in quantities equal to or down to eight tenths the amount of sulfur present. The formula is: $F_{1-x}S$, in which x varies up to 0.2; only in meteorites and in lunar samples is the end-member troilite (FeS) typical.

In general, rock-forming minerals contain a bit of everything ranging from major constituents (over 1 percent), minor constituents (below 1 percent), and trace elements (parts per million or less). Only by combination of subtle analytical methods can the complete chemical analysis be obtained. The principles underlying the distribution of chemical elements in and between minerals are covered in ELEMENTS, GEOCHEMICAL DISTRIBUTION OF.

Morphology. The shape of minerals obeys the rules of crystallography, which developed historically from the study of minerals. Museum collections naturally emphasize well-shaped crystals grown in cavities, but most minerals interfere with each other's growth in rocks. In analogous fashion, snowflakes contrast with block ice. Well-shaped crystals have characteristic angles between faces that are related to the dimensions and symmetry of the internal structure. Ideally, each group of faces related by the symmetry would be equally developed, thus providing the form; a set of six faces on a cube of galena or the prism faces on quartz are examples. Most minerals show

By courtesy of (top left) Joseph and Helen Guetterman, Belleville, Illinois, (bottom left) the Field Museum of Natural History, Chicago, (bottom right) the Harvard Collection; photographs, (top left, bottom left) John H. Gerard—EB Inc., (top right) Floyd R. Getsinger, (bottom right) Benjamin M. Shaub

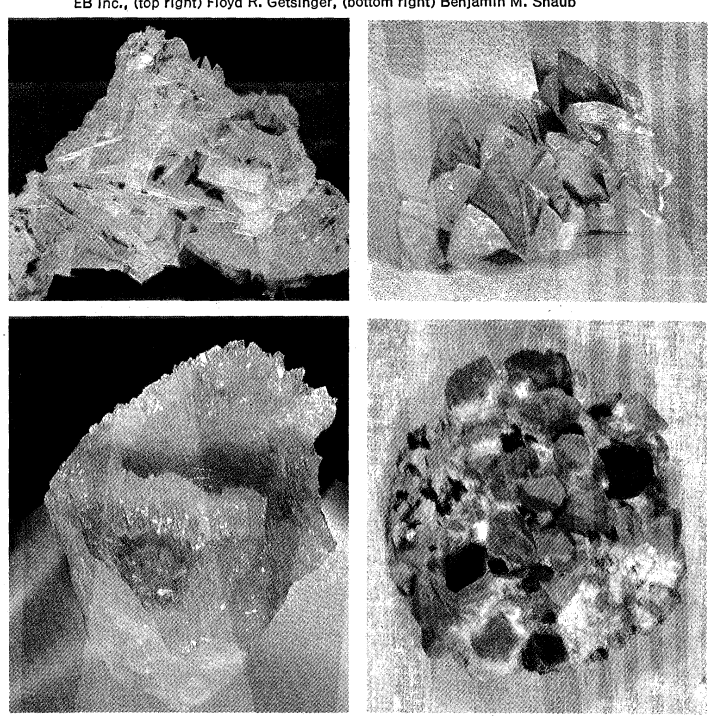


Figure 5: *Representative minerals that sometimes occur as euhedral crystals.* Although in the minority, they receive the most attention because of their beauty and because they reflect the basic crystal form and structure of a particular mineral. (Top left) Wulfenite, $PbMoO_4$, from Mexico. (Top right) Calcite, $CaCO_3$, from Chihuahua, Mexico. (Bottom left) Rose quartz from Minas Gerais state, Brazil. (Bottom right) Microcline feldspar, $KAlSi_3O_8$, variety amazon stone, with smoky quartz, from Colorado.

several types of faces, and the relative development of each form yields the habit. Quartz from volcanic rocks usually grows as a hexagonal pyramid, whereas quartz from low-temperature veins is dominated by a hexagonal prism. Some minerals almost always show the same habit (*e.g.*, the icosahedral habit of garnet), but others (*e.g.*, feldspar) show major variations from one locality to another that are only partly correlative with the temperature and other obvious physical and chemical factors.

Certain crystals have pseudo-symmetry and during their growth incoming atoms may adopt a new configuration rather than adopting the old one. Such multiple crystals are called twins, even though many twins consist of more than two units. Easily recognized twinned minerals are staurolite, pyrite (the "iron cross"), sanidine (Carlsbad twin), aragonite, and spinel; in thin section, the twinning of microcline and plagioclase is often characteristic. Twinning also results from deformation and symmetry inversion as a consequence of changed pressure and temperature conditions.

Classification in terms of structure and composition. Historically, mineral names were assigned haphazardly, with many errors and much redundancy. In spite of attempts to develop rational systems, the present conglomeration of proper names is useful, reasonably consistent, and fascinating from the etymological viewpoint. Since the 19th century, better understanding of crystal structure and chemical composition has led to slow refinement of mineral nomenclature. Since 1960, an international committee of nomenclature has reviewed new proposals for mineral names and attempted to remove redundancies. Generally speaking, a new mineral name is adopted only upon satisfactory evidence of (1) a significant change of chemical composition in a known structure type or (2) a completely new structure type of any composition. The discoverer has the right to propose a proper name (ending in "-ite") consistent with absence of confusion with an existing name. In complete solid solution (atomic substitution) series, a name is usually accepted for each of the constituents present. Gold and silver are such a pair, for example, and, if gold dominates, then gold is the name assigned for (Au, Ag); the name silver would be used for (Ag,Au). Minor constituents are represented by adjectival modifiers—*e.g.*, argentian tetrahedrite for minor Ag in $Cu_{12}Sb_4S_{13}$. Exceptions arise from such special features as changes of symmetry (*e.g.*, sanidine and anorthoclase in the feldspar group) and from historical precedent (*e.g.*, many subdivisions of olivine, plagioclase, and pyroxene). Names should be useful, reflecting the principal properties, and particularly complex mineral groups need many names even if they break some desired artificial system. Several thousand names are in common use, modified slightly from one language to another.

Mineral nomenclature

The first recorded classification was by the Greek philosopher Theophrastus. Subsequent schemes utilized mostly physical properties, but increasing chemical knowledge led to classifications based on chemical composition, such as that by a Swedish chemist, Axel Fredrik Cronstedt, in 1758. Major advances in both chemistry and mineralogy permitted James Dwight Dana, an American mineralogist, to develop his *System of Mineralogy* in 1837; the later editions of this work served as a primary reference source. Although ignorant of crystal structure, Dana proposed a system that correlates well with present knowledge. In this article, minerals are classified into 13 groups; those marked by an asterisk (*) are covered or partly covered in separate articles: *(1) native elements; *(2) sulfides, arsenides, and antimonides; (3) sulfosalts; (4) oxides and hydroxides; (5) halides; *(6) carbonates; (7) nitrates and iodates; (8) borates; (9) sulfates; (10) chromates; *(11) phosphates, vanadates, and arsenates; (12) tungstates and molybdates; and *(13) silicates. In this classification, the first group concerns single-element minerals with metallic or covalent bonding; the second two groups are principally metallic; the next two groups are principally minerals with ionic bonds, generally of equal strength; and the remaining groups have complex structures in which tightly bonded groups (*e.g.*, carbonate) are joined to ions by weaker bonds. The mineral groups are dominated by oxygen, which accords with its great geochemical abundance. In each group, isostructural minerals (*i.e.*, those with similar crystallographic structures) are placed together, and minerals are arranged on the basis of structural properties.

Dana's classification system

In the discussions that follow it is important to bear in mind that the Earth consists of three concentric shells—the crust, mantle, and core—which differ in chemical composition and in the conditions of pressure and temperature that prevail. Certain minerals that are known to form only under conditions of high temperature and pressure (*e.g.*, diamond) cannot form at the Earth's surface; others, which do form under surface conditions (*e.g.*, evaporites), cannot be expected to occur at depth in the mantle. For a detailed exposition of conditions in the Earth's crust, mantle, and core, see EARTH, STRUCTURE AND COMPOSITION OF.

Native elements. These are rare at the Earth's surface because most chemical elements combine rapidly with oxygen, sulfur, and halogens. A few metallic elements are highly resistant and form stable native metals (platinum, gold), and a few survive in special environments, although they are unstable in the presence of atmospheric gases or solutions (copper, mercury, arsenic, antimony, iron). Sulfur and carbon (as diamond and graphite) are metastable with respect to oxygen but persist in special environments. Diamond is truly stable only at pressures that are equivalent to depths of several hundred kilometres in the Earth, and its persistence testifies to the difficulty of the transformation to graphite at the Earth's surface. Iron characteristically occurs in reducing environments (where oxygen is lacking), as typified in certain meteorites and in the Moon. The crystal structures of the native elements are simple. Carbides, nitrides, and phosphides are normally classified with native elements, since the carbon (C), nitrogen (N), and phosphorus (P) atoms merely fill the interstices between the metal atoms. See further NATIVE ELEMENTS.

Sulfides, arsenides, and antimonides. During melting of planetary material, sulfur combines with metals such as iron to form a liquid that is immiscible, or will not mix, with a silicate liquid. The sulfide–metal melt produces the native iron and troilite (FeS) that are seen in meteorites and in volcanic rocks from the Moon and are inferred to occur in the Earth's core. During the complex chemical processes of chemical differentiation near the Earth's surface, much of the sulfide is oxidized and hydrolyzed to form sulfates or gases, and some is converted to sulfur. Surviving sulfides, especially those in veins and hydrothermal deposits (those resulting from high-temperature solutions derived from magma), represent important ores of copper, zinc, silver, lead, mercury, and other metals. Arsenic and antimony tend to be associated with sulfur, though their semimetallic nature results in somewhat different properties. For convenience, the complex sulfosalt minerals (treated below) are separated from the sulfides, arsenides, and antimonides, most of which have simple binary (two-element) compositions.

Importance of sulfides as ore minerals

The sulfides, arsenides, and antimonides mostly have semimetallic properties and structures based on packing of small metal atoms between close-packed, large sulfur, arsenic, and antimony atoms. Molybdenite (MoS_2) has a layer structure in which SMoS sandwiches are weakly bonded together, however. Understanding of the chemical bonding in sulfides is still primitive, and most mineralogists believe that the bonding includes metallic, covalent, and ionic components (see further SULFIDE MINERALS).

Sulfosalts. The general chemical formula of the sulfosalts is $X_aY_bS_c$, in which X is silver, copper, or lead; Y is arsenic, antimony, or bismuth; and S is sulfur. All of the 100 or so mineral species are rare and difficult to identify correctly without advanced techniques. The best known species are tetrahedrite-tennantite $[Cu_{12}(Sb,As)_4S_{13}]$; the ruby-silver group pyrargyrite-proustite $[Ag_3(Sb,As)S_3]$; enargite (Cu_3AsS_4); bournonite ($PbCuSbS_3$); and boulangerite ($Pb_5Sb_4S_{11}$). The crystal structures are related to those of simple sulfides

such as galena (PbS) but are modified severely to take account of the chemical differences between sulfur and the semimetallic arsenic, antimony, and bismuth. All have low thermal stability and occur typically in low- to moderate-temperature ore deposits. All occur as exploitable ores, often in association with other economic minerals, and solid solution is common.

Oxides and hydroxides. Ice and quartz are the principal oxides on the Earth's surface. Cuprite (Cu_2O) occurs in oxidized copper deposits, and periclase (MgO) occurs rarely in marble but may be a major constituent of the deep mantle of the Earth.

The spinel group

The spinel group of minerals, with the general formula AB_2O_4 (representing two constituents, *A* and *B*, combined with oxygen, O) contains a wide range of natural minerals and artificial products. The structure consists of cubic close-packed oxygen atoms with *A* and *B* atoms in 4-fold and 6-fold interstices, respectively, giving a structural formula AB_2O_4. To balance the electrical charges (eight negative charges are given by O_4), one divalent cation (II) and two trivalent ones (III) are needed. Normal spinels have II in *A* and two III in *B*, whereas inverse spinels have III in *A* and II + III in *B*; disordered spinels also occur. The distribution of atoms depends on subtle chemical factors and on temperature. Natural spinels contain principally divalent magnesium, iron, manganese, and zinc; trivalent aluminum, iron, and chromium; and quadrivalent titanium, but spinels can be synthesized from many more elements. Spinel of any colour can be synthesized; natural gem spinel (typically ruby colour) mostly comes from gravels in southeastern Asia formed by weathering of metamorphosed aluminous limestone. Spinels occur as octahedral crystals, often exhibit twinning, and may be magnetic, depending upon the properties of the ions in the *A* and *B* sites.

Magnetite and chromite (the only ore of chromium) occur as accessories in basic and ultrabasic rocks (igneous rocks rich in iron and magnesium). Gravity settling produces rich ore bodies. Magnetite weathers and concentrates in beach sands; it also occurs in veins and metamorphosed rocks. The zinc spinels occur as ores in limestones and schists.

Oxides with the general formula R_2O_3 have cations coordinated 6-fold in hexagonally close-packed oxygens. Corundum (Al_2O_3) and hematite (Fe_2O_3) are isostructures. Ilmenite ($FeTiO_3$) has divalent iron and tetravalent titanium atoms alternating in the aluminum sites of corundum.

Corundum tends to occur naturally as hexagonal barrels or plates in high-temperature rocks rich in alumina and low in silica. Its hardness (9 on the Mohs' scale, which ranges from 1 to 10; see below *Hardness*) results in corundum-rich rocks (*e.g.*, emery) being used as abrasives. Most commercial corundum is made synthetically in high-temperature furnaces by fusing bauxite, however. Clear, coloured crystals of natural corundum are prized gems, especially those showing asterism (a starlike effect visible in transmitted or reflected light). Gem-quality synthetic corundum is used in jewelry, clock and instrument bearings, and in crystal lasers. Coloured varieties result from substitution of metals for aluminum: chromium in red ruby and iron or titanium in blue or black sapphire. Asterism results from exsolution into oriented shapes obeying 6-fold symmetry.

Hematite, the most important iron ore, occurs in beds resulting from precipitation of colloidal (very finely dispersed) iron weathered from iron-bearing rocks. Complex chemical and biological reactions occur involving other iron minerals. Massive or powdered hematite is reddish, but museum specimens often occur as metallic flakes or as kidney-shaped masses. Powdered hematite is used in red paint and for polishing (jeweller's rouge).

Ilmenite is an accessory mineral of iron- and magnesium-rich rocks on Earth and in lunar maria. Its high specific gravity (4.7) results in gravity concentrations in igneous rocks and in beach sands; these are exploited as titanium ores.

Rutile and uraninite

The rutile group comprises rutile (TiO_2), cassiterite (SnO_2), pyrolusite (MnO_2), and stishovite (SiO_2). The tetragonal structure—consisting of three mutually perpendicular crystallographic axes, two of which are of equal length and longer or shorter than the third—has six-coordinated cations and three-coordinated oxygens. Rutile and cassiterite occur as stout dark prisms with a Mohs' hardness of about 6½. Rutile is an accessory in high-temperature rocks, and cassiterite occurs in hydrothermal veins. The high densities (4.2 and 7) and resistance to weathering result in concentration in beach sands and alluvium, forming principal ores of titanium and tin. Anatase and brookite are rare polymorphs of rutile. Synthetic rutile is used in jewelry.

Pyrolusite, braunite [$(Mn,Si)_2O_3$], manganite [(MnO(OH)], and psilomelane (hydrated barium manganese oxide) are principal manganese minerals of variable composition and crystal structure composing fine-grained ores of manganese. Such ores form typically in bogs from colloidal solutions derived from weathered rocks or in veins.

Uraninite has an imperfect crystal structure (similar to that of fluorite), and the composition ranges from UO_2 to U_3O_8. It occurs in pegmatites (very coarse-grained igneous bodies) and in veins. Alteration to other uranium minerals (uranyl hydroxides and silicates) and redistribution by sedimentary and solution processes lead to complex chemical variations and geographic distribution of uranium ores. Minor occurrences of rich pitchblende ore are now supplemented by low-grade sources of uranium minerals. Thorium and rare earths are major constituents of uranium minerals.

The hydroxide minerals

All hydroxides are low-temperature minerals typically formed from products of aqueous alteration or from hydrothermal veins. The crystal structures consist of hexagonal layers of cations bonded to six hydroxyls or oxygens. Interlayer bonds are weak, and this results in soft crystals. Brucite [$Mg(OH)_2$] typically occurs in magnesium-rich rocks subjected to low-temperature hydrothermal action or metamorphism (*e.g.*, serpentinite and magnesian marbles). Hydrated iron and aluminum oxides occur everywhere on the Earth's surface and represent the stable forms, with respect to which other iron and aluminum minerals are metastable.

The aluminum ore, bauxite, consists of boehmite [AlO(OH)], diaspore [AlO(OH)], and gibbsite [$Al(OH)_3$] plus iron oxides. Clay and laterite soil consist principally of aluminum hydroxides. Alumina is resistant to leaching and may precipitate as hydroxides that render the soil incapable of sustaining agriculture.

Limonite is an iron ore containing goethite [FeO(OH)] and other hydrated iron oxides of complex structure. The hydrated iron oxides result from weathering of iron minerals in aqueous oxidizing environments. Most deposits are sedimentary, but some specimens occur in veins as crystals or stalactites (deposits suspended from the roofs of caverns).

Halides. Fluorine substitutes in silicates of high-temperature rocks and occurs as fluorite (CaF_2) in veins. Fluorite occurs in cubes, often beautifully coloured (but rarely red) from various trace substituents or lattice defects (the two most common defects in crystals are partial slippage along a crystallographic plane and a void or missing lattice element). The variety Blue John from Derbyshire, England, is used for ornamental vases, but the softness (hardness of 4 on Mohs' scale) precludes its use as a gem. Fluorite is the principal fluorine mineral, and increasing demand as an industrial flux and for fluorinated organic materials of high stability has led to reworking of many old galena deposits, with which fluorite is typically associated. Fluorite and fluorine glasses are valuable for their high optical refringence (refraction of light).

Chlorine, bromine, and iodine tend to end up in solutions, and form a major part of seawater. Evaporation of aqueous bodies leads to sequential precipitation of halide minerals as each successive solubility limit is reached (see further EVAPORITES). The resulting sedimentary deposits are structurally weak and are squeezed into salt domes. Halides also occur in volcanic sublimates.

Halite, or common table salt (NaCl), and sylvite (KCl)

are isostructural minerals that occur as cubes or in massive deposits. Sylvite is more bitter than halite. Although transparent when pure, many specimens are coloured by impurities. Carnallite, a hydrous potassium magnesium chloride, is a deliquescent mineral that occurs at the top of many salt deposits.

Cerargyrite (AgCl) occurs in the oxidized zone of silver deposits as waxlike, ductile masses. Atacamite [$Cu_2(OH)_3Cl$] is a green secondary mineral in the oxidized zone of copper deposits in arid regions. Cryolite, a sodium fluoride (Na_3AlF_6), is a very rare mineral used as an electrolyte for industrial electrolysis of aluminum ores and now mostly synthesized from fluorite.

Calcite, aragonite, and dolomite

Carbonates. At low temperatures under oxidizing conditions, carbon combines with oxygen to form the triangular CO_3 group. Bonding to metal ions such as Ca, Mg, and Zn leads to about 100 carbonate minerals, of which calcite ($CaCO_3$) and dolomite [$CaMg(CO_3)_2$] are the dominant constituents of limestones and dolomites. Calcite and the polymorph aragonite form the skeleton of many organic creatures, particularly marine invertebrates. Many carbonate minerals contain water of hydration. Carbonates are typical weathering products of ores, resulting from reaction with carbonic acid: the copper mineral malachite [$Cu_2CO_3(OH)_2$] is an example (see further CARBONATE MINERALS; and LIMESTONES AND DOLOMITES).

Nitrates and iodates. In the geochemical cycle, nitrogen becomes present in water as nitrate ions—as an end product of volcanic emissions and the breakdown of organic matter. In very arid climates, nitrate is deposited as soda-nitre ($NaNO_3$) and nitre (KNO_3). Because of the very high solubility, other evaporite minerals such as halite are precipitated first. Alkali nitrates are also produced by bacteria and by reaction of organic matter with salts. Only the Chile deposits are really significant commercially. Most nitrate for fertilizer results from synthesis using atmospheric nitrogen.

Some rare iodates (*e.g.*, lautarite [$Ca(IO_3)_2$]) occur together with nitrates.

Borates. The boron ion is very small and exists in both triangular BO_3 and tetrahedral BO_4 groups. Furthermore, the groups condense into clusters, rings, and chains.

Borate minerals occur principally in dried-up basins (playas) fed by waters rich in volcanic emanations; many complexities result from resolution, reprecipitation, and burial of deposits. The principal commercial borates are kernite [$Na_2B_4O_7 \cdot 4H_2O$], borax [$Na_2B_4O_7 \cdot 10H_2O$] of mule-train fame, and colemanite [$Ca_2B_6O_{11} \cdot 5H_2O$]. All are white and have complex crystal structures. Thermal stability is low. Molten borax dissolves many substances, the metallic elements of which can be identified from the colour of the resulting glass bead.

A group of rare borates containing beryllium, aluminum, or magnesium occurs in contact metamorphic deposits (where various rocks are heated at the contact with an igneous rock body intruded at high temperature) and hydrothermal veins. All but one are free of water, but many contain hydroxyl. The rare gem sinhalite ($MgAlBO_4$) has an olivine-like structure with B in fourfold coordination. The ludwigite group occurs in contact-metamorphosed limestones.

Boracite ($Mg_3B_7O_{13}Cl$) occurs in evaporite beds with halides and sulfates, along with a complex group of hydrated magnesium and calcium borates.

Sulfates. Approximately 130 minerals contain sulfate ions. In the geochemical cycle, sulfides and volcanic emissions of sulfurous gases tend to become oxidized. Some sulfates result from local oxidation of ore sulfides, but large sulfate bodies result from evaporation of sulfatic solutions. Barite, celestite, and anglesite are isostructural minerals in which tetrahedral SO_4 groups provide 12 oxygen atoms around the large cations to yield the compositions $BaSO_4$, $SrSO_4$, and $PbSO_4$, respectively. Barite typically occurs in veins associated with lead ores as dirty, massive deposits. Its low solubility leads to sedimentary deposits from limestone weathering. Barite is readily identified from its heaviness (specific gravity of 4.5). Barite can participate in major solid solution with celestite, but the latter is usually pure. Celestite is a relatively rare mineral that may occur in a variety of sediments, probably as a precipitate from solutions; it also occurs in hydrothermal veins. Anglesite is a rare mineral formed in the oxidized zone of galena ores.

Anhydrite and gypsum

Anhydrite ($CaSO_4$) has a different structure because the smaller Ca cation is bonded to only eight oxygens of the SO_4 groups. The symmetry is orthorhombic (three mutually perpendicular crystal axes of unequal length), and three mutually perpendicular cleavages are characteristic. Anhydrite occurs in massive deposits interstratified with beds of gypsum, carbonates, and halides. Seawater precipitates anhydrite above 42° C (108° F), or at lower temperature upon evaporation. Gypsum is deposited at lower temperature or lower salinity. Some deposits of gypsum and anhydrite probably were derived by alteration of the other mineral; that is, by hydration or dehydration.

Gypsum ($CaSO_4 \cdot 2H_2O$) is an important economic mineral, being used extensively in wallboard and cements for the construction industry, as fillers in paper and paints, and as a soil dressing. Its softness (hardness of 2 on Mohs' scale) and the waxy surface of the massive variety alabaster permit use in statues and ornamental carvings. Most gypsum is massive and stained. Rare, transparent, well-developed crystals are called selenite. Fibrous forms are called satin spar. The weakly bonded structure contains water molecules and SO_4 groups in a monoclinic cell (three crystallographic axes of unequal length, two of which are perpendicular and the third inclined). Swallow-tailed twins are displayed commonly in mineral collections. Dehydration near 100° C (212° F) produces the hemihydrate used commercially as plaster of paris because it rehydrates to a solid aggregate of gypsum. Heating to 200° C (about 400° F) produces anhydrite. Gypsum occurs not only in evaporite deposits, as noted above, but also in low-temperature hydrothermal veins and as products of volcanic gases.

Epsomite ($MgSO_4 \cdot 7H_2O$) and kieserite ($MgSO_4 \cdot H_2O$) are just two of the six hydrates of magnesium sulfate deposited by appropriate solutions at characteristic temperatures and subject to hydration and dehydration reactions as conditions change.

Chalcanthite ($CuSO_4 \cdot 5H_2O$), antlerite [$Cu_3(SO_4)(OH)_4$] and brochantite [$Cu_4(SO_4)(OH)_6$] are the three common sulfates formed as alteration products of copper ores in arid regions. The copper ions cause rich blue (chalcanthite) and green colours (antlerite and brochantite).

Glauberite [$Na_2Ca(SO_4)_2$] and

polyhalite [$K_2Ca_2Mg(SO_4)_4 \cdot 2H_2O$]

are important minerals in salt deposits. Other notable sulfates include mirabilite ($Na_2SO_4 \cdot 10H_2O$), which occurs as an evaporation product of saline lakes and as a soil efflorescence, melanterite ($FeSO_4 \cdot 7H_2O$), a greenish-blue weathering product of iron sulfides, and alunite [$KAl_3(SO_4)_2(OH)_6$] and jarosite [$KFe_3(SO_4)_2(OH)_6$], which are isostructural. Alunite results from alteration by sulfatic gases in volcanic regions, whereas jarosite is formed by alteration of iron ores.

Chromates. Chromium mostly substitutes in iron and aluminum minerals. Under oxidizing conditions, crocoite ($PbCrO_4$), an orange-red monoclinic mineral, occurs as a secondary product of lead minerals.

Phosphates, vanadates, and arsenates. Although quantitatively insignificant except for apatite, there are more than 250 minerals in this group—resulting from an incredible variety of crystal structures. The complex ions condense into clusters and link with cations, hydroxyls, and water molecules into families of fascinating chemical and topologic interest (see further PHOSPHATE MINERALS). Here only the vanadates and arsenates are briefly considered.

Important arsenates include adamite [$Zn_2AsO_4(OH)$], erythrite [$Co_3(AsO_4)_2 \cdot 8H_2O$], and

annabergite [$Ni_8(AsO_4)_2 \cdot 8H_2O$]

which are oxidation and hydration products of zinc, co-

balt, and nickel ores. Mimetite [$Pb_5(AsO_4)_3Cl$] and vanadinite [$Pb_5(VO_4)_3Cl$] belong to the pyromorphite structural group. These minerals form by alteration of lead sulfides using components brought in from breakdown of arsenides and vanadium-bearing minerals.

Carnotite [$K_2(UO_2)_2(VO_4)_2 \cdot 3H_2O$] and tyuyamunite [$Ca_2(UO_2)_2(VO_4)_2 \cdot nH_2O$] are rare but widespread alteration products of vanadium and uranium ores; their yellow colour is quite distinctive.

Tungstates and molybdates. Wolframite [$(Fe,Mn)WO_4$] and scheelite ($CaWO_4$) are the principal ores of tungsten. Wolframite occurs in high-temperature hydrothermal veins and altered rocks, and scheelite occurs by contact metasomatism of carbonate rocks. Both minerals contain tetrahedral WO_4 groups bonded to metal cations.

Wulfenite ($PbMoO_4$) is a minor ore of molybdenum and occurs frequently in weathered lead ores. Spectacular orange-yellow square plates are familiar to mineral collectors. The crystal structure is similar to that of scheelite.

Occurrence of silicate minerals

Silicates. The high abundance of oxygen and silicon in the solar system, plus the separation of much of the iron metal and sulfide as an immiscible phase, leads to the dominance of silicates in the terrestrial planets. On Earth, the crust and mantle essentially consist of silicates with only trivial amounts of other minerals. There are approximately 600 silicate minerals, but the major rock-forming minerals at the Earth's surface fall into the principal families of olivine, pyroxene, amphibole, mica, feldspar, quartz, aluminosilicates, feldspathoids, clay minerals, and zeolites. Each of these families is represented by separate articles that describe the principal properties; a few general relationships are described here to provide a framework within the overall subject of minerals (see further SILICATE MINERALS).

At low pressures, silicon enters the tetrahedral "hole," or site, between four oxygen atoms to form the basic silicate building block. This tetrahedral unit can yield Si_2O_7 groups; rings and chains of SiO_3; and sheets and frameworks of SiO_2. The silicate units combine with metal cations, water molecules, carbonate ions, etc., to yield a great variety of structures. Most structures have an elegant geometric pattern of fundamental topologic significance.

With falling temperature, natural silicates tend to be more polymerized or occur in more forms. At high temperature, olivine is stable. Silicon and magnesium or iron atoms enter the interstices of oxygen atoms in a hexagonal close-packed pattern to form isolated SiO_4 groups, linked by magnesium atoms in 6-fold coordination. Zircon ($ZrSiO_4$) also has isolated SiO_4 groups, but zirconium (Zr) is in 8-fold coordination. Pyroxene, *e.g.*, diopside ($CaMgSi_2O_6$), has SiO_3 chains that are cross-linked by cations. Amphiboles such as tremolite [$Ca_2Mg_5Si_8O_{22}(OH)_2$] have Si_8O_{22} bands. The mica mineral muscovite [$KAl_2Si_3AlO_{10}(OH)_2$] has triple sandwiches of aluminosilicate sheets surrounding aluminate sheets. Feldspars, feldspathoids, and zeolites consist of an aluminosilicate framework into whose holes cations and water molecules enter.

Wide ranges of chemical substitution occur, governed by ionic sizes and charges. Aluminum commonly substitutes for silicon, and divalent cations such as calcium, magnesium, and iron show extensive or complete substitution. Coupled substitutions occur: thus albite ($NaAlSi_3O_8$) and anorthite ($CaAl_2SiO_8$) form the end members of the plagioclase feldspars.

Aluminum is a common constituent of silicates, substituting for silicon in tetrahedral coordination with oxygen or occurring in sites that are surrounded by six oxygens. The AlO_6 octahedron can form column or sheet structures. In the aluminosilicate polymorphs of Al_2SiO_5 (sillimanite, andalusite, and kyanite), columns of shared AlO_6 octahedrons are cross-linked by tetrahedrons and octahedrons. In mica and clay minerals, sheets of aluminate octahedrons are combined with silicate sheets plus hydroxyls and water molecules.

Water molecules enter silicates only at low temperature, as in the zeolites. With increasing temperature, hydrous silicates give way to hydroxylates or anhydrous silicates.

High-pressure assemblages

At high pressures, silicates invert to denser polymorphs or assemblages. Thus the feldspar albite ($NaAlSi_3O_8$) gives way to quartz and jadeite, the pyroxene structure of which has a tighter packing. Some hydroxylate minerals are stable at high pressure. In the Earth, increasing depth results in a greater pressure and temperature. In addition, there is a change of bulk chemical composition, first to a composition rich in magnesium silicate and, ultimately, to iron plus sulfide. Laboratory studies plus observations on rocks erupted from great depths and on meteorites suggest that garnet, pyroxene, olivine, and spinel structures are important in the upper mantle of the Earth. These structures have different chemical compositions from those found on the Earth's surface because there exists the tendency of metal ions to move toward a higher coordination as the pressure increases.

From the economic viewpoint, silicates form the major part of many inorganic building materials from granite to concrete. Kaolin and nepheline are used for ceramics; olivine, beryl, zircon, feldspar, and other silicates are used as gems. Silicates are generally unsuitable as ores for metals because of the difficulty of separating the oxygen from the metal.

THE PHYSICAL PROPERTIES OF MINERALS

Increasingly simple visual and tactile estimations of mineral properties are augmented by complex techniques from the domain of solid-state physics (See *Bibliography* for tabulation of data).

Mohs' scale of hardness

Hardness. The simplest estimate of hardness involves scratching by reference minerals arbitrarily chosen for the Mohs' scale in 1882 and placed in numerical order of increasing hardness: 1 talc, 2 gypsum, 3 calcite, 4 fluorite, 5 apatite, 6 orthoclase, 7 quartz, 8 topaz, 9 corundum, and 10 diamond. The scratch test must be made carefully because many minerals contain surface impurities. Quantitative estimates utilize the magnitude of indentation made by a known load applied to a known point (*e.g.*, the Vickers hardness test often applied to ore minerals). Each step on Mohs' scale corresponds approximately to a doubling of the indentation. Hardness is a directional property, especially of kyanite, for which it varies on the cleavage face from 4 to 5 when scratched parallel to the long, or *c*, axis, and 6 to 7 in a direction that is perpendicular to it. The octahedral surface of diamond is the hardest of all minerals. The hardness tends to correlate with the weakest of the bonds in the crystal structure, ranging from the van der Waals bonds of talc to the covalent bonds of diamond.

Cleavage, fracture, and parting. Strain beyond the elastic and plastic limit causes either irregular fracturing or cleavage along one or more planar surfaces. Fractures result from accidental irregularities, but cleavages are controlled by unequal (anisotropic) bonding strengths in different directions in the crystal structure; they always occur parallel to potential faces. The lubricating properties of graphite, talc, and molybdenite result from cleavage into tiny slippery flakes. Mica cleaves into thin, insulating plates that are useful in the electrical industry. In many minerals, chemical bonding is only weakly anisotropic, leading to imperfect cleavage. Parting resembles cleavage but is not an inherent property of the mineral; it results from various defects such as chemical changes and twinning.

Specific gravity. The specific gravity of a mineral is the ratio of the mineral density to that of water at 4° C (39° F), and it varies from 0.92 for ordinary ice to 21 for platinum. Direct measurements of the ratio of weight to volume are affected by voids and impurities. Suspension in heavy liquids of mineral powder followed by determination of the heavy liquid provides higher accuracy. The density can be calculated from the unit cell volume (determined by X-ray diffraction) and the chemical composition, assuming full atomic-site occupancy. Minerals are commonly separated in laboratories and industrial plants by suspension, or sinking, in liquids. Many monomineralic igneous rocks and economic-mineral deposits form by density separation.

Mineral deformation

Elasticity and strength. Under low stresses minerals are elastic, but under higher stresses some kind of permanent deformation occurs. In the Earth, velocity of seismic waves is related to these stresses (see EARTHQUAKES). The rigidity coefficients decrease with temperature and correlate with the type of chemical bonding and atomic packing. Metallic minerals are ductile (*e.g.*, will flow under high stress), and empirical relations for oxides and silicates are very useful.

The onset of permanent deformation occurs at lower stresses as the temperature rises. Controlled laboratory experiments, especially of carbonates and silicates, delineated the failure mechanism in terms of fracturing, twinning, and structural inversion. Particularly important are deformations induced by shock, produced naturally by meteorite impact and experimentally by explosions. The pressure and temperature rise very rapidly as the shock front passes, resulting in deformation, inversion to dense polymorphs, and finally melting as the peak pressure rises (see further METEORITE CRATERS). Intensity of meteorite impact and nuclear explosions is calibrated by the deformation: for most natural silicates, melting occurs by about 0.5 megabar. Occurrence of coesite and stishovite, the high-pressure forms of SiO_2, is used as a criterion for meteoritic origin of circular features.

Magnetic and electrical properties. Minerals may be ferromagnetic, or attracted by a magnet (iron), anti-ferromagnetic (magnetite; used in primitive compasses), paramagnetic (all iron-bearing silicates), or diamagnetic —that is, repelled by a magnet (many iron-free silicates). The position of the magnetic field at the time of rock consolidation can often be determined from the magnetic direction of the constituent minerals (see ROCK MAGNETISM). Minerals are separated in the laboratory and commercially by passing them down an inclined chute between the jaws of an electromagnet. Certain ore bodies are located from their magnetic field, and subtle features of chemical forces and site distribution of atoms also are determined from techniques based upon magnetic properties.

Mineral conductivity

Minerals range from electrical conductors (metals), to semiconductors, to nonconductors (mica and graphite). Minerals with good conductivity have metallic lustre, whereas most transparent ones have low conductivity. Natural and synthetic mica is widely used as an insulator. Nonconducting crystals with polar directions become electrically charged under compression (piezoelectricity) or upon heating (pyroelectricity). Quartz piezoelectric oscillators are used for generation of ultrasonic waves (militarily exploited for submarine detection) and provide highly stable frequency standards. Tourmaline pyroelectric crystals are used in pressure gauges.

Optical properties. Light incident on the surface of a mineral is partly reflected, partly transmitted, and partly absorbed. Minerals with a high reflectivity have metallic bonding, whereas those with a high transmissivity have ionic or covalent bonding. The optical properties vary with direction (except for isometric minerals—those in which properties are the same in all directions) and depend on the atomic nature of the crystal structure.

Colour of minerals

The colour of a mineral as seen by the naked eye or at low magnification is a broad property that often depends on elusive factors. The colour of a metallic mineral is governed by the interaction of light with the surface. Some metallic minerals such as gold are chemically stable, yielding an intrinsic colour. Most are unstable, yielding a coating of a different colour (silver, copper, and many sulfide minerals). The intrinsic colour is yielded by rubbing the mineral on a porcelain plate to obtain a streak. The colour of a transparent mineral may be intrinsic (*e.g.*, green of the birthstone peridot), arising from interaction of the electromagnetic radiation with electrons, or it may be extrinsic, arising from inclusions (*e.g.*, reddish or whitish orthoclase). Intrinsic colours can arise from major constituents (divalent iron in peridot and copper in chalcanthite) or from minor constituents substituting into the crystal structure (chromium in ruby), or even from lattice defects. The colour may vary with the direction of light propagation (pleochroism), as in cordierite and alkali amphibole. Some minerals show a wide range of intrinsic colours (diamond, corundum, fluorite, beryl). The causes of some typical colours (blue and green of kyanite, green of the amazonite variety of microcline) are highly controversial despite extensive research. Sub-microscopic inclusions causing complex diffraction yield the remarkable light scattering of moonstone and labradorite. Flakes of iron oxide cause the light scattering of aventurine. Star gems (rubies, sapphires, garnets) contain sub-microscopic inclusions oriented in crystallographic planes. Colour thus is often a poor guide to mineral identification.

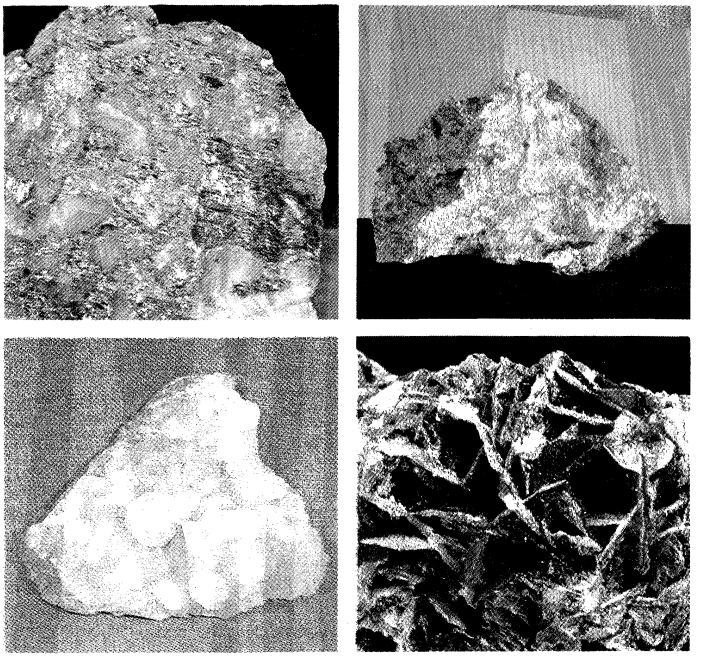


Figure 6: *Minerals in which colour is an intrinsic property of the species.*
(Top left) Brass-coloured chalcopyrite, $CuFeS_2$, from Ontario, Canada. (Top right) Red cinnabar, HgS, from Arkansas. (Bottom left) Yellow sulfur, from Racalmuto, Italy. (Bottom right) Green dioptase (a hydrous copper silicate) with orange-yellow wulfenite, $PbMoO_4$, from Arizona.
By courtesy of (top left, top right, bottom right) Ted Boente; photographs, John H. Gerard—EB Inc., (bottom left) Floyd R. Getsinger

The reflected light from polished surfaces permits ready distinction between many opaque minerals on the basis of brightness, colour, and anisotropy (variation of properties with crystallographic direction) and is mandatory for study of ore minerals and host rocks.

Transparent minerals (especially in rocks) are commonly identified by transmitted light, using polarized light passing through thin sections or through fragments mounted in oil.

Most minerals have anisotropic optical properties, and incident light splits up into two polarized waves with different velocities. The plane of polarization and velocity vary with direction, and the principal indices of refraction, orientation of the optic axes, and variation of indices with wavelength (dispersion) and absorption of light serve to identify most transparent minerals, but the procedures are highly technical.

Advanced studies utilize a microscope fitted with a multi-axis stage for orienting the mineral; this permits the measurement of all optical properties. Routine studies of thin sections of rocks rely on the skill of the microscopist, who mentally correlates the properties seen in different grains. The degree of relief, polarization colour, position of optical extinction, shape, twinning, nature of inclusions, and other factors are all used to yield a probable identification.

Thermal properties. The thermodynamic properties are a function of pressure, temperature, chemical composition, and distribution of atoms in the crystal structure. Measurements are made of the specific heat, melting point, latent heats for change of state, thermal conductiv-

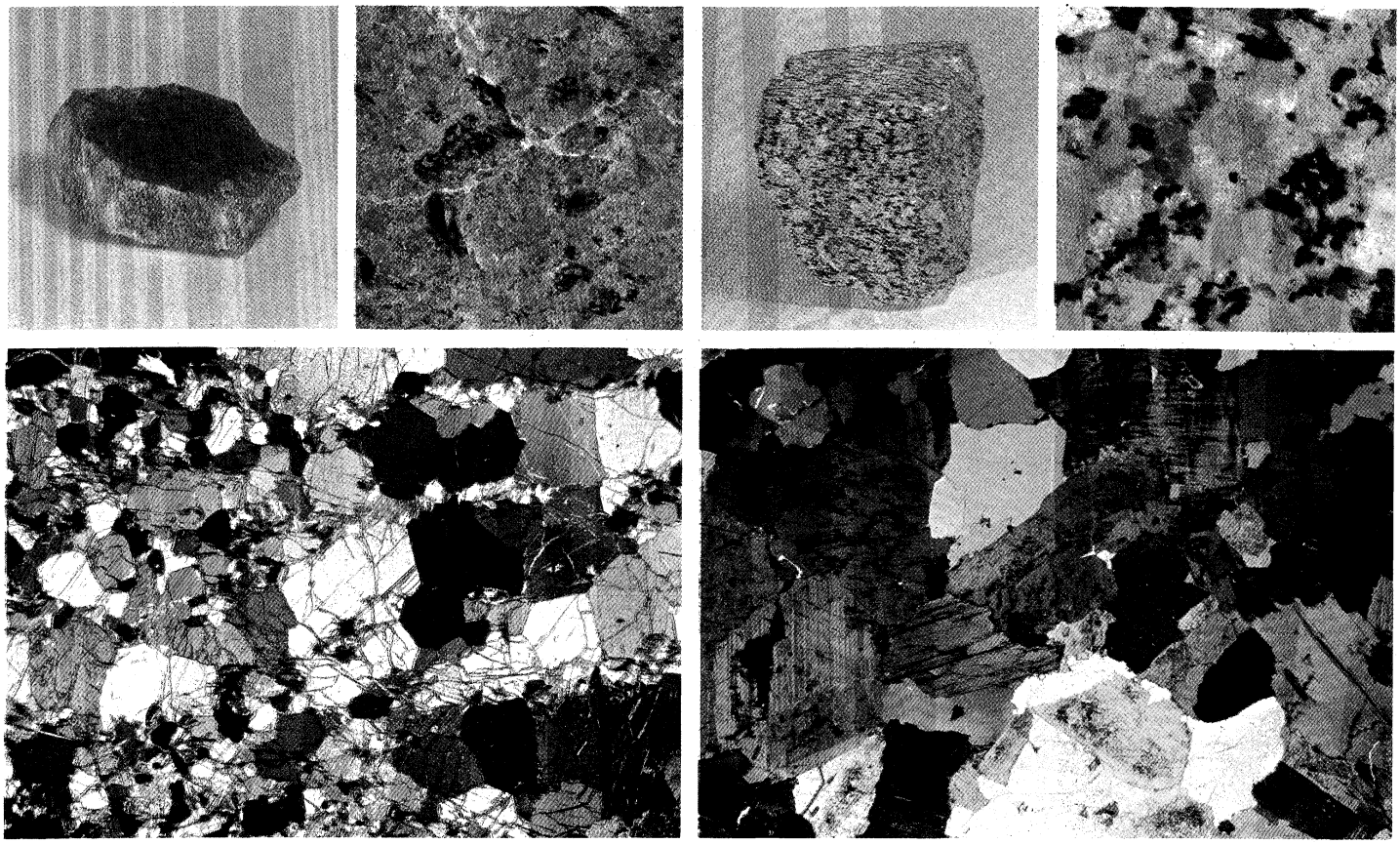


Figure 7: *Two types of crystalline aggregates at three levels of magnification.*
Most mineral crystals occur without faces (anhedral crystals) and as crystalline aggregates in the form of rocks. (Top left) Olivine crystals forming the monomineralic rock dunite, in which the individual crystals can scarcely be distinguished. (Centre left) Polished surface of the specimen (magnified 2×). Light green grains are olivine, dark spots are magnetite and serpentine. (Bottom left) Photomicrograph of the specimen in polarized light (magnified 12×). The different colours are due to the optical effect of the different orientations of the olivine crystals on the polarized light; most of the grains are olivine. (Centre right) Quartz, feldspar, biotite, and hornblende crystals forming the rock granite, a polymineralic aggregate. (Top right) Cut and polished surface of the rock (magnified 1½×). Large pink crystals are microcline feldspar, white ones are sodic plagioclase feldspar, small smoky ones are quartz, and black spots are biotite and hornblende. (Bottom right) Photomicrograph of the rock in polarized light (magnified 12×). White and gray crystals are quartz, light to dark bluish-gray ones are feldspar, those with irregular cross-hatching are microcline perthite, and those with finely ruled parallel lines are sodic plagioclase. Most of the green and brown crystals are biotite; the brown one in the centre is hornblende.
D.L. Weide

ity, and similar thermal properties. Differential thermal analysis involves detection of temperature discontinuities as heat is supplied at a constant rate: the discontinuities reveal structural changes that are diagnostic of some minerals (dehydration and structural readjustment in clays; bond-bending inversion in quartz; oxidation; structural breakdown).

Radioactivity. Minerals containing uranium and thorium provide commercial sources for nuclear energy and can be detected by radiation counters. The crystal structures of minerals containing thorium (zircon, allanite, thorite) commonly undergo a change of state by internal movement of recoiling atoms. Inclusions of these minerals in some other minerals are revealed in thin section by pleochroic halos; thus biotite is bleached.

Deliberate etching reveals the track of radioactive particles in minerals. The size of the track provides a measure of the energy of the particle. Study of minerals from the lunar surface is providing estimates of their exposure time to, and mass distribution of, cosmic rays.

Determining the ages of minerals

Several rock-forming minerals contain enough radioactive material to permit determination of the time since the radioactive material was incorporated in the mineral. The techniques involve accurate mass spectrometry and careful consideration of the thermal history of the mineral. The isotope potassium-40 decays to argon-40, but argon may be either lost or gained by diffusion, and thus misleading ages may be obtained. The uranium isotopes decay to lead isotopes, permitting a check on the concordance of the ages implied by the different decay series. The rubidium isotope, rubidium-87, decays to strontium-87, augmenting the abundance of the latter. Measurement of coexisting minerals with different amounts of rubidium and strontium permits estimation of the age at which the minerals formed together and the original isotopic ratio of the mineral system. Careful study permits selection of genuine ages from misleading ages, and the principal mineral-forming events on the Earth are dated (see DATING, RELATIVE AND ABSOLUTE). Isotope studies of meteorites and of terrestrial and lunar minerals yield crucial information on the chemical evolution of the solar system.

The distribution of stable isotopes between minerals, or between minerals and solution, depends on the temperature. The distribution of the isotopes oxygen-16 and oxygen-18 between silicates and oxides of rocks and the distribution of isotopes in biological carbonates yield estimates of their equilibrium temperatures.

Luminescence. The emission of light from a relatively cool mineral results from extremely complex electronic processes and from various energy sources. The luminescent colour bears no relation to the natural colour. Fluorescence excited by ultraviolet radiation permits rapid detection of certain minerals, including some uranium ores. Cathodoluminescence (excitation by accelerated electrons) is routinely used to identify some rock-form-

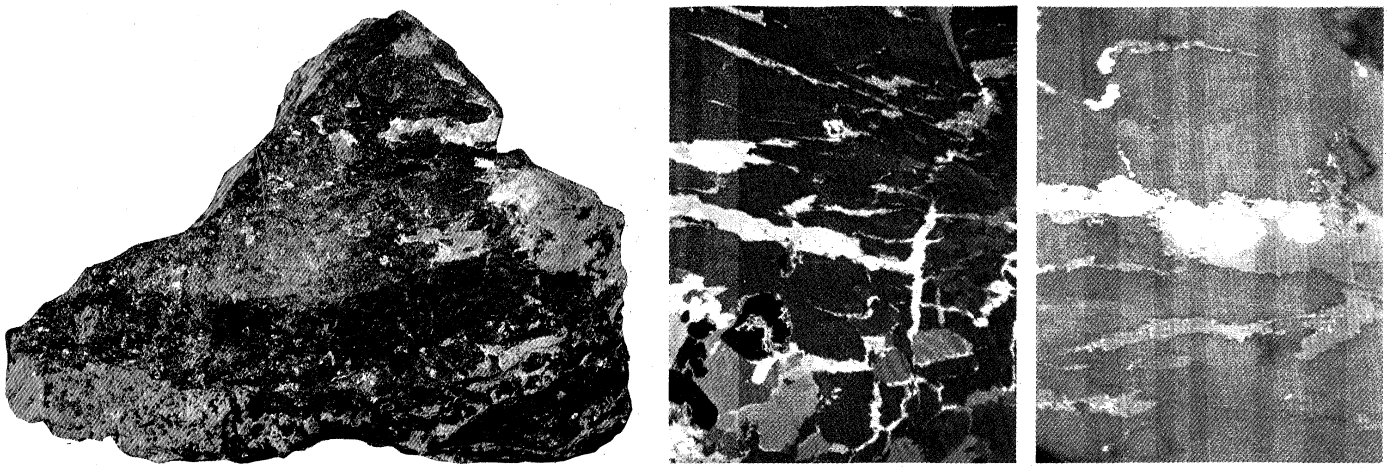


Figure 8: *Aggregation and intergrowth of ore minerals at megascopic and microscopic levels.*
(Left) Aggregate of copper ore minerals (from Arizona) including bornite, Cu_5FeS_4 (reddish purple to iridescent blue where tarnished); covellite, CuS (dark blue); chalcopyrite, $CuFeS_2$ (yellow); and quartz and altered rock (white and light gray). (Centre) Photomicrograph of polished surface of what appears megascopically to be covellite (magnified 200 ×). The specimen is actually bornite that has been partly replaced by extremely fine-grained covellite (greenish-gray). (Right) Photomicrograph of central portion of polished surface to left, showing covellitized bornite (pink to red) more clearly (magnified 400 ×). Much more bornite is present than is apparent at lower magnifications. Also present are chalcocite, Cu_2S (gray), chalcopyrite (yellow), and tiny crystals of pyrite, FeS_2 (pale yellow), in the chalcocite veins.
By courtesy of (left) Ted Boente; photographs, John H. Gerard, (centre, right) Robert H. Carpenter

ing minerals and to map variations of chemical composition as in carbonates and feldspars. Thermoluminescence is the emission of light released from stored energy levels

Benjamin M. Shaub

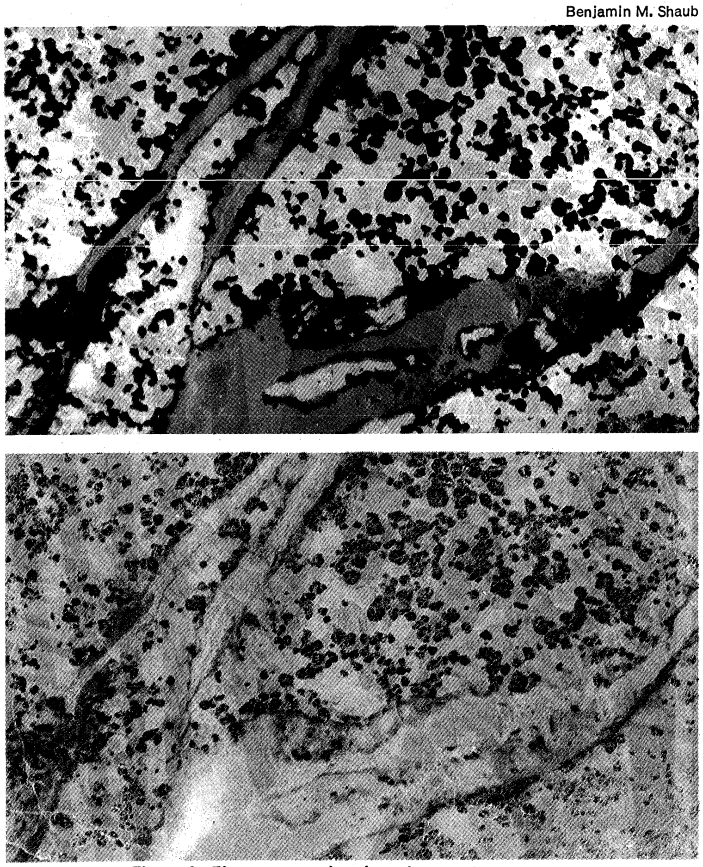


Figure 9: *Fluorescence in minerals.*
(Top) Specimen, in ordinary light, of calcite (calcium carbonate) with grains of black franklinite (a zinc, manganese, and iron oxide) and veins and grains of brownish-gold willemite (a zinc silicate), from Franklin, New Jersey. (Bottom) Fluorescent colours due to excitation by ultraviolet light. Red is calcite, green is willemite, black is nonfluorescent franklinite.

by heating: it provides a crude estimate of the radiation and thermal history of certain minerals and rocks. Triboluminescence is the emission of light induced by mechanical deformation, as in sphalerite and fluorite. The electronic processes are known for very few luminescent minerals (*e.g.*, manganese substitution in red luminescing calcite). The phenomenon is erratically developed for most minerals, and for each mineral species there are several causes, including chemical substitutions and physical defects.

THE OCCURRENCE OF MINERALS IN NATURE

General geochemical considerations

Minerals result from a sequence of complex processes that began with chemical differentiation of the solar system to form planets and that ended, in many cases, with crystallization in rocks or ore bodies controlled by trivial local factors. The occurrence of minerals, therefore, involves a general understanding of geochemistry, geophysics, and petrology (the study of rocks), as well as special factors involved in the formation of ore deposits. Although spectacular advances have been made in these subjects during the 20th century, and particularly since 1950, the factors controlling mineral genesis are often controversial, and ideas can be expected to undergo revision.

The continuing changes on Earth and in the solar system result in mineralogic changes such that mineral genesis must be considered in terms of kinetic rather than static factors; most minerals become transformed to other minerals by reaction with air and water when they reach the Earth's surface, for example.

The presence of water is extremely important because it permits concentration of elements to form valuable ore bodies. The extreme dryness of the Moon inhibits formation of such types of ores, and it appears likely that the number of mineral types found on the Moon will be small. Under dry conditions, mineral concentrations arise principally from the processes of liquid immiscibility and from gravity settling of crystals from liquid. Under aqueous conditions, these processes are supplemented by incongruent solubility, by differential resistance to weathering, and by fluid transport and deposition. Furthermore, biological processes profoundly affect weathering and deposition on Earth.

In the magmatic domain, the classification of elements into lithophile, chalcophile, siderophile, and atmophile groups for those elements with an affinity for silicate, sulfur, iron, and oxygen, respectively, is extremely important in providing a guide to whether an element enters a mineral containing oxygen or sulfide or whether it occurs as a native element or enters the atmosphere. In the low-temperature aqueous domain, the relation between ionic charge and ionic radius controls whether an element is transported as a soluble complex anion (for example, the

sulfate anion SO_4^{2-}) or a soluble cation (sodium, Na^+, for example) or whether it forms a hydrolysate, such as $Fe(OH)_3$, which is insoluble. The first two types may react to form a precipitate (*i.e.*, sodium sulfate [Na_2SO_4]). Finally, the availability of oxidizing material is important in determining the valence state of iron and manganese and in determining the extent and nature of alteration.

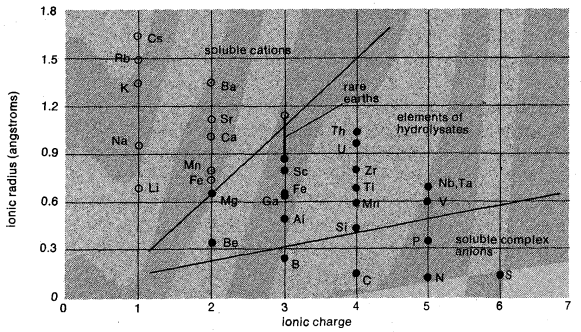


Figure 10: Consequences of ionic radius and ionic charge in geochemical processes (see text).

Derivation from silicate magma

Mineral associations in igneous and metamorphic rocks. Volcanic rocks on Earth are extruded igneous rocks and are readily proved to have crystallized from a liquid silicate magma containing some water and volatile gases. Rocks of similar composition but of coarser grain size are recognized to have crystallized at some depth within the Earth. There is considerable controversy about the origin of such deep (plutonic) rocks, however, and about the origin of the magmas responsible for volcanic rocks. Almost certainly many processes including melting, crystallization, annealing, remelting, and metasomatism (process of rock replacement), are involved. Particularly spectacular are kimberlite pipes (relatively narrow igneous bodies that extend downward into the Earth's crust) containing diamonds that probably were derived from depths of several hundreds of kilometres.

Igneous rocks are dominated by the silicate minerals (olivine, pyroxene, feldspar, quartz, feldspathoid, mica, amphibole); their differing proportions yield the various rock types. Accessory minerals (magnetite, ilmenite, sphene, zircon, apatite, allanite, fluorite, xenotime) occur in small quantities, but occasionally they are concentrated into ore bodies either by gravity separation or by later weathering.

During the crystallization of igneous rocks from silicate melts, the small amounts of water and volatile gases tend to concentrate, particularly in granites. This leads to the crystallization of especially large crystals. The concentrations may be small in volume (vugs) or may become large pegmatite bodies, commonly known as veins or dikes. Most pegmatites contain essentially the same minerals as the host rock, but a few pegmatites of great economic importance contain spectacular minerals of rare elements (*e.g.*, beryl, lepidolite, pollucite, cassiterite).

Metamorphic processes are commonly described as being either of contact or regional type. In the contact type, an intrusive magma heats an existing rock, producing a narrow zone of recrystallized minerals. Volatile gases from the magma may cause extensive alteration of the chemical composition (especially typical for magmatic intrusion of limestone), producing skarns. Such skarns may prove exploitable as ores.

Regional metamorphism involves burial (often of a sediment) followed by release to the surface after erosion. The pressure and temperature rise to a maximum and then fall. During the rise, the rock normally contains water, which acts as a catalyst for recrystallization. The water is essentially lost in the process, resulting in metastable persistence of minerals during cooling. Many sediments are rich in alumina and silica, resulting in the formation of aluminosilicates, cordierite, staurolite, and others, each of which forms at a specific range of temperature and pressure. Any type of rock may undergo metamorphism, however, resulting in extraordinary complexities that have been only partly resolved. Nevertheless, considerable progress has been made in estimation of pressure–temperature–time–composition relations for the great metamorphic cycles associated with sedimentation and mountain building (see further ROCK METAMORPHISM, PRINCIPLES OF; and METAMORPHIC ROCKS).

Typical metamorphic minerals in mineral collections are garnet, kyanite, staurolite, tremolite, idocrase, and wollastonite. Almost all silicates, and many other minerals, may occur in metamorphosed rocks; calcite in marble is a good example.

Veins, replacements, and secondary deposits

Mineral associations in ore deposits. The crust of the Earth is unstable, and this condition results in extensive deformation and fracturing associated with igneous and other activity. The fractures form channels for aqueous solutions that deposit minerals; many are of economic value and are called ore deposits. The mineral deposits are loosely classified by the supposed temperature of deposition: hypothermal (300°–500° C [600°–900° F]); mesothermal (200°–300° C [400°–600° F]); and epithermal (up to 200° C). Pressure is also an important parameter. Typical minerals in this classification scheme are as follows: hypothermal—cassiterite, wolframite, scheelite; mesothermal—sulfides of copper, zinc, iron, and lead; epithermal—sulfides of antimony, mercury, silver, lead, zinc, and silver and gold as native metals. Gangue minerals (those associated minerals that are not of economic value) commonly include quartz and carbonates, among others; a few, such as fluorite and barite, may be of commercial value, however.

Solutions may percolate through porous rock, resulting in the precipitation of new minerals and alteration of old ones. Some of these replacement deposits, such as those of copper sulfides, have economic value.

Primary sulfides are unstable in the presence of air and water. A vein of primary sulfides becomes leached at the surface, producing a gossan (*e.g.*, decomposed rock) of quartz and iron oxides. Above the water table, oxidized and carbonate minerals such as malachite, azurite, and cuprite are formed. Below the water table, secondary sulfides such as chalcocite and covellite are precipitated.

In volcanic regions, escaping aqueous solutions mix with surface water to produce hot springs and fumaroles. Most of the deposited material is silica, but some deposits contain sulfur, sal ammoniac, sulfides, or oxides.

Mineral associations in sedimentary rocks and precipitates. Upon exposure to water and air, most minerals break down to yield a residue and an aqueous solution. A few minerals such as quartz are highly resistant to alteration but are released by breakdown of other minerals to be concentrated into mechanical sediments. The products of breakdown proceed in several directions, depending on several chemical and mechanical factors. The hydrolysates form clays (mostly hydrated aluminosilicates), chert (silica), bauxite (alumina), and sedimentary iron silicates (glauconite, chamosite). Under oxidizing conditions, some iron and manganese become oxides. Under reducing (anaerobic) conditions, as in swamps, in the presence of sulfur-bearing material, iron sulfides and carbonates may be precipitated. Bacteria are increasingly recognized as important agents in both weathering and precipitation of such sediments.

Biological factors and the role of organisms

The soluble cations plus complex anions enter oceans and lakes, where they react with each other and with volcanic emanations plus the atmosphere. Biological organisms produce carbonates that yield limestone. Calcite and aragonite react with magnesium from seawater to form dolomite. Silica is incorporated by some organisms, such as the single-celled Radiolaria. Parts of the ocean floor and some lake beds are covered by nodules rich in manganese oxides containing other valuable elements. A host of complex organic and inorganic reactions occur in seawater, resulting in relatively small changes of composition. In lakes, the composition may change drastically as a result of variation of rainfall and of incoming and outgoing flows. Drying up of a basin leads to sequential precipitation of evaporite minerals as the successive solubility products are reached. The resulting layered beds of salt (*e.g.*, gypsum, halite, sylvite, carnallite) are of

great economic value, especially as they are often associated with deposits of sulfur and petroleum.

Beach sands are commonly dominated by quartz grains sorted by prolonged mechanical processes. Heavy minerals may be concentrated into deposits of economic value (*e.g.*, gold, cassiterite, ilmenite, diamond).

Knowledge of marine and lake sediments is rapidly increasing as a result of deep drilling, dredging, and seismic studies. Marine sediments are composed of minerals derived from: rock fragments transported by rivers and ocean currents; windblown debris, such as desert sand and volcanic ash; chemical deposition, as in the case of carbonates, zeolites, and manganese nodules; death of biological organisms; and volcanic intrusions. The resulting minerals are slowly yielding evidence on the time of their deposition and the nature of the chemical environment. Ultimately the history of the ocean basins and continents will be derived from the mineral properties. For example, the recurrent beds of evaporite minerals in the Mediterranean Sea indicate cyclic sedimentation and dehydration associated with the differential movement of Africa and Europe. The concepts of sea-floor spreading and plate tectonics will be checked by the isotopic ratios of the minerals in these areas.

Mineral associations in the Moon, planets, and meteorites. Study of the lunar rock samples returned by the Apollo and Luna missions suggests that the most important factors controlling the formation of minerals on the Moon are: (1) crystal–liquid differentiation of material rich in basaltic components and low in volatile material, (2) an absence of biological processes, (3) absence of hydrothermal processes and water transport, (4) a high degree of reduction of transition metals and europium to low-valence states, and (5) liquid immiscibility.

Minerals of the lunar surface

The lunar maria (surface depressions) contain surface lava flows consisting of dominant pyroxene, plagioclase, olivine, and ilmenite plus about 30 other minerals. Troilite and native iron separated from a liquid immiscible with the silicate liquid. Spinel and armalcolite (an acronym from the American astronauts Armstrong, Aldrin, and Collins) are early minerals. The final products of crystallization contain an immiscible rhyolitic (equivalent to granite in composition) liquid associated with small amounts of apatite, whitlockite, pyroxferroite (a new mineral), zircon, and various other rare minerals, some of which are not yet identified.

The lunar surface is covered by fragmented minerals and rocks, some deformed by meteorite impact and others thermally sintered into breccia (rock consisting of coarse, angular fragments). Fragments rich in one mineral (either plagioclase or pyroxene or olivine) suggest mechanical processes for concentrating minerals by density fractionation in magma. Satellite photography and other remote studies of the surface indicate complex mineralogic variations over the surface that will be resolvable only by detailed exploration. Information available by 1971 indicated highland regions dominated by pyroxene and plagioclase, but most mineralogists and petrologists expect to discover many complex rocks in future lunar exploration. Occurrence of ice in shaded regions and of valuable mineral deposits could not be ruled out in 1971, though the low content of hydrogen and of noble metals in Apollo rocks definitely indicates less favourable economic prospects than on Earth.

Martian and meteorite mineralogy

The mineralogy of Mars is highly speculative. Remote studies indicate polar caps composed principally of solid carbon dioxide (Dry Ice). The surface is heavily cratered in places but shows curious chaotic features elsewhere. The origin of the reddish colour is speculative; it may be due to iron oxide or to biological causes. Direct mineralogic studies may begin in 1975 with unmanned landings and continue around 1980–2000, at which time samples will be returned to Earth. The high density of the planet Mercury suggests a mineralogy dominated by iron. The high surface temperature (near 500° C [900° F]) and high pressure (100 atmospheres, mostly CO_2) of Venus should cause an unusual mineralogy. Asteroids are prime targets for mineralogic study.

The origin of meteorites is controversial (see METEORITE). Their mineralogical and chemical properties are being pursued vigorously in an attempt to determine the development of the solar system. Important concepts governing the genesis of their minerals are: (1) condensation from a solar nebula falling in temperature, (2) mechanical accretion, (3) crystal–liquid fractionation plus metal–silicate immiscibility, (4) mechanical disruption, and (5) shock metamorphism upon collison of meteorites with the Earth. The high state of reduction results in minerals not occurring naturally on Earth.

The iron meteorites typically contain kamacite and taenite, iron and nickel alloys, troilite (FeS), cohenite (Fe_3C), schreibersite (Fe_3P), and similar mineral species. The stony meteorites are dominated by olivine, pyroxene, and plagioclase. Carbonaceous chondrites contain low-temperature minerals, including layer silicates and water-soluble salts (hydrated sulfates). Application of advanced techniques has resulted in discovery of many new minerals, including ringwoodite and majorite, high-pressure polymorphs of olivine and pyroxene. Some of the minerals not found on Earth are: osbornite (TiN); sinoite (SiON); oldhamite (CaS); lawrencite ($FeCl_2$); farringtonite [$Mg_3(PO_4)_2$]; and kosmochlor ($NaCrSi_2O_6$). Rare diamond apparently results from shock upon Earth impact.

The studies of meteorites and planets are changing the old parochial concept of mineralogy as a study of material from the Earth's surface. New concepts of solar mineralogy are developing rapidly as a result of exploration of the solar system, and studies of light from intergalactic regions suggest the presence of graphite and silicate grains.

EXPERIMENTAL SYNTHESES OF MINERALS

Purposes and utility of synthetic minerals

Minerals are synthesized under controlled conditions to obtain information about how they form naturally and to produce commercially valuable products. Growth may occur under conditions of thermodynamic equilibrium, but commonly the first products of synthesis are metastable and only give way slowly to the stable materials. Even when the equilibrium conditions of mineral growth are known, care must be taken in the application to natural minerals. Kinetic factors are extremely important: indeed, if all mineralogic reactions achieved equilibrium, it would be impossible to obtain information about the mineralogic history of the solar system.

Processes and products of mineral synthesis. Essentially all minerals have been synthesized, or can be expected to be synthesized, if the appropriate conditions are used. Some minerals that require complex atomic ordering have not been synthesized, however, presumably because of extreme sluggishness of atomic movement at low temperature. The most important examples are microcline and low-temperature varieties of plagioclase; their high-temperature equivalents with disordered atoms can be synthesized easily.

Several gem minerals are synthesized commercially, and the first synthesis of diamond (by the General Electric Company about 1950) resulted from technical advances of high-pressure research. Probably the most important synthetic minerals from an economic viewpoint are zeolite molecular sieves, which have revolutionized industrial catalytic processes. The synthetic relatives of the extremely rare mineral faujasite (from the hilly area of Kaiserstuhl, near Sasbach, West Germany) and other synthetic zeolites are used for a variety of processes utilizing petroleum. The industrial zeolites are metastable products produced by rapid reaction of alkaline material with aluminosilicate gels at temperatures around 100° C (212° F). Other important industrial minerals are crystals such as ruby used in lasers or in electronic equipment. Synthetic alumina minerals are used for industrial bearings and in clock movements. Synthetic quartz is used in optical equipment and piezoelectric oscillators. Synthetic diamond is a better abrasive than natural diamond.

Synthesis of individual minerals

Individual minerals are synthesized by choosing conditions of pressure, temperature, and bulk composition for which only one mineral phase crystallizes. Sulfide minerals can be synthesized easily by heating the appropriate

source material in sealed silica tubes. Salts can be obtained by concentrating brine until the solubility product is exceeded. Refractory minerals such as corundum are synthesized by fusing alumina in a furnace using the Verneuil process, in which material melted by an oxyhydrogen torch crystallizes onto a seed forming a boule. Minerals stable only at elevated pressure, such as diamond, can be synthesized only in special high-pressure equipment.

Crystallization of some minerals proceeds rapidly only in the presence of a catalyst or flux. Fluxes may operate by producing a low-melting liquid from which the mineral crystallizes. The synthesis of diamond is promoted by catalyst or fluxes. Water is often an extremely effective flux; for silicates it increases crystallization velocities by many orders of magnitude.

Some minerals can crystallize only when the atmosphere is inert or controlled. Thus, ferrous minerals will grow only in the absence or near absence of oxygen gas: otherwise minerals like magnetite or hematite will form.

Phase equilibria of mineral systems. The equilibrium assemblages of coexisting minerals are governed by thermodynamic principles that are covered, according to the equilibrium temperatures and pressures involved, in the articles GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES; GEOCHEMICAL EQUILIBRIA AT LOW TEMPERATURES AND PRESSURES. Mineral assemblages up to depths of several hundreds of kilometres in the Earth (and for the entire Moon) can be simulated by high-pressure–high-temperature apparatus working under static conditions; the existing limits are several hundreds of kilobars (the bar is a unit of pressure equal to a barometric reading of 750 millimetres at 0° C [32° F]) and any necessary temperature. Experiments using explosives and other means of generating shock waves permit transient attainment of pressures up to one or two megabars and correspondingly high temperatures.

Mineral reactions at the Earth's surface generally occur in the presence of aqueous solutions and are governed by the oxidation–reduction conditions as well as the concentrations of cations and anions in the solution. Difficulty of obtaining equilibrium at low temperatures has led to a major attempt to calculate the stability of mineral assemblages from the measurable thermodynamic properties of the individual minerals. Combination of these calculations and careful observations of both synthetic and natural mineral assemblages has produced a fairly good understanding of mineralogic processes at the Earth's surface, but the details are very complex.

Interpretation of natural occurrences. The interpretation of natural occurrences in terms of synthetic data requires careful evaluation of kinetic factors. Fortunately, most minerals appear to retain their bulk composition during falling temperature, and solid-state reactions, including atomic ordering and exsolution, often give evidence about the nature of the cooling history. In general, mineralogists and petrologists find that careful study of coexisting minerals and of rock formations permits estimates of the maximum temperature attained by the minerals to within an accuracy of 100° C (180° F) or so. Pressure estimates are more difficult and tend to be qualitative or semi-quantitative. The inversion of calcite to the high-pressure form aragonite at several kilobars pressure and the supplanting of albite by jadeite plus quartz provide estimates of the minimum pressure attained in certain metamorphic rocks. The formation of aragonite in certain biological organisms at atmospheric pressure is a warning of the extent of nonequilibrium processes. But in general, the progress of mineral syntheses, especially after 1950, is one of the major advances of science and technology and is leading to a significant quantitative control on models for the formation of the Earth and solar system and major improvements in living conditions for the human race.

Minerals as geological thermometers and barometers

BIBLIOGRAPHY

Introductions to minerals: I. VANDERS and P.F. KERR, *Mineral Recognition* (1967); P.E. DESAUTELS, *The Mineral Kingdom* (1968); F.H. POUGH, *Field Guide to Rocks and Minerals* (1953).

Textbooks of mineralogy: I. KOSTOV, *Mineralogy* (Eng. trans. 1968); L.G. BERRY and B.H. MASON, *Mineralogy* (1959); C.S. HURLBUT, *Dana's Manual of Mineralogy*, 17th ed. (1959); E.N. CAMERON, *Ore Microscopy* (1961); E.W. HEINRICH, *Microscopic Identification of Minerals* (1965); P.F. KERR, *Optical Mineralogy*, 3rd ed. (1959); J. SINKANKAS, *Mineralogy: A First Course* (1966).

Standard references: C. PALACHE, H. BERMAN, and C. FRONDEL, *Dana's System of Mineralogy*, 7th ed., 3 vol. (1944–62), see also earlier editions for silicates and early data; W.A. DEER, R.A. HOWIE, and J. ZUSSMAN, *Rock-Forming Minerals*, 5 vol. (1962); W.L. BRAGG, G.F. CLARINGBULL, and W.H. TAYLOR, *Crystal Structures of Minerals* (1965).

Crystallographic references: J.F. NYE, *Physical Properties of Crystals* (1957); L.V. AZAROFF, *Elements of X-Ray Crystallography* (1968); F.C. PHILLIPS, *An Introduction to Crystallography*, 2nd ed. (1956).

Geochemical references: K.H. WEDEPOHL (ed.), *Handbook of Geochemistry* (in parts, 1970–); K. RANKAMA and T.G. SAHAMA, *Geochemistry* (1950); B.H. MASON, *Principles of Geochemistry*, 3rd ed. (1966); R.M. GARRELS and C.L. CHRIST, *Solutions, Minerals, and Equilibria* (1965).

Petrological references: F.J. TURNER and J. VERHOOGEN, *Igneous and Metamorphic Petrology*, 2nd ed. (1960); H.G.F. WINKLER, *Petrogenesis of Metamorphic Rocks* (Eng. trans. 1965); H.B. MILNER, *Sedimentary Petrography*, 4th rev. ed. (1962).

References to ores and the uses of minerals: R.L. BATES, *Geology of the Industrial Rocks and Minerals* (1960); E.W. HEINRICH, *Mineralogy and Geology of Radioactive Raw Materials* (1958); J.L. GILLSON *et al.*, *Industrial Minerals and Rocks*, 3rd ed. rev. (1960); C.F. PARK and R.A. MACDIARMID, *Ore Deposits* (1964).

Journals: *American Mineralogist* (1916–); *Mineralogical Magazine* (1876– ; including *Mineralogical Abstracts*, 1920–58). The international abstracting journal *Mineralogical Abstracts* (1959–) covers the most important articles on minerals; *Chemical Abstracts* provides a thorough coverage of mineralogic papers. *Bulletin Signalètique* provides an immediate brief abstract.

(J.V.S.)

Mining and Quarrying

Mining, in a broad sense, is the excavation of valuable mineral materials from the Earth's crust and includes minerals of organic origin, such as coal and petroleum. This article, however, describes only the mining of metallic and nonmetallic nonorganic materials; for coal and petroleum see the articles COAL MINING; PETROLEUM AND GAS EXTRACTION.

This article is divided into the following sections:

I. Development of mining techniques
  Early history
  Development of modern excavating technology
    The Americas
    Australia
    Europe
    Africa
    Soviet Union
II. Underground mining
  Prospecting and exploration
  Choosing a mining method
    Rock mechanics
    Evaluating geologic features
    Engineering-design factors
    Geology, rock structure, and geometry
  Mining operations
    Explosives
    Drilling
    Hoisting
    Loading and haulage
    Ventilation and lighting
  Mining methods
    Intervals between levels
    Underground opening support
    Open stopes
    Supported stopes
    Top slicing and sublevel caving
    Block caving
III. Surface mining
  Basic open-pit mining techniques
  Quarrying
IV. The mining industry
  Organization and management
  Conservation, ecology, and mining

Quarrying is the process of obtaining stone from surface excavations, such rock material being produced either in regular blocks (dimension stone) by cutting methods or in variable-sized fragments (crushed stone) by blasting with explosives. Excavation of gravel is considered by some to be a form of quarrying, and both are types of open-pit, or surface, mining.

The preliminary treatment of minerals, rocks, and ores after they leave a mine varies widely. Rock to be used for aggregate is crushed and screened. Most ores must be crushed before they are treated further. Some low-grade ores, including copper, are concentrated before smelting or further refining. Low-grade taconite (siliceous iron) ores must be ground, concentrated, and pelletized before they can be smelted to recover the iron.

Few minerals can be employed in their raw state, and most must be treated or refined before they are used. Nonmetallic minerals also may require purification. Hence, the development of mining has been necessarily paralleled by the development of treatment processes known collectively as extractive metallurgy. Modern mining has thus been made possible by the development of technology to a high degree both in mining and in the treatment of mined products. Mineral resources have played a dominant role in the political power of nations throughout recorded history, and the current world political situation emphasizes the importance of mineral resources in establishing high standards of living as well as influencing world politics.

I. Development of mining techniques

EARLY HISTORY

Archaeological discoveries indicate that mining was practiced in prehistoric times. Apparently, the first mineral man used was flint, largely for weapons. Certain minerals, such as ochre, black pigments (manganese ore), or chalk, were also mined and used as pigments.

Placer deposits

Gold was one of the first metals utilized, being mined from stream beds of sand and gravel where it occurred as a pure metal (placer deposits) because of its chemical stability. Although chemically less stable, copper occurs in native form and was probably the second metal discovered and used. Silver was also found in a pure state and at one time was valued more highly than gold.

According to historians, the Egyptians were mining copper on the Sinai Peninsula as long ago as 3000 BC, although some bronze (copper alloyed with tin) is dated as early as 3700 BC. Iron is dated as early as 2800 BC; Egyptian records of iron-ore smelting date from 1300 BC. Found in the ancient ruins of Troy, lead was produced as early as 2500 BC.

One of the earliest evidences of building with quarried stone was the construction (2600 BC) of the great pyramids in Egypt, the largest of which (Khufu) is 230 metres (750 feet) on a side and contains approximately 2,300,000 blocks of two types of limestone and red granite. The limestone is believed to have been quarried from across the Nile. Blocks weighing as much as 32,000 pounds (14,500 kilograms) were transported long distances and elevated into place, and showed precise cutting that resulted in fine-fitting masonry.

Agricola's *De re metallica*

One of the most complete early treatments of mining methods in Europe is by the German scholar Georgius Agricola in his *De re metallica* (1556). He describes detailed methods of driving shafts and tunnels. Soft ore and rock were laboriously mined with a pick and harder ore with a pick and hammer, wedges, or heat (fire setting). Fire setting involved piling a heap of logs at the rock face and burning them. The heat weakened or fractured the rock because of thermal expansion or other processes, depending upon the type of rock and ore. Crude ventilation and pumping systems were utilized where necessary. Hoisting up shafts and inclines was done with a windlass; haulage was in "trucks" and wheelbarrows. Timber support systems were employed in tunnels.

DEVELOPMENT OF MODERN EXCAVATING TECHNOLOGY

Great progress in mining was made when the secret of black powder reached the West, probably from China in the late Middle Ages. This was replaced as an explosive in the mid-19th century with dynamite, and since 1956 both ammonium nitrate fuel-blasting agents and slurries (mixtures of water, fuels, and oxidizers) have come into extensive use. A steel drill with a wedge point and a hammer were first used to drill holes for placement of explosives, which were then loaded into the holes and detonated to break the rock. Experience showed that proper placement of holes and firing order are important in obtaining maximum rock breakage in mines.

The pneumatic hammer

The invention of mechanical drills powered by compressed air (pneumatic hammers) increased markedly the capability to mine hard rock, decreasing the cost and time for excavation by several-fold. It is reported that the Englishman Richard Trevitheck invented a rotary, steam-driven drill in 1813. Mechanical piston drills utilizing attached bits on drill rods and moving up and down like a piston in a cylinder date from 1843. In Germany in 1853, a drill that resembled modern air drills (see below) was invented. Early drills required frequent repairs, as many as 200 being required to keep 16 in operation. Piston drills were superseded by hammer drills run by compressed air, and their performance improved with better design and the availability of quality steel.

Developments in drilling were accompanied by improvements in loading methods, from handloading with shovels to various types of mechanical loaders. Haulage likewise evolved from human and animal portage to mine cars drawn by electric locomotives and conveyers and to rubber-tired vehicles of large capacity. Similar developments took place in surface mining, increasing the volume of production and lowering the cost of metallic and nonmetallic products drastically. Large stripping machines with excavating wheels used in surface coal mining are employed in other types of open pit mines.

The Americas. Perhaps the first mining by Europeans in the Americas after Columbus was for copper in Cuba in 1554. Many small deposits of copper and iron were discovered in the 13 colonies that became the United States; all were producing small amounts of iron by the end of the Revolutionary War. Iron was first produced from the extensive Lake Superior deposits in 1846. Though much of the richer iron ore had been exhausted by the early 1970s, lower grade taconite ore is now being mined and concentrated by special processes.

California gold rush

While gold has been found in several places in North America, and the existence of gold in California was previously known, its discovery in 1848 had marked effects upon immigration and economics in the United States. The first gold mined was in placer stream deposits that proved to be extensive; some deposits were later mined by dredges. The gold deposits were traced upstream to the source from whence they had been eroded, which led to the discovery of the Mother Lode, a rich quartz vein of unusual length in the Sierra Nevada Mountains. Gold was later discovered in Alaska in 1880 on the beaches and in streams and in vein deposits in the mountains inland, in both Alaska and British Columbia.

One of the richest silver mines in America was located at the Comstock Lode in 1859, near what is now Virginia City, Nevada, an area that had been worked for placer gold. Several valuable deposits were discovered in Colorado beginning in 1859, yielding gold and silver from districts known as Leadville, Cripple Creek, and others.

Lead mines existed in what is now Missouri as early as 1687. The mines of southeast and southwest Missouri have supplied large amounts of lead and other metals. More recent discoveries in south central Missouri indicate the possibility of a lead belt across the state.

Lead was discovered in Bingham Canyon, Utah, in 1863, leading ultimately to the discovery and development of the great Utah copper open-pit mine. The successful exploitation of massive low-grade copper deposits of the western United States and elsewhere in the world was made possible by the discovery of the process of mineral flotation, which permits the economical concentration of the ore before it is smelted.

Wide-scale development of mining in Canada has been fairly recent. Iron ore was found in Newfoundland and

World Production of Major Metals and Nonmetallics, 1969

metal	country (percent of total production)
Aluminum	U.S. (37.9); U.S.S.R. (12.1); Canada (11.0); Japan (6.3); Norway (5.6); France (4.1); others (23.0)
Antimony	South Africa (27.9); Bolivia (20.1); China (18.3); U.S.S.R. (10.1); Mexico (4.8); others (18.8)
Asbestos	Canada (45.9); U.S.S.R. (26.8); South Africa (7.9); China (5.0); others (14.4)
Barite	U.S. (25.7); West Germany (12.5); U.S.S.R. (7.4); Italy (6.4); others (48.0)
Beryl	Brazil (39.1); India (18.1); U.S.S.R. (17.4); Argentina (7.2); South Africa (4.4); Uganda (4.0); others (9.8)
Bismuth	Japan (19.4); Peru (18.0); Bolivia (17.4); Mexico (15.8); Canada (8.5); China (6.5); Australia (5.3); others (9.1)
Cadmium	U.S. (33.6); Japan (16.2); U.S.S.R. (13.5); Canada (5.6); Belgium (5.0); West Germany (4.6); others (21.5)
Chromite	U.S.S.R. (33.3); South Africa (23.4); Philippines (9.2); Turkey (8.7); Rhodesia (7.1); India (4.2); others (14.1)
Cobalt	Zaire (49.3); Finland (8.1); Cuba (7.6); Morocco (7.6); U.S.S.R. (7.4); Canada (7.2); others (12.8)
Columbium-Tantalum	Brazil (65.4); Canada (20.6); Nigeria (11.5); others (2.5)
Copper	U.S. (23.4); U.S.S.R. (15.0); Zambia (12.5); Chile (11.6); Canada (8.3); Zaire (6.0); others (23.2)
Diamond	
Gem	South Africa (30.3); South West Africa (15.2); Angola (13.8); U.S.S.R. (13.4); Zaire (4.4); others (22.9)
Industrial	Zaire (48.1); U.S.S.R. (19.3); South Africa (14.4); Ghana (8.1); Sierra Leone (4.3); others (5.8)
Feldspar	U.S. (29.7); West Germany (12.6); U.S.S.R. (10.8); Italy (9.2); France (7.4); Norway (5.1); others (25.2)
Fluorspar	Mexico (26.1); U.S.S.R. (10.5); Spain (8.5); France (7.9); Thailand (7.9); Italy (6.8); China (6.6); U.K. (5.3); U.S. (4.4); South Africa (4.0); others (12.0)
Gold	South Africa (67.4); U.S.S.R. (13.5); Canada (5.2); others (13.9)
Gypsum	U.S. (17.9); Canada (12.5); France (10.2); U.K. (9.6); U.S.S.R. (9.4); Spain (8.0); Italy (6.6); others (25.8)
Iron ore	U.S.S.R. (25.7); U.S. (12.5); France (7.7); Canada (5.6); China (5.5); Australia (5.4); Brazil (4.6); Sweden (4.6); others (28.4)
Lead	U.S. (14.4); Australia (13.9); U.S.S.R. (13.9); Canada (9.4); Mexico (5.3); Peru (5.1); others (38.0)
Magnesium	U.S. (45.1); U.S.S.R. (22.4); Norway (15.9); Canada (4.8); Japan (3.1); others (8.7)
Manganese	U.S.S.R. (38.8); South Africa (12.2); Brazil (10.9); India (8.9); Gabon (7.6); China (5.5); Australia (5.1); others (11.0)
Mercury	Spain (22.5); Italy (17.1); U.S.S.R. (16.5); U.S. (10.3); Mexico (7.9); Canada (7.0); China (7.0); Yugoslavia (5.0); others (6.7)
Molybdenum	U.S. (63.2); Canada (19.2); U.S.S.R. (9.5); Chile (6.8); others (1.3)
Nickel	Canada (40.0); U.S.S.R. (21.9); Cuba (7.3); others (30.8)
Phosphate rock	U.S. (44.6); U.S.S.R. (21.1); Morocco (12.5); others (21.8)
Platinum group	U.S.S.R. (62.0); South Africa (28.5); Canada (7.9); others (1.6)
Potash (K_2O equivalent)	U.S.S.R. (19.1); Canada (17.2); West Germany (15.6); U.S. (15.3); East Germany (13.8); France (11.7); others (7.3)
Selenium	U.S. (44.5); Canada (25.8); Japan (15.8); Sweden (6.4); Zambia (2.1); others (5.4)
Silver	Mexico (14.9); Canada (14.5); U.S. (14.5); U.S.S.R. (12.8); Peru (11.8); Australia (8.5); others (23.0)
Tellurium	U.S. (54.8); Canada (24.4); Japan (11.9); Peru (8.9)
Tin	Malaysia (32.3); U.S.S.R. (12.1); Thailand (9.3); China (8.9); Indonesia (7.7); Australia (3.5); others (26.2)
Titanium	Australia (30.0); U.S. (24.0); Canada (19.0); others (27.0)
Tungsten	China (24.4); U.S.S.R. (19.8); U.S. (13.1); North Korea (6.6); South Korea (6.0); Bolivia (5.6); Canada (4.5); others (20.0)
Uranium oxide	U.S. (51.4); Canada (18.1); South Africa (17.0); France (7.8); others (5.7)
Vanadium	U.S. (49.1); South Africa (25.3); Finland (13.1); Norway (8.5); others (4.0)
Zinc	Canada (22.4); U.S.S.R. (11.5); Australia (9.5); U.S. (9.4); Peru (5.9); Mexico (4.8); others (36.5)

Source: U.S. Department of Interior, *Minerals Yearbook*, vol. 4, 1969.

gypsum in Nova Scotia. Much of the mineral wealth is found in the Canadian Shield (pre-Cambrian rocks, 570,000,000 years old). Large deposits of ilmenite (titanium-iron ore, Tahawus, New York), gold (Kirkland Lake), copper (Sudbury Basin), and iron (Lake Superior) are currently being mined, as well as gold deposits in the West (Yellowknife).

In South America, Chile is noted for its deposits of sodium nitrate (1880–1930) and copper. The economic manufacture of synthetic sodium nitrate has, however, decreased the value of Chilean deposits. Copper was mined by the Spaniards as early as 1601, but most production now comes from massive low-grade deposits. Chile also has several rich iron-ore deposits, as well as some sulfur. Bolivia is the world's second-largest producer of tin. Gold, silver, and other metals and minerals are produced in other Latin American countries.

Australia. Gold was discovered in 1851 in New South Wales and later in Victoria, but after 1857 opportunities for miners to make a living decreased. The Broken Hill, Mount Isa, and Kalgoorlie gold mines have been most productive. Copper, silver, zinc, and tin are also produced in varying amounts. Australia leads the world in production of opals. In the past decade, 1960–70, active exploration led to discovery of extensive new deposits of metallic and nonmetallic mineral deposits.

Europe. France has notable iron deposits, while England's deposits of coal and iron have made it a source of steel. Open-pit operations in western Europe include mining of iron, limestone, granite, and shale. In France the Minette iron deposit, previously exploited by open-pit methods is now mined by underground techniques. Ammonium nitrate fuel oil and slurry blasting agents have been widely adopted in some European countries, though their use is restricted in some. Stripping, mining, loading, and transport are practiced in much the same manner as in the Americas, with some differences due to geography and local conditions.

Africa. The continent of Africa is the source of about 84 percent of the world's diamonds (major producers are South Africa, South West Africa, Angola, Zaire, Ghana, Sierra Leone) and over one-half of the gold (South Africa). Cobalt (65 percent of world supply; Zaire and Morocco), chrome ore (30 percent; South Africa and Rhodesia), and manganese (25 percent; South Africa and Gabon), all important in strengthening steel, are other products of this area. The huge copper-mining operations in the Shaba Province of Zaire became a source of political difficulty following that nation's independence in 1960. In western Africa, high-grade iron deposits, magnetite and hematite, are mined by the open-pit method. Other mining practices generally resemble those employed in Europe.

Soviet Union. Besides extensive reserves of coal, the Soviet Union probably has a greater variety of minerals than any country outside the United States. The Krivoy Rog deposits, similar to the Lake Superior deposits, have produced large tonnages of iron, which is also found in the Ural and Crimea areas. The Soviet Union leads the world in manganese reserves and has deposits of copper, lead, zinc, and aluminum, as well as nickel, gold, and platinum. Nonmetals include potassium and phosphate salts for production of fertilizer, asbestos, talc, magnesite, salt, and gems. See the Table for a listing of leading mineral producers.

II. Underground mining

PROSPECTING AND EXPLORATION

In the past, many deposits of valuable minerals were found by means of outcrops, but few such deposits near the surface remain to be discovered. Currently, most new deposits are located by means of geological techniques,

often supplemented by geophysical and geochemical methods. Prospecting is also being conducted to find deposits under the ocean. An understanding of the processes by which deposits were formed is a valuable aid in prospecting, since geologic controls can furnish information as to the likely position of ore.

The first step in modern prospecting is the compilation of all existing geological information on the area of interest. A detailed study and mapping of surface features follows; these are correlated with known structures that are projected to depth where possible. All evidences of mineralization, favourable structures, or environment for deposits are carefully recorded and employed as indicators of ore. If evidence of the possible presence of valuable deposits is found, a drilling program is carried out to obtain more information about subsurface geology and to intersect ore structures if possible. Data obtained in prospecting may also be employed in the exploration or evaluation stage that follows.

Indicators of ore

Types of geological evidence are both regional (of broad extent) and local (of limited extent). For metallic deposits, the presence of batholiths (large, irregular masses of igneous rock that have melted or forced their way into surrounding strata), certain volcanic and igneous rocks, the age of rock relative to mineralization periods, presence of favourable rocks, and climatic conditions may be important. Local surface features that can provide valuable clues are presence of rock alteration, trace mineralization in surface rocks, and quartz and other gangue (non-ore-bearing) minerals. Rock fractures, particularly fault zones, are important, as well as beds or zones that are favourable for mineralization. Many deposits "bottom out" with depth because pressure or other factors were unfavourable for mineral deposition.

Geochemical prospecting is the chemical analysis of rocks to determine the presence, usually in very small amounts, of possible trace elements. The patterns of dispersion of certain elements have proven to be valuable in conjunction with geologic data in locating and defining the limits of valuable ore bodies.

Geophysical methods employed vary with the local conditions. Seismic techniques, which involve transmission of low-frequency waves through the Earth's crust and recording of their behaviour, are utilized to determine underground structure, particularly in horizontal-bedded deposits. Magnetic methods determine the presence of magnetic bodies. Electrical methods indicate the presence of conductive materials. Gravity methods detect the presence of more dense bodies. Geiger counters ascertain the presence of radioactive minerals.

Obtaining cores

Prospect drilling is often performed with diamond drills to obtain cylindrical specimens of the rock and ore, called cores. The drill bit is a hollow cylinder embedded with commercial-grade diamonds on one end and attached to a core barrel. The drill and core barrel are rotated with the bit against the rock by means of hollow drill rods through which water flows to keep the bit cool and wash out the cuttings. Diamond drilling is of greatest value where the rock can yield intact cores that can then be employed to map underground geology or to give solid samples of valuable rock or ore.

When a valuable deposit is discovered, prospecting phases into exploration, which is planned to yield necessary data concerning size, shape, and value to furnish information to begin development. Careful trenching (digging a trench to obtain samples of ore lying near the surface), drilling, and sampling are carried on to give reliable information before large capital expenditures are made or the necessary plant is purchased.

Modern computers are often employed to calculate possible risks, profits, and other results of further exploration. Sample values obtained for the material to be mined are usually evaluated by use of the theory of probability, with average values weighed by the amount of ore they represent or by a similar method.

A mineral deposit is a "wasting asset," and, hence, its present value is calculated on the basis of the fact that the worth of the mine will be zero at the end of a certain period of time. The cost of capital equipment must be amortized accordingly, and a reliable estimate of the life of the mine is made. Rates of interest and return of investment are calculated to ensure adequate profits before the time a deposit is mined out.

CHOOSING A MINING METHOD

A space in an underground mine from which ore has been removed is called a stope. Depending upon conditions, stopes require varying amounts of support to prevent the earth and rock above (called overburden, or capping) from collapsing. In some mines, ore is extracted from several different underground levels. In this case, intervals must be left between the levels to prevent collapse. Figure 1 is a composite sketch illustrating various terms used in underground mining.

From R.S. Lewis, *Elements of Mining*, 3rd ed. (©1964); used with permission of John Wiley & Sons, Inc.

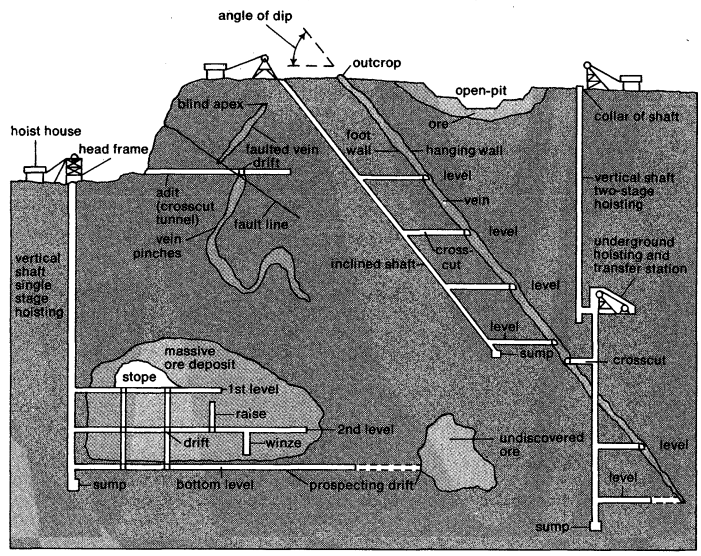


Figure 1: Illustrations of terms used in mining.

Rock mass stability

The stability of rock masses in which ore deposits occur in many cases determines the method of mining employed, although other factors may dictate the method used. Geologic and rock-mechanics data (see below) are the most reliable guides for evaluation of the adaptability of mining methods to a given deposit. Ideally, a method of mining is employed that will yield the largest net financial return consistent with safety and optimum extraction of valuable minerals.

Rock mechanics. The term rock mechanics is used to define the study of the behaviour of rock materials as they relate to man-made engineering projects that involve rock in place. Thus, studies are made of the physical properties of rock specimens that are taken from a rock mass. Many intact rocks behave elastically under load, and the theories of elasticity and the strength of materials are applied. Compressive and tensile strength are measured, as well as the strength under confining pressure (see MATERIALS TESTING).

Dynamic rock mechanics deals largely with waves caused by explosives and their effects in rock breakage, with rock breakage in drilling, with mechanics of rock bursts (sudden collapses), and with effects of large, underground explosions.

Methods have been devised to measure stresses, strains, and displacements in rock, both in open pits and underground. These studies are made to determine the stability of rock structures, support requirements, the behaviour of subsiding (settling) masses in caving and longwall methods of mining, and for increasing the efficiency and safety of mining.

Evaluating geologic features. Geologic details are carefully evaluated so that preliminary openings may be placed to best advantage; permanent openings may be located in stable rock; stopes may be planned for optimum stability and ground control; mine openings may be maintained at minimum cost; underground water may be

avoided; overall mining methods, including caving procedures, may be planned; and safety can be achieved.

Geologic features that influence the choice of mining method include the size and shape of the ore body; the depth and type of overburden; the location, strike (horizontal course), and dip (the angle at which the deposit is inclined from the horizontal) of the deposit; the strength and physical character of the ore; the strength and physical character of the surrounding and overlying rock; the presence or absence of aquifers (water-carrying rocks); and the grade and type of ore.

Engineering-design factors. Important factors in engineering design of underground mines include strength of ore and wall rocks; shape, horizontal area, volume, and boundary regularity of the ore body, and the thickness, dip or pitch or both of the deposit and ore shoots; continuity of the ore; depth below the surface; nature of the overburden; and position of the deposit relative to surface improvements, drainage, and other mine openings. It is usually necessary to minimize those factors of lesser importance in order to satisfy the requirements of the more critical ones.

The strength of both ore and wall rocks fixes the safe size of openings, the length of time they will stand, and the support required. Stability depends not only on the strength of the solid rock but also on the existence of fractures, joints, and other discontinuities and their geometric arrangement.

Ore body shapes

Shapes of ore bodies vary from massive to tabular (having two parallel surfaces) and from bedded deposits to pipes, shoots, dikes (fissures in older rock filled with igneous material), and sills (strata of igneous rock found between beds of other rock), depending on their origin and the character of the host rock structure. The attitude (inclination) of a deposit directly influences the mining method. A tabular deposit dipping at a high angle, for example, may be removed without use of supports if the walls are moderately strong. The roof of a similar deposit dipping at a low angle, however, may need support.

A regularly shaped deposit may require less support. Mining operations in an irregular deposit will produce stopes with uneven walls, overhanging slabs, and projecting columns of wall rock that may require support if the rock is fractured or broken.

A small ore body in firm rock may be mined without support, but larger ones in the same kind of groundmass may require it. There is a limit to the length or width of an unsupported rock arch or span that will possess sufficient strength to stand.

Usually, as depth increases, a method must be changed because openings require more support. Shallow deposits are subject to relatively low lithostatic (weight of overlying rock) pressures. The vertical pressure on mine openings usually increases proportionately with depth. The direction of pressure in walls or roofs of an excavation depends on the attitude of the deposit and local geologic factors and must be considered in stope support. In a steeply dipping tabular deposit, waste (broken-rock) filling will furnish adequate support against side pressure, whereas in a wide deposit, in which the critical pressure acts downward, filling cannot be relied on to prevent all movement. Thus, a cut and fill method can be employed where the pressure is from the sides and the back is strong and self-supporting, but if the critical pressure is vertical, square setting (a heavy square framework of timbers) supplemented by filling or some other method of support is necessary (see below *Underground opening support*).

Geology, rock structure, and geometry. In addition to their value in exploration, geological theories of the genesis of mineral deposits are also valuable guides in mine operation, influencing prospecting, exploration, development, and exploitation. In prospecting, geology is a vital tool in finding new ore bodies and is most effective in the hands of an experienced mining geologist. In exploration and development, the mine engineer should have a knowledge of strength of materials, rock mechanics, and of the structural features of the deposit that will affect the production of ore and the margin of profit. Most of these factors can be treated in only a semiquantitative manner, because the behaviour of rock structures is complex. Design details may also be considered in relation to bedded deposits, veins, and massive deposits.

Bedded and sedimentary deposits that have not been disturbed by faulting, folding, or other action are, like flat-lying beds of coal, simple to mine. Support problems are often easily solved, and a systematic method of pillar support is employed. Poor rock in the immediate roof, however, such as shale, creates hazardous conditions. The most serious problem in this type of mine may be determination of the number, size, and position of pillars required for support.

Dangers of shale roofs

Many commercial mineral deposits are found in or closely associated with veins in fractures, shear, and fault zones. Deposits vary from tabular veins to pipelike bodies and shoots and may be regular or irregular, wide or narrow, continuous or discontinuous. Thus, many different methods are used to mine them.

The large, disseminated copper-ore bodies of the western United States are typical massive deposits; those with enough overburden or capping to prohibit mining by open-pit methods are often mined by block-caving (see below). Deposits of porphyry coppers are usually extensive, flat-lying, and relatively regular in outline; though low-grade, their reserves are usually measured in millions of tons. The concentration of valuable minerals has been the result of geologic changes that left behind a weak, leached overlying rock mass.

MINING OPERATIONS

These include use of explosives, drilling, hoisting, loading and haulage, and ventilation and lighting.

Explosives. The theory of detonation of high explosives is based upon the laws of conservation of mass, momentum, and energy applied to plane shock waves, combined with thermodynamics, equations of state for gases, chemical equilibria, and material-balance equations (see EXPLOSIVES). In a detonation wave, the energy of chemical reaction is utilized to heat the detonation products and to maintain the thermodynamic processes of shock-wave propagation at high temperatures (to 5,500° K [9,400° F]) and pressures (to 200,000 times atmospheric pressure). The distinguishing feature of a detonation is that it is a reactive shock—*i.e.*, a shock wave supported by an accompanying chemical reaction.

Most common commercial explosives are made up of compounds containing carbon, hydrogen, nitrogen, and oxygen, some of the compounds serving as fuels and some as oxidizers (agents that bring about chemical combination with oxygen). The compounds must be of such a character that they will react rapidly, releasing chemical energy at a very high rate. Thus, one stick (eight ounces, or 225 grams) of an average-energy dynamite may have a work capacity of 40,000 horsepower.

Ammonium nitrate–fuel oil explosives

Development of ammonium nitrate–fuel oil blasting agents in the late 1950s has reduced explosive energy costs to as little as one-third the cost of dynamite. Though it had been known since the late 1800s that ammonium nitrate mixed with a fuel would explode in large diameters, the corrosiveness, affinity for water, and tendency of ammonium nitrate to cake, plus its large critical diameter (the minimum diameter at which it will propagate a stable detonation), made it unusable. In the 1950s it was found that commercial-grade ammonium nitrate fertilizer prills (small spherical grains) would detonate in reasonably small diameter holes when they were mixed with finely divided carbon or about 6 percent by weight of fuel oil.

Whereas solid ammonium nitrate crystals are difficult to detonate, prills are cracked and fissured because the last 5 percent of moisture is evaporated after the ammonium nitrate is in prilled form. This leaves much more surface area exposed to chemical reaction, which is the primary reason for the success of ammonium nitrate–fuel oil as a blasting agent; *i.e.*, it results in a smaller critical diameter. Equivalent in strength to about a 40 percent dynamite, ammonium nitrate–fuel oil cannot be detonated by blasting caps but requires a strong booster to initiate an explosion. This feature classifies it as a blasting agent rather than an

explosive. Thus, it is safe to handle, can be mixed at or near the blasting site, can be loaded in dry holes in bulk from automatic loaders, and costs much less than dynamite. Disadvantages mentioned above include its affinity for water, corrosive action on metals, and tendency to cake. It cannot be used in wet holes because of its solubility except in waterproof bags or as a slurry.

While the solid density of ammonium nitrate is 1.725 grams per cubic centimetre, the bulk density of prills varies from 0.85 to 0.95 gram per cubic centimetre. Detonation velocity increases almost linearly with bulk density, and the question arose during its development whether the increase in density by filling pore spaces between prills with water could be compensated by other thermodynamic factors. It was found that, if ammonium nitrate, trinitrotoluene (TNT), and aluminum were mixed with water, the increase in density caused an increase in detonation pressure, and the favourable heat of reaction of aluminum plus other thermodynamic factors gave detonations of high velocity, higher pressure, but lower temperature. Because of these characteristics, slurries of various compositions have found wide application in rock and ore that is difficult to break. They may also be used in wet blast holes.

Blasting caps

Explosives are usually detonated with electric or nonelectric blasting caps that contain a small amount of powerful, relatively sensitive explosive. Detonating fuse, a train of explosive with a small critical diameter that will propagate a detonation, is also employed. Blasting agents, which are insensitive to blasting-cap initiation, are detonated by means of a booster, a charge of up to eight ounces (225 grams) of high explosive, which is itself initiated by a blasting cap.

An explosive confined in a borehole transmits much of its explosive energy to the surrounding rock in the form of a compressive "stress pulse." When this reflects from a rock surface in tension, it causes the rock to break in slabs, inasmuch as rock is only about one-tenth as strong in tension as it is in compression. The type, amount, and spacing of explosive charges are determined from experience or by experimentation.

Drilling. Drilling for blasting purposes (Figure 2) is performed with several types of drills. Percussion drills impact the rock with a chisel-shaped or button-faced bit

By courtesy of Gardner-Denver Company, Quincy, Ill.

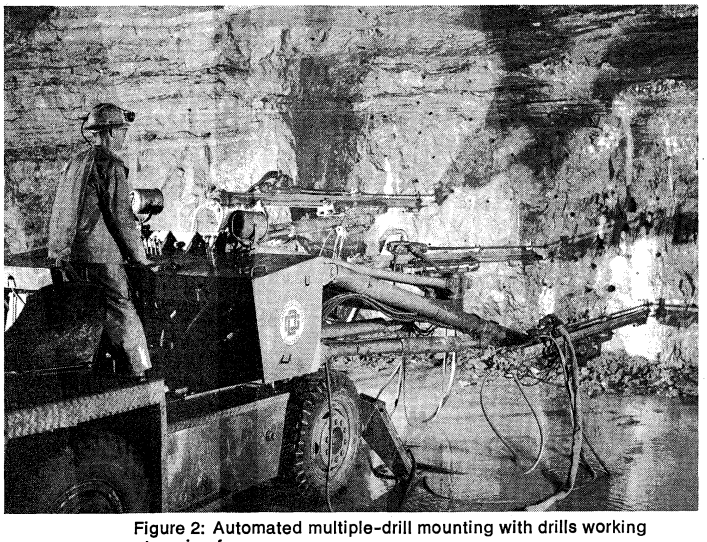


Figure 2: Automated multiple-drill mounting with drills working at a mine face.

(containing button-shaped inserts of extremely hard tungsten carbide) that chips the rock with repeated blows as it is rotated. This type of drill usually employs compressed air for power. Drag-bit rotary drills cut by shaving, plowing, or abrading with a rotating cutting head using blades, diamonds, or hard steel shot as cutting agents. Drag action is provided by thrust and rotation. Rotary-bit drills utilize a cutting head mounted with roller cones studded with teeth or hard metal buttons; they cut the rock by indenting, shaving, fracture, and crushing. Rotary–percussion is a combination of the two methods described above.

Factors influencing mechanical drilling are the condition of the equipment, depth, and diameter of the hole, rock properties, geologic conditions, and the stress in the rock.

In open pits, larger holes are drilled for explosives than underground; most large holes in surface mining are made with rotary bits. Some drilling a few decades ago was performed with churn drills, which fragment the rock in the bottom of a hole by repeatedly raising and dropping a heavy drill bit. Cuttings are removed from the hole with a long pipe with a closure device on one end. Until 1949 most rotary drilling employed drag bits. Modern rotary drills attack the rock with roller cones with hard teeth or knobs, the bit assembly being supplied with energy by means of rotating drill rods. Rock is broken by processes of abrasion, chipping, and scraping, or a combination of these. Circulating water is used to clean the hole and carry the cuttings to the surface. The drill is operated by a rig that may be mounted on a truck, trailer, or crawler. Power is furnished by a gasoline or diesel engine or electric motor. Hole diameters for surface mines may vary from nine to 15 inches (23–38 centimetres), with the spacing between holes being 15 to 40 feet (5 to 12 metres).

Flame-jet drilling

In hard rock that spalls (chips at the surface) readily when heated rapidly, flame-jet or jet-piercing drilling may be employed. The drill consists of a jacketed blowpipe that is water cooled. Fuel oil and oxygen are burned in a combustion chamber generating a high-temperature flame. The heat spalls the rock, and the cuttings are carried from the hole by the steam from the cooling water. Quartzitic rocks containing 20 percent or more of quartz are usually most favourable for spalling. Holes cut by flame jets may be somewhat irregular and require a free-flowing dry explosive or slurries to fill them completely. Flame jets may also be used to cut deep channels in dimension stone rock that spalls readily when attacked by heat.

Most underground drilling and blasting ulitizes small holes with diameters of one to three inches (3 to 8 centimetres), which are cut with either pneumatic or noncoring diamond drills, the latter often being employed for deeper holes. Also, both of these types of equipment can drill in any direction.

Pneumatic drills, powered by compressed air, consist of a heavy cylinder with a reciprocating piston that strikes a hollow drill rod with a bit on the opposite end. The drill rod fits into a chuck that rotates; water is conducted through the machine and drill rod to cool the bit and wash cuttings from the hole. Pneumatic drills may be hand-held (jackhammer) or mounted on a pneumatic leg (stopers) or on a column or jumbo (drifter). Modern jumbos for mounting drills are mobile, and the drilling by two or more machines may be controlled from a console on the vehicle. Pneumatic drills are employed primarily for shorter holes because of the inefficiency of transmitting impact energy through long rods.

For drilling very large holes, such as the shaft in a uranium mine, drills of the type shown in Figure 3 may be employed.

Hoisting. Hoisting from underground mines was first accomplished with a windlass. A later type of hoist was the horse whim, comprised of a rope or cable around a drum turned by a bevel gear attached to a sweep to which a horse was harnessed. Mechanical hoists are powered by internal combustion engines, compressed air motors, or —usually—by electric motors.

Most hoists have a double drum to hoist from a two-compartment shaft. The cables are wound on the two independent but mechanically connected drums in opposite directions to permit hoisting in balance. The cables run up over sheaves (grooved pulley wheels) on the top of a headframe to cages (see below) or skips attached on each. As one cage or skip is hoisted, the other descends, thus utilizing the weight of the empty to assist in hoisting the loaded one. To keep the peak power demand at a

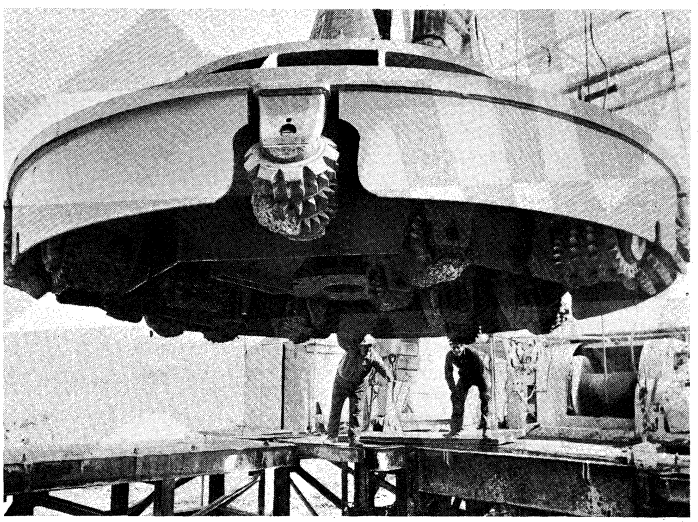


Figure 3: Shaft drilling rig for moderately hard rock. Drilling bit is 16½ feet (five metres) in diameter and weighs nearly 40 tons (36,000 kilograms). It is being used here to bore the shaft of a uranium mine.
From *Engineering and Mining Journal* (June 1970); © McGraw-Hill Inc.

lower level, a control device that utilizes energy stored in a large flywheel is employed on some hoists.

Hoisting drums may be either single or double and cylindrical or conical in shape. A conical shape permits a low starting torque on the motor and thus helps hold down the peak power load. If a single-compartment shaft is used, a counterweight may be employed, similar to that in a building elevator system, which in turn permits a smaller size motor.

Hoisting ropes

Most hoisting ropes are made of multiple-strand, twisted steel wire that must be strong enough to lift loaded skips and cages in addition to sustaining the acceleration and friction forces of the hoisting system. The hoisting drum must be large enough to keep the bending stresses of the cable around the drum below critical values. Hoisting ropes must be inspected frequently and discarded before they are weakened below the allowable strength.

Loading and haulage. Early loading of ore and rock was done by hand into baskets or similar containers that were then carried out of the mine. Wheelbarrows were employed in early mining and are currently used in small openings; animal haulage, which found early use, is still employed in small mining operations. Hand loading, or mucking, is likewise used at the present time, particularly where the cost of mechanical loaders is not justified.

Modern loading in horizontal room and pillar mines (in which a large horizontal ore body is mined, leaving some ore behind in the shape of pillars to support the roof), drifts (horizontal openings in or near an ore body and parallel to its long dimension), and tunnels is accomplished with various types of mechanical loaders mounted on either track, tractor tread, or rubber tires. Manual pushing of small cars on narrow-gauge tracks is called hand tramming. Broken ore and rock may also be moved along the floors of stopes or along tunnels by means of a scraper, a large hoe drawn back and forth by means of a cable through an anchored sheave (pulley wheel), the system being powered by a small hoist. Broken material is scraped into chutes or over ramps into mine cars.

Mine cars are heavy sheet-metal boxes on wheels with capacities from about 2,000 to 20,000 pounds (900 to 9,000 kilograms). They are usually drawn by electric locomotives, either battery-powered or powered through a trolley to an overhead wire. All are designed so that the broken rock or ore can be dumped outside of the mine or into a bin cut in rock, often lined with timber (skip pocket) at the hoisting shaft. For large mines on one level, truck haulage may be used. Sufficient ventilation must be provided to dilute or carry out the exhaust gases from the internal-combustion engines.

Ventilation and lighting. Underground mines must be adequately ventilated to provide a breathable atmosphere, to carry out noxious or explosive gases, and to provide cool air in hot mines. One primary purpose of mine ventilation is to make certain that there is enough oxygen in the working places for men to breathe. In gassy bituminous-coal mines, large amounts of air are required to dilute methane gas released from the coal. Noxious gases are generated in some metal mines by the action of acid water on rocks and minerals as well as by explosives used for blasting. Radon gas must be removed from uranium mines because of its radioactivity. Many mines are excavated in rock masses where temperatures are as high as 140° F (60° C) or more, sometimes with high humidity conditions. The air pumped through many such mines is cooled and dehumidified, a costly process.

The main ventilation system of a mine usually requires a minimum of two openings to the surface, with a fan at or near one opening. Air may either be forced through the mine by a positive pressure or pulled through by exerting suction on one opening. The fresh air is guided through the mine by stoppings, doors, and bulkheads. Auxiliary ventilation by small fans and ventilation ducts picks up fresh air and forces it into dead-end drifts, raises, stopes, and other areas that cannot be ventilated by the main air current. If fires are burning in a mine (old timbers, etc., in mined-out areas), these areas must be sealed off from the oxygen supply in the air currents and objectionable gases contained by bulkheads or stoppings.

Auxiliary ventilation

Costs of mine ventilation and air conditioning may constitute a significant part of the cost of producing ore, thus affecting the cost of the final mine product to the consumer, the grade of mineral that can be mined, and the margin of profit to the mining company.

For primary ventilation of a mine, a fan must be chosen that will deliver the necessary amount of air at the proper pressure. The amount of air a fan can deliver varies with the resistance to airflow encountered in the mine. This resistance changes with the extent of excavation, so fans have been designed whose characterstics can be changed by altering the pitch of their blades, or the fans may be designed to operate efficiently over a fairly wide range of pressure and quantity of air flowing.

Natural ventilation will take place in a mine if the two openings are at different elevations at the surface and if the outside temperature is different from the mine temperature, the flow being caused by the difference of weight of cold and warm air (thus producing convection currents). Natural air currents often reverse direction from summer to winter.

Air in workings must be treated to minimize the presence of harmful dust. Fine coal dust may cause "black lung" disease, while fine silica and asbestos may cause silicosis or asbestosis if breathed over a long period. Some dusts may be settled with a water spray, but others are difficult to remove from the atmosphere. Maintaining the quality of mine air is essential to the health of the miners and has been the subject of intensive research and development. Conditions have improved markedly in the past few years. Most countries have comprehensive legislation regarding health and safety in mines; labour unions may refuse to work under dangerous conditions. In view of the vital role that mining plays in the economic and strategic welfare of a country, the importance of supporting mining operations with competent scientific and engineering research is evident.

Most of the shaft stations and main haulage ways in mines are lighted with electric lights. Mine locomotives, trucks, and similar vehicles have lights comparable to those of transports operating on the surface at night. Workmen underground have lights, usually worn on their caps. Early lights were candles, wick lamps, or carbide lights. Most miners currently have electric lights powered from a battery on their belts through an electric cord. Good lighting is essential to safe working conditions as well as to efficient production.

MINING METHODS

In mining, the intervals between different underground levels and underground opening supports must be carefully selected. Mining methods include open stoping, sub-

level stoping, supported stopes, top slicing and sublevel caving, and block caving.

Intervals between levels. In ore deposits of hundreds of feet in vertical dimension, the interval between haulage levels should be the distance that gives the least cost per ton; that is, the value of ore to be mined from a given level must exceed costs of driving the level. The main cost items include sinking shafts and drifts (horizontal openings in or near an ore body, parallel to the course of its long dimension); making crosscuts to the ore; level equipment such as fans, track, air and water lines, and power lines; mining pillars; raises (vertical mine openings driven from one level to the other) and ore passes needed; maintenance of drifts, crosscuts, and raises for the life of the level; hoisting ore to the surface; pumping; handling ore and supplies in stopes; ventilation; waste filling; supervision; and interest on capital represented by broken ore in the stopes.

Underground opening support. In underground openings where the rock in the roof and walls is too weak to stand alone, it must be supported. The simplest type of timber support is a one-piece set called a stull or post (Figure 4A and B). Two-, three-, and four-piece sets

From R.S. Lewis, *Elements of Mining*, 3rd ed. (©1964); used with permission of John Wiley & Sons, Inc.

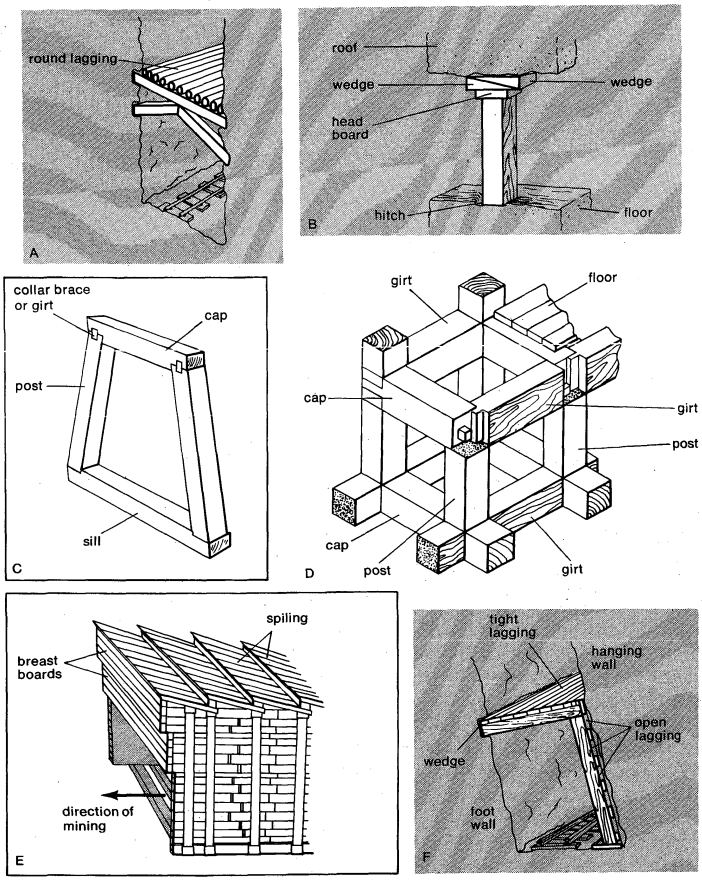


Figure 4: *Types of mining timbering.*
(A) Reinforced stull. (B) Post. (C) Four-piece set, sawed timber. (D) Square-set. (E) Timbering in running or loose ground. (F) Two-piece stull set.

(Figure 4C and F) are employed in small drifts and tunnels. In stopes that require close support, square sets (Figure 4D) are employed, either in narrow veins, wide veins, or larger, irregular deposits. Timber support may also be required where horizontal openings meet the shaft at a shaft station that must usually provide for two hoisting compartments, a manway, and compressed air and water lines in the shaft, in addition to chutes to load ore and support for headroom for unloading materials. Many modern shafts are circular; support may be provided with steel or timber. Concrete may be used in shafts and permanent openings as well as steel sets. In very heavy grounds, collapsible steel arches or sets have been employed.

Rock bolting

Rock bolting has found broad application, particularly in openings that must be supported for long periods. A rock bolt is a steel rod, threaded on one end and provided with a means for anchoring in a hole in the rock on the other end. The bolt is anchored in the hole and a plate or other retainer pulled against the rock with a nut on the threaded end. Rock bolts are employed in drifts and other types of openings to hold the rock structure together, to reinforce strata, and to hold key pieces of rock in place and thus stabilize the structure.

A fine aggregate mixture of portland cement, sand, and water called shotcrete may be used to stabilize the rock around an opening when it is sprayed on in sufficient thickness, and it hardens in place. The shotcrete fills in small voids, holds key fragments in place, and furnishes arch support for the opening. It is used mostly in tunnels.

When a drift is driven in running (loose) ground, a method of timbering called spiling, or forepoling (Figure 4E), is used to provide close support to the ground.

Open stopes. Pillars of ore or waste rock are left to support the back of stopes in deposits of low-grade ore, in which it is cheaper to leave such pillars than to use artificial support. The required support depends on the depth of the deposit or the total weight on the pillars; the nature of the roof, which controls the length of unsupported span; the nature of the floor, which controls the bearing area and consequently the size of the pillar; and the strength of the ore, which determines the minimum size of pillars.

This method is applicable to geologic conditions that require a minimum of artificial support, the walls and roof span or both being self-supporting. Small ore bodies may be mined from wall to wall without pillars. Where ore bodies are larger, pillars of ore are left to keep the roof span to a safe dimension. Pillars may be spaced regularly, as in deposits of salt or phosphate, or randomly spaced in irregularly occurring ore (Figure 5), and the method may be applied to either horizontal or inclined deposits. Its greatest application is in flat-lying, bedded deposits.

The pillars are usually roughly circular, square, or rectangular in section and flared at the top and bottom to increase contact area with the roof and floor. In weak ground, the spacing is less, and the pillars are larger than in strong ground. The ore left in the pillars is sometimes recovered later. Some extraction of high-grade pillars has been accomplished with the roof supported by concrete columns, which must be preloaded by forcing them against the roof before it collapses.

In the development and mining of a flat-lying deposit, one or more openings are driven in a selected direction, and rooms are excavated usually at 90° to the openings. The pattern of layout of rooms may vary considerably, however. Pillar heights vary from a few to over 100 feet (30 metres), depending upon the thickness of the deposit.

The geology of typical open-stope mines is notable for simplicity of structure, lack of faulting, lack of alteration,

From R.S. Lewis, *Elements of Mining*, 3rd ed. (©1964); used with permission of John Wiley & Sons, Inc.

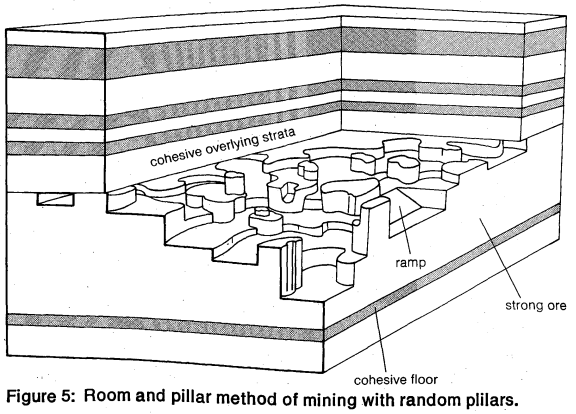


Figure 5: Room and pillar method of mining with random plilars.

and competency of the beds and the strength of the overall structure. Although dimensions of the deposits are variable, the method has uniform application throughout the deposits as long as the ore and rock are strong enough to support large openings. The placing or size of pillars is governed almost entirely by the presence of faults, fractures, or weak beds. Where weaknesses do not occur, the pillars may be placed regularly. Natural pillars may also be left to support the walls of a deposit that dips steeply. In this case, levels are driven from a shaft, and the ore is mined between the levels with natural inclined pillars left for support. Pillars may later be taken out ("robbed") and the walls allowed to collapse.

Robbing pillars

Sublevel stoping has many variations but is best suited to deposits that are steeply inclined and have strong ore and stope walls. In this method, there are two horizontal drifts (the main levels) 100 to 200 feet (30 to 60 metres) apart vertically. These drifts are connected by raises (vertical openings) that are separated about the same distance. At intervals of 20 to 25 feet, sublevel drifts, between the main horizontal drifts, are driven from the raise to the end of the stope. The main volume of ore is mined from the sublevel drifts. The ore and stope wall must be strong after the ore is excavated to furnish footing for workmen and support for sublevel drifts. After the main blocks of ore have been mined, the pillars are usually removed after which the walls may collapse.

Supported stopes. In shrinkage stoping the back (roof) of the stope, which is usually ore, is unsupported during mining, but the walls are supported by broken ore. Deposits may or may not need support during mining. The ore should be strong, but the walls may be weaker than those mined by sublevel stoping.

The first stages of shrinkage stoping are similar to those of sublevel stoping. There are no sublevels, however, and horizontal sections of the ore are mined from the bottom upward. The broken ore is left in the stope to support the walls during mining, and to serve as a floor for miners removing the ore overhead. Since the ore takes up more space when broken, about a third is drawn off as it is mined. After a stope is completed, the remaining ore is taken out.

Since the ore is unsupported during mining, it must be competent (*i.e.*, it must not yield or fail under stress). If walls are too weak, rock will slough off, diluting the ore. Hence, in deposits minable by shrinkage methods, the overall structure is slightly weaker than the ore and rock in which sublevel stoping is used. The footwall (the wall of rock under a vein of ore) of the stope should be steep enough to permit drawing the ore by gravity.

Shrinkage stoping has the advantage that blocks of ore may be quickly developed and mined, but considerable broken ore must be left in the stopes during mining, sorting may be difficult, and some sulfide ores may catch fire before they are removed.

Cut-and-fill method

In the cut-and-fill method (Figure 6), the ore is removed in parallel or stepped slices, and, as each slice is removed, a layer of waste (broken rock) fill is placed in the stope, leaving working headroom above it. Its applicability is similar to that of shrinkage stoping. The ore must be strong so that the miners can work under it. The walls do not need to be as strong, since the fill provides support. The fill may be in the form of broken rock from the mine. In some mines a sand and water mixture is pumped through pipelines into the stopes and the water is drained off. Because it settles less, sand fill is usually more satisfactory than broken rock. Sometimes the ore is of such a type that dilution cannot be allowed, and, if the ore oxidizes readily, causing difficulty in subsequent processes, shrinkage stoping cannot be used.

Shrinkage stoping is also used widely in steeply inclined deposits where broken ore can be removed by gravity and the stopes filled from the level above. In such deposits, even minimum subsidence (gradual settling) should not occur adjacent to the ore body; thus side support is given by immediate filling. Occasional slabs or loose sections of the back are supported by timbers placed on the fill. In large ore bodies, cut and fill stopes are kept small by spaced vertical pillars that are mined later, commonly by square-set stoping.

Before blasting, a horizontal or inclined layer of plank (lagging) is often laid over the waste to keep it from becoming mixed with the ore. The ore is drawn off at the level below and the waste is brought down from the level above.

Cut-and-fill stoping permits sorting and mining of irregular deposits and keeps ore dilution to a minimum. Ore production is intermittent, however, since filling is required, and the method is costly.

Square-set stoping, which employs timber frameworks (square sets) for support underground, is used most with ores that are weak, and the surrounding rock is also weak. The method is used when caving methods cannot be used, either because of the deposit or because of its high value or both. Sets can be extended vertically or horizontally depending upon the shape of the ore body, thus lending flexibility to operations. Basically the sets provide temporary support and a means of reaching the working face. Permanent support is accomplished by means of fill placed in the sets after a tier is worked out. Besides being used in mining heavy ground, square sets are commonly employed in mining pillars between cut and fill stopes. Sets are standardized as far as possible to simplify timber framing.

Among the advantages of square-set stoping are that irregular ore bodies may be worked while masses of waste or lean ore can be left in place; waste may be sorted out and left in the stopes; the grade of ore can be controlled, since each new face can be sampled and assayed before the ore is drilled (this flexibility is important in mines where the ore varies in value); and filling the sets as soon as possible leaves only a small space open at a time. Disadvantages include high cost; slow ore extraction and low productivity per man; large timber requirements and consequent fire hazard; and the highest accident rate of any underground-stoping method.

Disadvantages of square-set stoping

Under many geological conditions in which square set mining is employed, the ore and rock structure approach the behaviour of unconsolidated materials. The instability is caused by faulting, fracturing, and folding and the destructive alteration that accompanies mineralizing processes.

Top slicing and sublevel caving. In top slicing, horizontal slices are made, work proceeding downward from the top of the deposit. The capping or overburden, immediately above the top of the deposit, caves after each slice is removed. During ore excavation, the capping is supported by timber.

Top slicing is most adapted to deposits that are large, horizontal, and weak. It is essential that the capping be such that it will cave. Development openings consist of a series of drifts and crosscut drifts below the mining level and raises to the top of the ore. Sublevel caving is similar, but in sublevel caving, alternate slices are mined, and the weight of the capping is used to assist in mining of the ore. The capping should be somewhat stronger than that for top slicing.

From R.S. Lewis, *Elements of Mining*, 3rd ed. (©1964); used with permission of John Wiley & Sons, Inc.

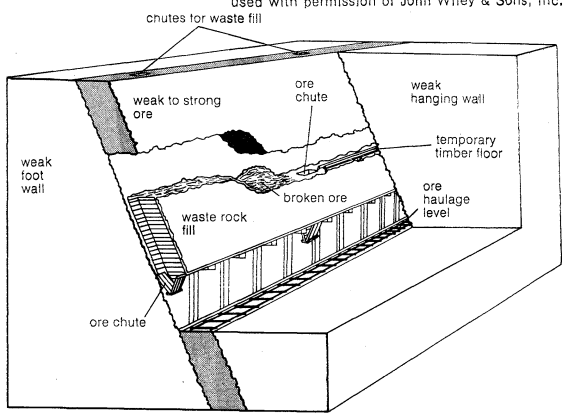


Figure 6: Cut-and-fill stoping, which is adapted to moderately weak ore and wall structures.

For both methods, the rock capping must be weak enough to cave when undermined. The caving formations should not be such that there could be sudden, dangerous collapses of large masses. Sublevel caving is more adaptable to irregular deposits than in top slicing.

Since considerable timber is used for supports in top slicing, an adequate supply of tough, low-cost timber should be available.

Development of top slicing

Top slicing was developed in the United States from caving methods used in iron mines of northern England. In the deposits, soft hematite was covered with glacial drift (loose soil and rock deposited when glaciers retreated). Initially, a longitudinal drift was driven through the ore body and raises containing a chute and a manway were placed at 100 foot (30 metre) intervals. From the top of each raise, a drift was driven a distance of 50 feet (15 metres) in either direction. The drifts were timbered. Slicing took the form of a succession of crosscuts and caves, a process that was repeated until an entire horizontal slice was removed.

Top slicing is safe for mining heavy ground and is economical where timber is plentiful and inexpensive. High extraction of ore is possible with little dilution. If the mine shuts down for any reason, stopes may be blasted down, and the ore body will remain in good condition for resumption of operations. Top slicing can be employed under loose material and does not require as clean a mat (mixture of broken rock and timber) as sublevel caving. On the other hand, top slicing cannot usually be employed without subsidence. Ventilation of such a mine is difficult, and large amounts of old timber create a fire hazard. Furthermore, the period of development prior to production is relatively long, and the rate of production is inflexible. Timbering and covering down (laying a rough timber floor) consume a large part of the work shift; hence, rock drills and scrapers are idle much of the time. Finally, sudden collapse of large blocks presents an element of danger; sorting of waste cannot be done in the stopes, nor can small blocks of waste be left in place.

Sublevel caving may have been developed from top slicing. In this method every other slice is mined by driving crosscuts (slice drifts). The ore between the crosscuts as well as that in the slice above is then mined and the overlying material caved. The method is applicable when the ore is weak and moderately soft and the capping material is hard enough to cave with a minimum of fine material. The method is used in irregular and inclined ore bodies that would not be suitable for working by top slicing. Because the ore must be removed before the capping caves, the capping must cave more slowly than is the case with top slicing.

The interval between levels ranges from 150 to 200 feet (45 to 60 metres). Sublevels are also driven from the tops of raises (vertical mine passages between levels) from the main and transverse haulage drifts. Raises are commonly spaced about 50 feet (15 metres) apart but may be farther apart in narrow ore bodies. Broken ore is pulled to the raises by scrapers. When all the ore is removed from a set and the capping is caved, the next set is caved, and so on, back to the raise at the sublevel.

The advantages of sublevel caving are that cost of mining heavy ground is comparatively low; cost of timber is lower and less development is required than for top slicing; ore may be mined more rapidly; the percentage extraction of ore may be high with little dilution; and the method can be applied to wet and sticky ore not suitable for block caving.

Disadvantages of this method include: there may be more dilution of the ore than in top slicing and square-set stoping; there is little sorting of ore in the stopes; stopes are difficult to ventilate; there is danger of fire from accumulation of timber; and low-grade ore in the capping or near the boundary of the deposits cannot be mined.

Block caving. The block caving method evolved from the methods used to mine pillars in Lake Superior iron mines. It is particularly applicable to very large ore bodies that have cappings that are sufficiently weak to be caved. Preliminary work consists of a series of evenly spaced crosscuts below the level of the ore body, from which a series of raises, known as main, branch, and finger raises, are extended up into the ore. The mass of the ore is then weakened by undercutting, and side cutting is sometimes necessary. The ore is further weakened, usually by blasting, just above the point where the ore is drawn off. The result is that the weight of the ore in addition to the weight of the capping crushes the ore and forces it down through the raises to the haulage way. The movement of the upper mass often affects the strength and stability of the lower mass.

Porphyry–copper deposits where both the ore and capping are weak lend themselves particularly to the block caving method. Large deposits of relatively strong ore are also mined by this method, using appropriate procedures.

Advantages of block caving

The cost of mining by block caving is low, approaching that for open-pit mining, and a high rate of production is attained once production is started. Mining conditions can be standardized, making for safe, efficient operation, and the accident rate is fairly low. On the other hand, capital expense is large, with a long period of time before a mine produces ore. Extracted ore is diluted with waste, and there is some loss of ore; drawing of the ore must be carefully controlled; and, finally, low-grade ore in the capping and at the margin of the ore body is lost.

III. Surface mining

BASIC OPEN-PIT MINING TECHNIQUES

Surface mining, or open-pit mining, is utilized to produce coal, copper, iron ores and nonmetallic mineral products such as clay, gypsum, phosphate rock, several types of stone, sand, and gravel. In recent years in the United States, as much as 88 percent of copper ore and 90 percent of the iron ore produced has been mined by open pit, whereas these values are 40 percent and 50 percent, respectively, for other countries.

Just as in underground mining, mineral deposits mined by surface methods are found by prospecting, evaluated by exploration, then developed, and finally placed in production.

For coal and metallic ores, the ratio of overburden to ore that can be moved is greater than for nonmetallics. Some of the factors that are important in excavation are drilling and blasting, cost of explosives, improvement of efficiency and size of excavation equipment, and methods of haulage.

Prospecting for deposits that may be mined by open-pit methods is conducted in much the same manner as for deeper deposits mined by underground methods. Surface geologic structures are mapped and studied. Drilling, geophysical methods, or both, may be employed to determine details of subsurface geology, to ascertain the presence of valuable minerals, and to make preliminary estimates as a basis for the exploration phase that follows.

When a new mineral deposit is found, it must be explored so that its size, value, shape, overburden, and other related factors can be determined. This information is utilized to ascertain whether it is economical to mine the deposit, to plan the method of operation, to predict profits, and to determine economical limits of mining and the other essential elements. Methods of drilling and sampling are adjusted to the type of deposit, and the results of sampling are evaluated by statistical methods to yield reliable evaluation. It is estimated that only one in 100 mining ventures develops into a successful operation. The exploration of a deposit, if properly done, will yield most of the information necessary to determine whether it may be mined at a profit.

Drilling programs

Drilling programs must be carefully planned; the type of drill chosen depends upon the material to be drilled. In reasonably intact rock and ore, diamond drilling yields more detailed geologic information as well as mineral value. Unconsolidated materials must be drilled by other methods. The first step of a drilling program is further verification of prospecting information; the second is outline drilling, a method of determining the approximate dimensions of the deposit; and the third step is sampling drilling, to yield enough detailed information about the deposit for economic appraisal. Drillings from all three phases are sampled and tested for valuable contents. Drill

holes must be properly spaced and values of samples correctly utilized to evaluate the deposit. Total ore reserves determine the life of the mine.

Factors that must also be considered in the overall operating feasibility of a prospect are geography, land and mineral rights, sociological and political factors, geology, mining conditions, ore-treatment requirements, and economic analysis.

Ownership of mineral land must be proven and stable before a deposit can be explored. Political and sociological factors must be conducive to profitable mining. Geological factors include the occurrence, size, mineral content, depth, and value of a deposit and must be evaluated sufficiently to justify expenditure of capital for development. Mining conditions refer to labour, working conditions, elevation, and related factors. Ore-treatment requirements are critical and include answers to questions of whether the ore must be concentrated before it is smelted or, if nonmetallic, whether treatment will be required before a marketable product can be produced. Finally, an economic evaluation of all of the above must be made, including investment costs, again taking into account that a mineral deposit is a wasting asset.

After a deposit has been explored and its grade, total tonnage, depth of overburden, size, shape, and other factors determined, pit layout must be planned and correlated with all mining operations. In addition to the factors described above, types of ore, cutoff (the lowest grade of ore that is profitable to mine), rate of production, pit slope, bench height (a bench is a level surface worked separately in an excavation), road grades, and related items must be considered. Careful sampling is required to fix the boundaries between capping and ore and to define waste inclusions. The ratio between worthless overburden and valuable ore must be calculated in order to ensure that a profit can be made. Flatbedded deposits are much easier to evaluate than irregular massive deposits. Pit slopes must be gentle so that rock and ore will not slide, and material near geologic fault planes sloping with the pit must be carefully excavated.

The industrial use of mined material determines some mining procedures. Concentration of low grade metallic ores requires that dilution be minimized. Nonmetallics that require purification are in the same category. Samples of the mined material must be representative for proper design of treatment processes. Some mining companies employ computers to calculate pit dimensions and other layout factors.

For stripping overburden and mining, ore equipment is purchased to suit operating conditions. Types of excavators include power shovels, draglines, scrapers, and bucket-wheel excavators. Haulage equipment includes bulldozers, scrapers, trucks, trains, and conveyors of various types. The combination of shovels and trucks is commonly selected.

The total production system usually includes drilling, blasting, loading, and haulage. Modern procedures include analyses of such cycles by advanced statistical techniques that require the use of computers. Successful open-pit mining thus requires modern management techniques and application of good industrial-engineering principles, particularly in large operations. Small or relatively simple operations do not require such complex analysis.

Production drilling is required to make holes for explosive in rock that is too hard to excavate without blasting. The two common methods of rock drilling are mechanical and thermal. Other novel methods, such as sonic, chemical, laser, electron beam, and water jets, are currently in the conceptual or research stage.

Blasting in an open-pit mine

In surface mines, blasting rounds (groups of holes) are spaced according to the size and shape of the benches, rock resistance, joint and fracture systems, and similar factors. Blasting is conducted to give the best breakage with a minimum displacement of the broken material. Large boulders require costly secondary breakage, and a pile of broken rock close to the new bench face is less expensive to load. If blasting is near populated areas, ground vibrations and air blast must be minimized.

Blasting with nuclear explosives is being considered for several purposes but mostly for large scale underground mining or for excavation of canals and harbours. Proper burial of the explosive device minimizes atmospheric radiation and the unit explosives costs decrease with increase of the size of the charge for nuclear explosives.

Pit operation is performed in cycles of drilling, loading explosives, blasting, and loading and haulage. In addition to power shovels, draglines and clamshells are employed in many countries, as are continuous excavators, consisting of a bucket wheel on the end of a long boom, which excavates loose material and dumps it in a conveyor system, which in turn conveys ore to haulage units and waste to a spoil bank.

Large mines are organized and equipped for maintenance of all types of mining equipment, while, in small mining operations, maintenance may be performed by operating crews.

Removal of water and flood control are important in many mines, depending upon rainfall and the presence of water-bearing rock. Slope stability in large mines may present problems. Research in rock mechanics and geology is often essential in preventing slides and predicting rock mass movement. In some mines, seismic Geophones have been installed in the rock at critical points to detect minute failures as evidenced by microseisms (tiny earth tremors). Their intensity and magnitude are indications of rock failure and possible slides.

Some valuable minerals, such as gold and cassiterite (tin mineral), are found in gravel deposits, where they have been deposited after being eroded from their place in primary deposits. Deposits of this type (placer deposits) are often exploited by dredges or hydraulic mining, as is overburden over iron-ore deposits in some countries.

Dredging

A dredge is a floating barge with a boom at one or both ends and engines and a concentrating plant if the latter is required in the barge or hull proper. A bucket-line dredge has one boom with a bucket line that may be lowered into the water in which the dredge floats. The dredge is pivoted about the far end on a spud (a heavy, pointed column driven into the floor of the water body), the bucket line continuously excavating material as it is swung by means of tension lines. The excavated material is conveyed through the concentrating plant if it contains valuable mineral; waste is conveyed up the rear boom to the top of the spoil bank. Dredges for moving waste material have no concentration plants on board. They are operated either in a pond in a gravel deposit (California, Malaysia) or offshore (Indonesia and the Philippines). Mined materials are screened and the valuable minerals recovered by washing the finer gravel over riffles, grooved sluices, where the heavy minerals settle out by gravity.

Hydraulic mining consists of washing the material from its position by a large stream of water under high pressure from a large nozzle (known as a hydraulic giant). It is used for mining terrace placer deposits that cannot be mined by dredging or other means. A large supply of water is necessary. Finer materials are transported to a sluice where valuable (heavy) minerals are recovered in riffles. Debris from the operation is accumulated downstream, preferably in ponds behind dams that will not be washed away by floods.

Recently, in the alluvial tin mines in Malaysia, mining practice has evolved almost entirely to an adoption of a system of breaking ground with water jets and removing the loosened material with gravel (suction) pumps.

Offshore operations include mining of gold in Alaska, tin-bearing sands off the coast of Thailand, tin off the shores of the Indonesia island of Sumatra, iron sands off the southwestern tip of Kyushu Island (Japan), cement sands (shell) west of Akranes (Iceland), and gravel from the English Channel.

QUARRYING

One meaning of the word quarry is "squared stone," hence, quarrying was initially applied to the mining of dimension (cut) stone. More recently, quarrying has been used to refer to the mining of several types of nonmetallic rock from shallow open-pit mines.

All three types of rock are quarried—igneous, metamorphic, and sedimentary. The most common igneous rock is granite; several sedimentary rocks are quarried, and marble, quartzite, and slate are common metamorphic types. The two products of quarrying are dimension stone and crushed stone. The former is employed for buildings, monuments, and decorative purposes. Such rock usually must be of uniform grain size, possess desirable coloration, and have lasting qualities related to its intended use. Sandstone is used for building, granite and marble for memorials, slate for roofing, and quartzite for flagging.

When usable rock is found, the first part of mining consists of stripping the overburden and weathered rock. Large blocks are then separated from the parent mass, sometimes with low-powered explosives. High-energy explosives shatter the rock and cannot be used.

Quarrying dimension stone

For most dimension-stone quarrying, the success of the art depends largely upon taking advantage of joints and cleavage planes in the rock. Granite has a "rift" along which it may be split with comparative ease. The planes perpendicular to the rift are the "head" and the "grain," along which it splits less easily. The rift of sedimentary rock is the bedding plane (*i.e.*, parallel to its layering), while some rocks, such as diabase, diorite, and porphyry (trap rock), have no rift. Joints in granite may be at obtuse angles, but, where they are perpendicular, the excavation is termed a block quarry.

A usual method of splitting granite by hand is to drill a row of holes about three-quarters of an inch (2 centimetres) in diameter, 3 to 5 inches (8 to 13 centimetres) deep and 6 to 8 inches (15 to 20 centimetres) apart. In each hole are placed two "feathers" and a plug. The feathers are half-round shims or wedges between which is placed a wedge plug. The plugs are successively driven with light blows of a hammer, creating a line of tensile stress between the holes that causes the rock to split. Granite and some limestone split with shallow holes, while sandstone and marble require deeper holes. Holes may be drilled with hand tools or with pneumatic drills. In softer rocks the holes may be farther apart.

Special quarrying methods

Special methods of quarrying include machine channelling, broaching, the Knox system, and flame-jet channelling. Machine channel bits are made of hard steel mounted to cut a channel in softer rock (that is, other than granite or hard metamorphic rocks). Broaching consists of drilling a row of holes and cutting the rock out between them. In the Knox system, deeper holes are drilled, grooved on opposite sides, and then loaded with black powder, which is fired to break the rock. For some types of rock containing quartz, flame-jet channelling has been employed. The rock must have the proper spallation (thermal breakage) characteristics for this method to be applied. Wire saws are also utilized; these employ a wire that runs as a belt over the rock, cutting by abrasion as sand or hard grains and water are fed onto it. Blocks are cut to the size desired, hoisted by derricks or other devices and transported away from the quarry for further cutting or sale.

Crushed stone is mined by the methods used in similarly shaped deposits of metallic ores; that is, the overburden is removed, then the rock is drilled and blasted. The broken material is loaded and hauled to a crusher, where it is crushed, screened, and sized. It is used for such purposes as road building, railroad ballast, and concrete aggregate. Some limestones have special uses, such as chemical applications, making portland cement, and for flux in blast furnaces.

IV. The mining industry

ORGANIZATION AND MANAGEMENT

The organizational structure of mining companies varies, but similar types are found in coal mining, open-pit mining, and underground mining. A large company usually has a corporate structure with departments or divisions for operations, engineering, sales, exploration, land, purchasing, financial, and legal aspects. All of these are administered by a centralized executive office headed by the president of the board and the president of the company.

A typical organization at a mine consists of a superintendent, an assistant superintendent, master mechanics, a chief electrician, pit or mine foremen, shift bosses, blasting foremen, engineers, clerical staff, and other assistants.

Strong support of research in various phases of mining has proved to be important to progressive management and successful operation of most mines. Most large mining companies throughout the world have research departments whose responsibility is to perform investigations that will solve problems in health and safety, rock mechanics, operations efficiency, and related matters.

CONSERVATION, ECOLOGY, AND MINING

The conservation of mineral resources, reduction and prevention of environmental pollution, and the preservation of wilderness areas are interrelated in varying degrees. Heavily populated countries are particularly concerned about the problems involved.

Inasmuch as mineral deposits are wasting assets that cannot be replaced, it behooves every nation to conserve them and use them wisely. An available quantity of minerals and mineral products is a vital factor in maintaining high national economic standards, in political power among the nations, and in the ability of a nation to develop its self-defense. Iron, coal, steel, copper, aluminum, and steel-alloying metals are basic necessities, with others of slightly less importance following. During wartime, many of the ores yielding these metals are classified as strategic minerals.

In the United States, many mining and smelting operations have contributed in the past to pollution of streams and the atmosphere. Legislation, both national and state, has been passed to improve these conditions, and many mining companies have taken steps to correct pollution problems. The Bureau of Mines is supporting research and development efforts to assist the mineral industries in the reduction of pollution due to mining and processing of minerals.

Effects of conservation regulations

Because of the nature of financial risks involved in the exploration and development of mineral deposits, many mining companies have found that regulations increase mineral-production costs; for example, regulations may prohibit temporary or permanent roads, aircraft landing strips or heliports, or the use of motorboats, motor vehicles, or other motorized equipment. An aircraft may not be allowed to drop equipment, supplies, or personnel. A permit is required for all types of prospecting except walking over the land. The validity of an unpatented claim may be contested, even though the valuable mineral may be far below the surface. Permits may be refused on the ground that the use of equipment is not consistent with desires to preserve various wilderness areas. Hearings may be called by interested groups to prevent the issuance of a permit, and repeated reconsideration of decisions may hold up prospecting indefinitely. (For further discussion of the ecological aspects of mining, see CONSERVATION OF NATURAL RESOURCES.)

BIBLIOGRAPHY. AGRICOLA, *De re metallica* (1556; Eng. trans. by H.C. and L.H. HOOVER, 1950), is a classic work on early mining in Europe. ROBERT PEELE and JOHN A. CHURCH (eds.), *Mining Engineers' Handbook*, 3rd ed., 2 vol. (1941), is a standard American work on basic surface and underground mining to 1941. SETH D. WOODRUFF, *Methods of Working Coal and Metal Mines*, 3 vol. (1966), gives a recent treatment. ROBERT S. LEWIS and GEORGE B. CLARK, *Elements of Mining*, 3rd ed. rev. (1964), is a general text that treats both theory and practice. GEORGE B. CLARK (ed.), *Symposium on Mining Research*, first to sixth (1956–61), deal with mining research, explosives, and blasting.

Recent works on surface mining include: EUGENE P. PFLEIDER (ed.), *Surface Mining* (1968); INSTITUTION OF MINING AND METALLURGY, *Symposium on Opencast Mining, Quarrying, and Alluvial Mining* (1965); and *Pit and Quarry Handbook* and *Purchasing Guide* (annual). MELVIN A. COOK, *The Science of High Explosives* (1958), is a comprehensive treatment of theory and experimentation. E.I. DUPONT DE NEMOURS AND COMPANY, *Blasters' Handbook*, 15th ed. (1966), deals with applied blasting techniques. JAMES TAYLOR and P.F. GAY, *British Coal Mining Explosives* (1958), discusses permissible explosives. U. LANGEFORS and B. KIHISTROM, *The Modern Technique of Rock Blasting*, 2nd ed. (1968), gives

an analysis of blasting theory and practice. KUMAO HINO, *Theory and Practice of Blasting* (1959), treats applications of experimental results. Two of the most recent books on mine ventilation are A. ROBERTS (ed.), *Mine Ventilation* (1960); and HOWARD L. HARTMAN, *Mine Ventilation and Air Conditioning* (1961).

Valuable journals and reports on mining theory, policy, research, and practice include: *Mining Magazine, Mining Journal, Mining Engineering, Canadian Mining Journal, Institution of Mining and Metallurgy Bulletin and Transactions, Mining Congress Journal,* Bureau of Mines publications (*Bulletins, Information Circulars,* and *Reports of Investigations*), *Annales des Mines, South African Mining and Engineering Journal, Transactions* of the *American Institute of Mining and Metallurgical Engineers,* and the *Engineering and Mining Journal.*

(G.B.Cl.)

Minnesota

Near the heart of the North American continent, Minnesota lies at the headwaters of the Mississippi River, which flows southward to the Gulf of Mexico; of the Great Lakes, which drain eastward to the Atlantic Ocean; and of the Red and Rainy rivers, which flow northward into Hudson Bay. The state's 84,068 square miles (217,736 square kilometres) are bounded on the east by Lake Superior and Wisconsin, on the south by Iowa, and on the west by North and South Dakota. The northern boundary with the Canadian provinces of Manitoba and Ontario includes a small hump that makes Minnesota the most northerly of the 48 coterminous U.S. states. The state is dominated by a single large metropolitan area, the Twin Cities of Minneapolis and St. Paul (the capital) and their suburbs, in which nearly one-half of the 3,800,000 Minnesotans live.

In the popular image, Minnesota is a land of rugged winters, innumerable lakes, Scandinavians, flour mills, and iron mines. The variation from cold winters to warm summers and the nearly 5,000 square miles of inland water—more than that of any other coterminous state except Texas—do, indeed, serve as pacemakers for the style of life in Minnesota. About one in four Minnesotans has some Scandinavian blood, but Germans make up the largest ethnic stock. Flour milling, symbolic of the state's economic dominion over the vast grain belt to the west and northwest, reached its peak in 1920, when Minneapolis was the world's leading producer of flour; but since then it has declined into a relatively unimportant role in the economy. At one time Minnesota's Mesabi Range produced more than one-half of all the iron ore in the world, but the high-grade ores in the range have been almost completely depleted. (For information on related topics, see UNITED STATES; UNITED STATES, HISTORY OF THE; and NORTH AMERICA.)

THE HISTORY OF MINNESOTA

Diversity of Indian tribes

Until the mid-19th century, two major Indian tribes occupied what is now Minnesota: the Chippewa in the north and east, the Sioux in the south and west. Between the time of European exploration and statehood, the Chippewa had pushed the Sioux southward and southwestward onto the prairieland, with the Chippewa occupying the forested areas of the state. Indians of tribes from as far away as the Appalachians and the Rocky Mountains met in a sacred place of peace in southwestern Minnesota to quarry a hard red rock used for making peace pipes; this area is preserved as the Pipestone National Monument.

European settlement. The Kensington stone, found in west central Minnesota in 1898, bears inscriptions allegedly made by Norsemen who penetrated the continent in the 11th century, but its authenticity remains in doubt. The earliest verifiable Europeans in the area were the 17th-century French explorers searching for a Northwest Passage. The first white settlement was made where the French fur traders known as voyageurs had to leave Lake Superior to make a nine-mile portage (land detour) around the falls and rapids of the Pigeon River, the longest portage in their transcontinental water route.

The first permanent U.S. settlement was at Ft. Snelling, a military outpost established in 1819 overlooking the Mississippi and Minnesota rivers; the site has been restored as a state park. Immigration into the region was slow during the next 30 years, but, once the value of the vast forestlands of northern and central Minnesota was realized, lumbermen from New England, seeking to exploit the pine, balsam, and spruce forests that covered nearly 40 percent of the land, led a large wave of permanent settlers. An additional quarter of the state was covered with noncommercial hardwood.

Territory and state. Minnesota became a territory in 1849, its boundaries reaching as far west as the Upper Missouri River, but most of its 4,000 white settlers were located in the Fort Snelling–St. Paul area in the east. Lumbering grew rapidly, and major sawmills were built at Stillwater, on the St. Croix River, and at the Falls of St. Anthony, on the Mississippi, the major source of waterpower in central North America. The two villages at the falls merged in 1872, the village of St. Anthony on the east bank being absorbed into the larger and more aggressive city of Minneapolis on the west bank.

Minnesota became a state in 1858, its boundaries being cut back from the Missouri to the Red River, the present western boundary. The most rapid period of settlement in Minnesota was during the 1880s, when homesteaders—many recruited from Germany, Sweden, and Norway—rushed into western and southwestern Minnesota. In the same period, lumbering was at its peak and flour milling, using the power at St. Anthony falls, was becoming important. Both Minneapolis, as the lumber, milling, and retail centre, and neighbouring St. Paul, as the transportation, wholesaling, finance, and government centre, tripled in population during the 1880s. The rivalry between these two cities became particularly intense after the census of 1880, when Minneapolis surpassed St. Paul in population.

Tapping the ore beds

Commercial iron-ore production began in Minnesota in 1884 at Soudan, on the Vermillion Range. After the huge iron reserves of the Mesabi Range were discovered at Mountain Iron in 1890, large-scale production began, and the population along the Mesabi Range and in the port cities of Duluth and Superior grew rapidly in the next two decades. Most of the valuable pine, balsam, and spruce in central and northeastern Minnesota was cut before 1900, after which time lumbering declined rapidly in importance. Wood products, however, remained important in northern and northeastern Minnesota.

The 20th century. The mechanization of agriculture and changes in transportation technology had significant effect in Minnesota. The former allowed greater productivity for farm workers, resulting in increased farm size and a decline in farm population. New modes of transportation improved the accessibility of all parts of the state, thus accelerating urbanization and extending the commutation zones for many miles around the growing urban areas. These two factors have had a particularly important impact on the density and location of population and the economic activity in the state in the 20th century. The impact was apparent after World War I, with the great popularity of the automobile, but was repressed during the Depression years and World War I. Since World War II the impact of these factors has been particularly dramatic; urban areas have been growing while rural areas have been thinning out.

THE NATURAL AND HUMAN LANDSCAPE

Surface features. Although it straddles three continental watersheds, Minnesota's elevations range only from 600 feet above sea level, at Lake Superior, to 2,300 feet, at Eagle Mountain. Most of Minnesota was at one time covered with glaciers, and the surface of the land today has been shaped by the movement and the melting of those glaciers. Among the major remnants are the many thousands of lakes, 10,809 of which are more than 25 acres in area. The greatest glacial moraine development in Minnesota stretches from Albert Lea, in the south, northward to the Twin Cities; then westward and northward through Willmar, Glenwood, Alexandria, and Detroit Lakes; eastward to Park Rapids and Leech Lake; and, finally, south to Mille Lacs Lake. The greatest of the lake plains is that of Lake Agassiz, which existed in gla-

cial times along the Minnesota–North Dakota border and is now known as the Red River Valley. In northeastern Minnesota there are many areas of bare rock, stream valleys, and deep, clear lakes that were scoured by glaciers from the landscape. Extreme southeastern Minnesota was the only unglaciated part of the state. There, streams have cut their way through layers of limestone over many thousands of years, and steep, rocky bluffs rise 500 feet or more above the valleys.

Climate. The surprisingly great variations of temperature and growing season that occur not only seasonally but also from one part of the state to another reflect the fact that the state stretches from the edge of the sub-Arctic forest to the heart of the Corn Belt. The northeastern part of the state lies in the moist northern Great Lakes storm belt, while the western border lies at the edge of the semi-arid Great Plains.

In July, average daily maximum temperatures range from 85° F (29° C) in southern Minnesota to 80° F (27° C) in the northwest to 70° F (21° C) along the shore of Lake Superior; minimum in the same areas are, respectively, 60° F (16° C), 55° F (13° C), and 50° F (10° C). Average daily January highs range from 25° F (−4° C) in the south to 15° F (−9° C) in the north, minimums from 5° F (−15° C) to −5° F (−21° C). The average frost-free periods range from less than 90 days in parts of the north to more than 160 days in parts of the south. The average annual precipitation ranges from more than 30 inches (750 millimetres) in the southeast and the extreme northeast to less than 20 inches (500 millimetres) in the northwest. Average seasonal snowfall ranges from less than 40 inches (1,000 millimetres) in the western part of the state to more than 70 inches (1,800 millimetres) in the extreme northeastern tip. Most parts of Minnesota have continuous snow cover for at least 90 days, from about mid-December to mid-March.

Vegetation and soils. Minnesota stands astride one of the major physical geographic boundaries in the world, the sharp transition from forest to prairie in the heart of North America. The natural vegetation of Minnesota may be divided into three general categories: needle-leaf forests, hardwood forests, and tallgrass prairie. The needle-leaf forests originally occupied the northeastern third of the state and included pine, spruce, and fir, with tamarack in the bog areas. A belt of hardwoods extends from southeastern Minnesota to the Canadian border, passing through the Twin Cities and lying immediately to the south and west of the coniferous forest. Known as the big woods, this hardwood forest averages some 40 to 80 miles in width and consists primarily of oak, maple, and basswood. South and west of the hardwood forests lies the tallgrass prairie.

Belts of forest and prairie

The most fertile soils in the state are found in the grassland regions of the south and west, where they are rich in organic matter and high in soluble mineral plant food. Soils formed beneath the coniferous forest in northeastern Minnesota are light coloured, acidic, and low in organic matter, while those in the hardwood areas are intermediate in colour and natural fertility.

Human imprints. During a century of white occupancy, virtually all of the prairies have been cultivated. The coniferous forestlands, mostly cut over by 1920, are covered again by aspen, birch, and jack pine, while much of the big woods has been cleared for crops and pasture. Minnesota reached its peak in cultivated farmland in 1945. Since that time, the agricultural frontier has retreated, and farms have been abandoned in the less fertile areas in north central and northeastern Minnesota. The big woods area is now primarily a dairying area. Within 100 miles or so of the Twin Cities, dairying has intensified in recent years, but beyond that—and particularly in west central and northwestern Minnesota—dairying has declined in importance. The prairie areas of southern and southwestern Minnesota have a characteristically Corn Belt crop and livestock agriculture.

THE PEOPLE OF MINNESOTA

Ethnic and religious groups. The New England Yankees who first settled Minnesota were businessmen and lumbermen who helped establish the institutions and many of the traditions that remain important in Minnesota today. The first major immigrant groups in the latter half of the 19th century were German, Swedish, and Norwegian; they were primarily farmers and tradesmen.

The German settlers dominated the push up the Mississippi, continuing into the central and south central parts of the state. Norwegian settlers moved westward across the southern tier of counties, forming the major ethnic group in west central Minnesota and the Red River Valley. The major area of Swedish settlement is in five or six counties immediately north of the Twin Cities area. Substantial numbers of Finns live in northeastern Minnesota, Poles in central Minnesota, Bohemians south of the Twin Cities area, French and French-Canadians just north of the Twin Cities and in northwestern Minnesota, Dutch and Flemish in parts of southwestern Minnesota, Icelanders in northwestern Minnesota, and Danes and Swiss in scattered pockets.

The ethnic blend

These ethnic groups influenced the distribution of religious denominations throughout the state. Central and south central Minnesota are heavily Roman Catholic, reflecting the German, Polish, and Bohemian populations. A strip across southern Minnesota and through western Minnesota is heavily Evangelical Lutheran, reflecting the Norwegian population. Swedish communities are generally Augustana Lutheran. Many of these ethnic clusters have retained a high degree of purity in the rural areas. They have been sources of population movement to urban areas, but they have attracted few in-migrants to alter the original stock.

The larger and growing urban areas—the Twin Cities area, in particular—have been built from many different ethnic groups and have become major areas of cultural contact and variety. These urban centres also attracted populations not found in significant numbers in the rural areas: Ukrainians, Irish, Italians, Greeks, Jews, blacks, and Spanish-Americans. The Indian population, numbering about 23,000 in 1970, was primarily Chippewa; of that number, nearly one-half lived in urban Twin Cities neighbourhoods. Less than 1 percent of the population was black, over 90 percent of whom lived in the Twin Cities area.

Demographic trends. Although Minnesota has experienced a net out-migration during the middle decades of the 20th century, its population has grown through natural increase—the preponderance of births over deaths—and it has experienced population movements similar to those of the nation as a whole. The percentage of the population living in the Twin Cities metropolitan area is increasing because of natural increase and in-migration from the rural areas of surrounding states. In 1970, the Twin Cities area was the nation's 15th largest metropolitan area; only two larger metropolitan areas exceeded it in rate of growth over the 1960s. Most urban centres in the state with a population of more than 5,000 share in this rural-to-urban migration, while most towns of fewer than 1,000 inhabitants struggle for their existence. The lake region and other scenic areas of the state, meanwhile, are experiencing an increasing number of permanent year-round residents.

THE STATE'S ECONOMY

Basic resources. The economic growth of early Minnesota was related closely to the exploitation of the primary natural resources: agriculture, minerals, and timber. The activities, in turn, stimulated the growth of such ancillary activities as railroads, processing of natural-resource products, and services. Today, in Minnesota as throughout the nation, employment in the natural-resource-based activities, together with the railroads and associated manufacturing sectors, generally is declining. Nevertheless, about one in 15 employed persons in Minnesota is engaged in agriculture—more than twice the national average—and railroads account for about one in 75 employed persons. As recently as 1951, Minnesota produced 82 percent of the nation's iron ore, and, as the high-grade ores have been depleted, the seemingly unlimited reserves of low-grade ore, or taconite, have been

developed. Minnesota now supplies some 40 to 50 percent of the nation's iron-ore needs, mostly with low-grade processed ores. Other mining activities include sand and gravel, granite, and limestone quarrying.

Components and trends. Fewer than one in four employed persons in Minnesota is engaged in manufacturing (a figure slightly below the national average), but the rate of manufacturing growth exceeds that of the nation. The Twin Cities area dominates the state in manufacturing and accounts for over one-half of the state's employment, but in recent years growth has been greater outside this area. The service and government sectors have grown rapidly, with a modest growth occurring in most manufacturing not tied to natural resources. The net effect of these changes is an employment growth in Minnesota slightly below the national rate.

Minnesota is a net exporter of agricultural commodities, iron ores, and manufactured goods tied to natural resources, such as meat, dairy, and grain products, pulp and paper, and primary metals; it is a net importer of most manufactured goods not tied to natural resources —with such exceptions as electrical machinery and instruments. No coal, oil, or natural gas is produced in Minnesota, and the geologic formations are such that the discovery of such minerals is highly unlikely.

Multistate domination by Twin Cities

Probably no region in the country is so dominated by one urban centre as the upper Midwest by the Twin Cities in wholesaling, finance, and cultural activities. This multistate region includes western Wisconsin, Minnesota, northern Iowa, North and South Dakota, and eastern Montana. No urban area in this region has as much as one-tenth the population of the Twin Cities area. The continued rapid growth of the Twin Cities area has been partly a result of this regional dominance and partly a result of the diversified economic base that contributes to its own economic stability. At the same time, the Twin Cities area in many ways is becoming more closely tied economically to other major urban centres in the nation and the world and less directly dependent upon processing the products of the region and servicing the region.

Transportation. The movement of people and goods in Minnesota and the upper Midwest focusses on the Twin Cities area. Regional and transcontinental rail and highway systems radiate outward from the Twin Cities, tying the larger towns, smaller towns, and hamlets into one interdependent network. Independent of this major transportation network is the rail system of northeastern Minnesota, which brings iron ore and taconite products for transshipment by boat at the Lake Superior ports of Duluth and Superior. Wheat from the Dakotas and Montana also has been an important product transshipped from rail to boat at Duluth. Since the opening of the Great Lakes waterway to ocean vessels in 1959, products of the upper Midwest are carried directly to locations throughout the world.

River transportation was the first important mode for the movement of both people and goods in many parts of the state. Barges on the Mississippi now carry bulk products to and from the major inland ports at St. Paul and Minneapolis. Coming upstream are such bulk products as coal, oil, and salt; transported downstream are grain, sand, and gravel.

The Twin Cities area, also the air hub of the upper Midwest, is served by several scheduled commercial airlines. In 1970, more than 5,000,000 commercial passengers used the Minneapolis–St. Paul International Airport, which is supplemented by a satellite network of eight additional airports. Fourteen other communities in Minnesota are served by scheduled air service.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. *State level.* Minnesota's constitution, dating from the attainment of statehood in 1858, is lengthier than that of many states. The executive branch of government consists of the governor, a lieutenant governor, a secretary of state, an auditor, a treasurer, and an attorney general, all of whom are elected independent of one another for four-year terms. The heads of the 35 administrative agencies are appointed by the governor.

Executive, legislature, and judiciary

The Minnesota legislature meets in odd-numbered years for 120 legislative days. The House of Representatives comprises 135 members elected for two-year terms; the Senate, 67 members elected for four-year terms. Members of the legislature are elected without party designation but are organized into liberal and conservative caucuses.

The judicial branch comprises the Supreme Court, district court, probate court, municipal courts, and justices of the peace. Seven justices—a chief justice and six associates—constitute the Supreme Court; each is elected for a six-year term.

Local government. The state's 87 counties range in size from 155 square miles of land area to 6,092 and in population from about 3,400 to 960,000 in 1970. The 1967 Census of Governments reported 850 municipalities and 1,817 townships. Counties and municipalities provide most of the local governmental services, while the responsibilities of townships are limited to maintenance of local roads and certain welfare services. The number of school districts has been declining steadily through consolidation, especially in rural areas. Special districts, much fewer in number than in many states, provide for sewers, conservation, water supply, fire protection, parks, airports, and mosquito control.

Revenues and regional cooperation. One of the most difficult problems facing Minnesota is meeting the increasing cost of government and increasing the sources of revenue that support government. Traditionally, Minnesotans have been willing to tax themselves at a high rate to obtain needed public services, but inequities have existed in the abilities of local governments to provide those services. Several important legislative acts have been based on an equalization concept. In 1967 a bill establishing a metropolitan council in the Twin Cities metropolitan area was passed, and in 1969 the concept of regional cooperation was extended to ten other multicounty areas of the state. Gradually these other regions are becoming established on a voluntary basis. In 1971 a revenue-sharing bill required that 40 percent of the property-tax revenue from newly located industries within the seven-county Twin Cities metropolitan area be returned to the area and redistributed to municipalities and school districts. Another important tax bill passed in 1971 stipulated that state funds rather than local property taxes were to be used to pay most costs of operating local school districts.

The Twin Cities Metropolitan Council, the 15 members of which are appointed by the governor, is responsible for the development of certain area-wide services that local government is unable to provide, including sewage and water systems, transportation, and major land uses. In so doing, it plays a coordinating and regulatory role among the more than 300 local governmental jurisdictions within the Twin Cities area.

The social and political milieu. Among Minnesota's most significant characteristics is its political environment and high degree of citizen involvement in public affairs. These traditions were laid in Minnesota by the early settlers from New England, who brought their town-meeting form of government. This foundation was reinforced by the Scandinavian and German immigrants, with their ambition and high regard for education. As a result, government has always been accepted as the legitimate means for public decision-making in Minnesota. From those early years, business has played an important role as a strong participant in public decisions.

Citizen involvement in politics

The traditions of citizen involvement can be seen in the many neighbourhood and community organizations and ad hoc issue-related groups in the state and in the relatively large number of Minnesotans in national political prominence. Minnesota has been a leader in such national movements as those to guarantee the rights of women, homosexuals, and American Indians. The average Minnesotan is usually well informed about public issues. As a result, the political parties usually have focussed their attention on the visible public issues, thus bringing about a balanced political system in which neither major party has been dominant over an extended period.

Higher education. The University of Minnesota was established in 1851, with its main campus in Minneapolis.

Together with the smaller St. Paul campus, the Twin Cities campus has the largest daytime enrollment of full-time students on any campus in the country—over 50,000 in the early 1970s. Smaller branches of the university are located in Duluth and Morris, with four-year programs, and in Crookston and Waseca, with two-year programs.

To the six four-year Minnesota state colleges in smaller cities across the state, the legislature in 1971 added a seventh in the Twin Cities metropolitan area. This "college without a campus" utilizes diverse physical facilities throughout the area, bringing higher education to the neighbourhoods. In addition, the state operates large systems of junior colleges and of post-secondary vocational–technical schools. Fourteen private four-year colleges and four private junior colleges supplement the public system.

Health and welfare. The Twin Cities and Rochester serve as national health-care centres. The Mayo Clinic in Rochester, famous for its comprehensive approach to health care, has served patients from around the world since the late 19th century. The University of Minnesota Hospitals complex in the Twin Cities area has been a pioneer in medical research, while numerous hospitals across the state provide an effective network of medical care.

According to a study by the Midwest Research Institute, Minnesota ranks first among the U.S. states in the quality of its health and welfare. The report pointed out that a high standard of general medical services is provided in Minnesota, that children's health and welfare programs are among the best in the nation, and that Minnesota leads all states in old-age-assistance programs.

CULTURAL LIFE AND INSTITUTIONS

Recreation. Cultural life in Minnesota is highly diversified and seasonal. Many activities are outdoors oriented: swimming, boating, canoeing, camping, hunting, fishing, and winter sports. More fishing licenses are sold in Minnesota than in any other state, and Minnesota has more boats per capita than any other state. A high proportion of Minnesotans own second homes on one of the state's many lakes and streams, which are easily accessible to the population centres of the state. Voyageurs National Park, authorized by Congress in 1971, is being developed in the water wilderness along the Canadian border.

Prevalence of outdoor life

The arts. Whereas "outstate" Minnesota—beyond the Twin Cities area—serves as the outdoor playground for the state, the Twin Cities area serves as the centre of institutional culture. The best known musical organization is the Minnesota Orchestra, which was formed in 1903 as the Minneapolis Symphony Orchestra and acclaimed as among the finest symphony orchestras in the nation. Other important musical organizations include the Schubert Club, the Civic Orchestra of Minneapolis, the St. Paul Chamber Orchestra, the Center Opera Company, and the Duluth Symphony. Colleges and universities throughout the state make substantial contributions to the arts within their communities and regions.

In the early 1970s the Twin Cities had nearly a dozen resident professional theatres. The best known of these was the Tyrone Guthrie Theatre Company, formed in 1963. The Childrens Theatre Company is nationally recognized as one of the finest of its kind. The Minnesota Dance Theatre and the Northern Theatre Ballet Company are resident dance companies in the Twin Cities.

The Minneapolis Institute of Arts, the Walker Art Center, and the Minnesota Museum of Art are among the most important art museums in the state. Other major museums that serve the Twin Cities area include the St. Paul Arts and Science Center, the Bell Museum of Natural History, the Minnesota Historical Society, the American Swedish Institute, the Natural History Museum and Planetarium of the Minneapolis Public Library, the University of Minnesota Landscape Arboretum, and the Como Zoo and Conservatory.

The communications media. In the early 1970s, about 30 daily newspapers were published in Minnesota, while four other dailies published in urban centres in adjoining states had significant circulations in Minnesota; weekly or semi-weekly papers numbered some 350. The four Twin Cities dailies, particularly the Minneapolis newspapers, dominate circulation throughout Minnesota, western Wisconsin, northern Iowa, and most of North and South Dakota. Fifteen commercial television stations and three educational stations serve the state, and part of the state is served by nine commercial television stations and one educational television station in adjoining states. There are about 150 radio stations within the state, including nearly 60 FM broadcasters.

PROBLEMS AND PROSPECTS

In Minnesota's gradual shift away from a natural-resource-based economy, farms are growing larger, the farm population is declining, and the rural nonfarm population is increasing, particularly in the scenic lake areas of the state. Small towns struggle for their existence, while the Twin Cities metropolitan area is spreading and extending its influence into the rural areas of the state. The cores of Minneapolis and St. Paul are growing older and share most of the typical problems of urban areas across the country.

These changes across the state raise a number of questions that must be confronted during the 1970s: the density and distribution of population and economic activity; and the equitable provision for and financing of public needs in the areas of health care, education, transportation, communications, housing, and environmental management. The outcome of the legislative steps toward regional cooperation in these areas will be critical for Minnesota and may provide insights into issues of importance to the nation as a whole.

BIBLIOGRAPHY. THEODORE C. BLEGEN, *Minnesota: A History of the State* (1963), an authoritative history; JOHN R. BORCHERT and DONALD D. CARROLL, *Minnesota Settlement and Land Use, 1985* (1971), current geographic patterns of population and major land use, with projections; JOHN R. BORCHERT and DONALD P. YEAGER, *Atlas of Minnesota Resources and Settlement*, rev. ed. (1969), maps of social, economic, and demographic conditions; KENNETH CARLEY, *The Sioux Uprising of 1862* (1961), a readable, accurate account with illustrations and maps; HELEN B. CLAPESATTLE, *The Doctors Mayo*, 2nd ed. (1954), a history of the Mayo Clinic; HIRAM M. DRACHE, *The Challenge of the Prairie: Life and Times of Red River Pioneers* (1970), a look at family, social, religious, and economic life in the Red River Valley area; JOHN FISCHER, "The Minnesota Experiment: How To Make a Big City Fit To Live In," *Harper's*, 238:12–32 (1969), on the political environment that led to the Twin Cities Metropolitan Council; WILLIAM W. FOLWELL, *A History of Minnesota*, rev. ed. 4 vols. (1956–69), the most authoritative, comprehensive history; MERRILL E. JARCHOW, *The Earth Brought Forth* (1949), an authoritative account of the pioneers of Minnesota agriculture; LUCILE M. KANE, *The Waterfall That Built a City: The Falls of St. Anthony in Minneapolis* (1966), a lively history of Minneapolis; MINNESOTA, SECRETARY OF STATE, *The Legislative Manual of the State of Minnesota* (biennial), a reference book on state government; G. THEODORE MITAU, *Politics in Minnesota*, 2nd ed. (1970), an insightful discussion of parties, issues, leaders, the legislative process, and the political environment; GEORGE M. SCHWARTZ and G.A. THIEL, *Minnesota's Rocks and Waters*, rev. ed. (1963), a popular description and interpretation of Minnesota landscape, with suggested field trips.

(N.C.G.)

Minorities and Ethnic Groups

The phenomenon of culturally or racially distinct minority groups living in the midst of an alien society is a familiar one, but what exactly is meant either by "minority" or by "ethnic group" is something that scholars, politicians, and those who have to take practical action in social matters have seldom found easy to describe.

The nature of minorities. The most common general description of a minority group used is of an aggregate of people who are distinct in race, religion, language, or nationality from other members of the society in which they live and who think of themselves and are thought of by others, as being separate and distinct. Separation, too, often implies that the members of such a minority are excluded from taking a full share in the life of the society

because they differ in certain ways from the dominant group—a situation that tends to develop attitudes of discrimination and prejudice toward the minority in question, attitudes that may also be assumed by other groups or minorities in the society. The minority itself is likely to respond with strong attitudes of group loyalty and to develop forms of behaviour that, by design or not, help to segregate its members still further from the rest of society.

"Minority" and "minority group"

The sociological employment of the word minority, then, largely agrees with common usage in denoting a distinct, separate group of people who are different in certain easily recognized aspects from the majority. But the term minority group implies rather more; for to a sociologist a social group of any kind is an aggregate of people with defined aims and rules of membership and with its own obligatory rules of behaviour and a subculture that publicly mark it off from the rest of society. It is the use of particular forms of cultural behaviour, used as criteria or emblems, that marks off a minority group from other types of group.

The word minority denotes by implication a part of a larger whole, but a minority group in the sociological sense is not always a numerical minority of the population. In parts of the southern states of the U.S., blacks form a clear majority of the population, but they are nonetheless a minority group in relation to the numerically smaller dominant group of whites. A similar situation existed in East African towns under the former colonial rule of the British. Under British rule immigrants from India had settled in East Africa as traders and skilled artisans. Most of them lived and worked in the towns, where they formed the overwhelming majority of the population, although, in east Africa as a whole, the Indians numbered less than 1 percent of the total African population. Despite appearances in the towns, the Indians were clearly a "minority group." In South Africa, to take another example, the Bantu population, although many times more numerous than the dominant white group, are nevertheless a "minority group."

A minority group is generally but not always politically less influential than other sections of the population, but its political position is only one factor in distinguishing a minority from other types of grouping that may also occupy subordinate positions. What marks off a minority group in the eyes of members of a society are a number of other distinctive features: race (as, for example, in South Africa or the U.S.), culture, language, religion, or economic function. To ordinary citizens it is characteristics of this kind that matter when they think of minority groups. A few examples of different kinds of minority groups should make these various points clearer.

After the Norman conquest of England in AD 1066, the native Anglo-Saxon inhabitants still constituted the majority of the population, but their culture and the social forms sustaining it were subordinated to Norman-French political, social, and cultural institutions. Like Africans, many Asians, and American Indians under the former European colonial empires, they lived as a linguistic and cultural minority group.

Religious minorities

The demarcation of minorities by religion is possibly more familiar from historical sources than from contemporary situations, where, especially in the Western world, the adherents of separate religious persuasions are more commonly regarded as categories of the population than as coherent groups of people with distinctive religious, cultural, economic, and political aims, which they pursue in their own interests. But even today such situations are not uncommon. In the northern parts of west Malaysia, for example, Thai-speaking Buddhists form an important religious and cultural minority group, whereas a few miles across the border in southern Thailand it is the Muslim Malays who are a religious and cultural minority. In India and Pakistan the confrontation of Hinduism by Islām led to the emergence of minority groups whose differences, though expressed primarily in a religious idiom, had important political and economic aspects. When the English consolidated their conquest of Ireland in the 16th and 17th centuries, a small number of English-speaking Protestants gradually spread out through a much larger Irish-speaking population of Roman Catholics and subordinated them religiously, culturally, economically, and politically. But only in some contexts were the Irish regarded as a purely religious minority group: in most situations they were seen, and felt themselves, to be primarily political and economic subordinates.

Economic minorities

Any social group formed for particular aims is likely to develop distinctive cultural features, but a group whose ends are primarily economic seems to do so less frequently, unless its organization is also intimately associated with other principles, such as race or kinship. Shopkeepers in any society are people with a number of shared interests. They are usually united in common associations, such as chambers of commerce, but it is unusual for them to constitute a minority group in the strict sense. In certain circumstances, however, they may do so. During the 19th and 20th centuries large numbers of Chinese people emigrated to Southeast Asia. Their motives in leaving China were almost exclusively economic: it was hard for them to earn a living at home. In Southeast Asia most of them eventually became traders, forming closely knit economic groups. The indigenous inhabitants of Southeast Asia, however, as much as the Chinese themselves and their colonial rulers, always saw these groups as Chinese first and only secondly as economic. In other words, the immigrants lived as Chinese ethnic groups with economic functions rather than as economic groups, most of whose members happened to be foreigners.

During the same period there was also a large emigration from the subcontinent of India to places as far apart as the Caribbean, Fiji, East and South Africa, Mauritius, and Great Britain. Like the Chinese, these Indian emigrants left their homes for economic reasons and settled abroad as trading or labouring minorities, but, as in Southeast Asia, the people among whom they settled regarded them primarily as cultural and ethnic groups.

The form of a minority group and its position in a particular society depend on many variable factors. Its members, as has been seen, may be bound to one another by common religious ideals or by similar political, cultural, or economic interests, or the tie may be one of race. In East Africa the Indian immigrant settlers distinguished themselves from the Africans and the Europeans first by the differences of their Indian languages and their Indian culture, but they also used the criterion of race to separate themselves, especially from the Africans. In Great Britain the dominant English group uses race as the main criterion to segregate ethnic minorities such as immigrants from the Caribbean, India, Pakistan, or East Africa. But however a minority group may be demarcated, for it to maintain itself as a group its organization normally must include political, economic, and cultural aspects. The emphasis placed on these various factors determines the form that the minority group takes and the place that it is assigned in society. Further, the emphasis itself varies with factors such as the group's size, its ideology, and, to some extent the outcome of these two, its wealth and political power.

The nature of ethnic groups and other types of minority groups. A social group, as noted earlier, may be founded on almost any kind of common interest, but, as also noted, strict reference to minority groups should be made only when the interests that bind the members also serve to separate them from the rest of society in ways that are felt to be significant by everybody. In most Western countries Freemasons are an organized minority of the male population, but to most people their characteristics do not mark them off as a minority group in quite the same way as do the distinctions of race for blacks in the U.S., for example.

"Race" and ethnic groups

Physically men vary widely, and in some societies selected bodily characteristics, lumped together under the broad heading of "race," are used as marks of social differentiation and for the formation of minority groups. If race is thought to be important, then it is likely to be used in forming groups that today are more usually called ethnic groups than minorities. (Some writers, especially in the U.S., have applied the term ethnic group also to immigrant groups who are distinguished by cultural differences

in language and national origin but who have no distinguishing physical characteristics. They can, if they wish, sometimes shed their particular cultural characteristics and disappear into the wider population in ways not open to the members of ethnic or racial groups who carry the marks of difference on their persons.)

Not every society whose population is racially diverse elects to make use of race for social purposes. In much of the Greco-Roman world, in many parts of past and modern Islām, and in some areas of Latin America a man's racial attributes may have as little significance as does the possession in northern Europe of dark or fair hair. The variations in the use of physical characteristics for social purposes can be placed illuminatingly on a scale. At one end are those societies in which racial attributes, if they are noticed at all, merely place people in ethnic categories, in much the same way as the males and females of a population are often separated for social purposes. Among the ruling classes of Mexico, for instance, the physical differences between people of predominantly Caucasian and Amerindian descent are very apparent, but in many situations of social life they are unimportant compared with the possession of a common pattern of social positions in society and a shared culture based on the Spanish language. Such a society is at the opposite end of the scale contrasted to one like South Africa, which makes extensive use of racial criteria in the allocation of group membership.

An important characteristic of a social category is that it can easily become a group if its members are led to hold values in common and to follow rules binding on all members. In the same way a social group, whose members lose their common purpose and sense of corporate identity, can become no more than a social category or even disappear altogether. Indian immigrants in South Africa, for example, were heterogeneous in origin and social composition, and in the earlier stages of their settlement they were loosely united only by some common, but vague, qualities of "Indianness." Nevertheless, they did constitute a distinct and recognizable element in the population, thus making them a social grouping more cohesive than a mere category but less cohesive than a fully organized group. Later changes in South African law compelled the Indians to assume more and more common, obligatory rules of behaviour that were different from those of other sections of the population, thus consolidating them into an increasingly corporate and separate ethnic group.

CHARACTERISTICS OF CULTURAL DIVERSITY

It is clear that any society may include (and that most have and do include) a large number of subgroups, or constituent groups. The manner in which the constituent groups are arranged and related to one another determines the social form or structure of the society. Minority and ethnic groups do not in principle differ from other social subgroups; however, their aims, the rules by which they recruit members, and their internal organization may well be different and impose special forms of behaviour on their members. Just how a minority group is related to the other sections of a society depends in part on the structure of the society itself and in part on the structure of the minority group.

An integral aspect of the structure of any social group is the set of ideas and beliefs held by its members, which largely determines their aims and, to a greater or lesser degree, guides their actions. The ideology of a minority group is an important element in the place that it is allowed to occupy in a society, but as important, and often more significant, is the ideology of the dominant group in the society. Its members and those of the other constituent groups exercise all kinds of political, economic, and other pressures to produce a situation that as far as possible satisfies them. But it is rare for any group, whether dominant or not, to have the power to impose a situation that is wholly of its own choosing: the process is almost always a two-way, reciprocal one.

There are, roughly speaking, two ways in which the members of a society may try to solve the problems posed by the presence of a minority group. They may either attempt to eliminate it altogether or decide to tolerate it. Both methods, separately or combined, with good grace or bad grace, have been used many times in the past. They are still being used today.

Elimination of minorities: assimilation, suppression, ejection. When the physical features, aims, or behaviour of the members of a minority group are unacceptable to the rest of society, there may be attempts to get rid of the minority altogether—either by forcing or encouraging its members to adopt the dominant group's culture and so be assimilated or by completely ejecting the minority from the society.

Assimilation

Assimilation (sometimes called acculturation) has been defined as those phenomena that result when groups of individuals having different cultures come into continuous first-hand contact, with subsequent changes in the original culture patterns of either or both groups. The process is a familiar one in the modern world. During the 19th and 20th centuries very large groups of immigrants, many of them speaking foreign languages, were encouraged to settle in the United States, but the possibility that they might develop into permanent, compact minority groups was felt to be a threat by the largely Protestant, English-speaking dominant group of American society. Prolonged and successful efforts were therefore made by means of school education and ideological indoctrination to assimilate the immigrants and their children as rapidly as possible. A similar policy of directed acculturation has also been successfully followed in Thailand, where the governing elite felt that the large and economically powerful minority of immigrant Chinese traders presented a political and cultural danger to the Thai people. The government, therefore, by means of legal inducements and restrictions on business and family life, combined with compulsory Thai education, have, without physical force and in a relatively short time, almost completely eliminated the Chinese minority as an ethnic group.

Ejection and suppression

The contrasting method of ejecting an unacceptable minority (and coercing those who chance to remain) is perhaps only a little less common than peaceful assimilation. A combination of ejection and suppression has been used many times in many parts of the world. In the 17th century Protestants in France were banished or driven into concealment if they refused to change their religion. In the same century the small minority of Christians in Japan was proscribed and eradicated. In modern Indonesia the Chinese trading minority, as in Thailand, was seen as a threat to national interests; the traders were therefore violently driven out of the country or forced rapidly to adopt most aspects of Indonesian culture. In the newly independent countries of East Africa the dominant African group, considering the long-settled Indian traders to be a threat to their interests, physically ejected as many as they could and severely limited all spheres of activity open to those who remained. Perhaps the most dramatic example of extermination or attempted extermination was the fate of the Jews under Nazi rule.

Assimilation necessarily takes time. It is usually a reciprocal process, and it is not always the culture of the dominant group that survives. The Greek invaders who followed Alexander III's armies and set up independent kingdoms in northwestern India disappeared almost without trace. For 200 years after the Norman conquest of England, the French language and culture were supreme, but English, a modification of Anglo-Saxon, was the language finally spoken in the country. The dominant culture evolved into a combination of French, Norman, and Saxon institutions; the same eclectic process can be seen at work in any postcolonial country in the world today.

Pluralism: toleration of minorities. A minority may be tolerated either because its aims and organization are not felt to be sufficiently objectionable to be eradicated or because the dominant group in the society, even if it dislikes the minority, may think that it is unable for political or ideological reasons or that it is morally wrong to assimilate, eject, or suppress it.

The rulers of several countries in western Europe in the 12th and 13th centuries found it hard to tolerate Jews, but so many of the rulers relied on Jewish trade and finance

that they could not afford to destroy them. The rulers were consequently forced to allow Jews to live partly on their own terms, difficult though it was in that particular feudal society to accommodate non-Christian aliens or persons without feudal status. In England this was done by segregating them in ghettos; in Spain and Portugal, where the feudal system was less rigidly organized, their freedom was greater.

During the 20 years after 1950, immigrants from the Caribbean, Pakistan, and India settled in Britain. The British did not welcome their arrival, even though their services were essential for running the economic and social system, nor did the British like certain aspects of the immigrants' culture, which combined with their physical appearance made the new minority groups socially very visible. But the ideology of the British people made it equally difficult for them seriously to consider eliminating these new groups either by ejection or by assimilation, and so they were reluctantly compelled to accept forms of cultural, social, and ethnic heterogeneity that had not before existed in their society.

The toleration of ethnic and cultural minority groups can produce a wide range of different types of social organization, most of which involve some form of separation. A minority may be completely segregated (as in medieval Jewish ghettos) even when separatism may not otherwise be a major structural principle in the society. Alternatively, separatism may run right through the system, as it does, for example, in Switzerland. In that country separate linguistic and cultural groups are effectively segregated by living in different districts, although all are united by a central administrative system, which, unlike that in most multiracial and multicultural societies, is not controlled by one dominant group. Other less extreme forms of separatism are produced by discrimination of varying degrees of intensity and by excluding a minority group from goods valued by members of the society. A society divided into minorities resolves the problems of pluralism by balancing the factors that make either for the complete elimination of the minorities on the one hand, or for complete separatism, short of fragmentation of the whole society, on the other hand.

TYPES OF SOCIETIES WITH CULTURAL DIVISIONS

The pluralist model

Pluralistic situations are common in all parts of the world and have given rise to much discussion. The term pluralism is used by political scientists for the view that political, cultural, and social systems are most usefully considered as being made up of separate but interdependent groups whose qualities together produce a characteristic and valuable moral order or unity. After World War II, pluralism was used and developed to describe and interpret societies divided into cultural and ethnic minorities and held together by the political power of one dominant group. J.S. Furnivall, who first began to limit the term in this way, wrote in 1948:

> In Burma, as in Java, probably the first thing that strikes the visitor is the medley of peoples—European, Chinese, Indian and native. It is in the strictest sense a medley, for they mix but do not combine. Each group holds by its own religion, its own culture and language, its own ideas and ways. As individuals they meet, but only in the market-place, in buying and selling. There is a plural society, with different sections of the community living side by side, but separately, within the same political unit. Even in the economic sphere there is a division of labour along racial lines.

The stratification model

Whether or not such "plural societies" are usefully considered as a special type is a matter that has provoked much argument. For one thing, critics would contend that all complex societies are more or less culturally heterogeneous and that to list alternative values and institutions within individual societies does not often help to distinguish one type of complex society from another. For reasons of this kind most social scientists have preferred to use the traditional models of stratification in analyzing social diversity. A society in which technology and the economy are able to support a large population is usually divided into broad divisions or strata—classes, castes, or estates—that form a hierarchy of wealth, power, and prestige. Such strata are marked off from one another by special rules of behaviour or forms of subculture, which are to a greater or lesser extent obligatory on all the members of each stratum, and which give them a graded access to goods (such as wealth, power, and prestige) that are valued by everybody in the society. Even though these broad divisions may not always correspond with the sections of a plural society or with ethnic or minority groupings, they do correspond often enough to merit further consideration of stratified and unstratified forms of society in reference to minorities and ethnic groups.

Relatively stratified societies. In any system of stratification the forms of behaviour and the rights of graded access to valued goods, which mark off one stratum from another, must be both objectively observable and subjectively accepted by most of the members of the society, all of whom must also occupy only one place in the system at one time. These criteria are sufficient to distinguish a system of ranked social groups, as, for example, that of the upper, middle, and lower classes in England during the 19th century, from an ordering of ranked statuses or positions, such as was also found in England in the same period in the rules of precedence that placed the queen's subjects in descending order below her, from the degrees of peerage to the degrees of commoner status. These criteria alone are, however, insufficient for distinguishing different types of stratification, any of which may be found in a society where ethnic or other forms of minority group are component elements of the system.

Closed stratification. A system of social stratification may be closed or open. If an individual's place, once assigned, can in theory or practice never or rarely be altered, the system is closed. If, in contrast, social mobility from one stratum to another is permitted, then the system is open, and the relation of one stratum to another is necessarily competitive. In a closed system not only is place assigned but shares in scarce and valued resources are also differentially assigned to the separate strata. As places and resources are graded and fixed, it follows that if a closed system is to work the relations between the strata must be cooperative to a considerable degree.

Caste systems

A traditional Hindu caste system was a small self-contained hierarchy of endogamous groups in which an individual was given a permanent position at birth. Relations between the groups were governed by religious rules, the breach of which was believed to produce a state of dangerous impurity that carried ritual, legal, and other penalties.

The English word caste refers to two separate systems of social relations. In the one, the traditional fourfold division of Hindu society into four *varṇa* or castes (priests, soldiers and administrators, businessmen, and labourers) is sometimes described in Sanskrit literature in terms of what appear to be ethnic differences, but at least in historical times the *varṇa* have not constituted a system of stratification. Sociologically speaking they are categories of value, which refer to all parts of the Hindu world and which people use, among other things, for placing themselves and the subcaste group to which they belong in the hierarchy of districts outside their own local caste system.

The word caste also refers to the several thousand small, self-contained, local hierarchies of endogamous subcaste groups, or *jāti*, of which Hindu society was made up and of which a Hindu state might include a great many. Each *jāti* was named and ranked, though the ranking might not be the same even in adjacent local hierarchies. Traditionally each subcaste had a particular service assigned to it, so that the sons of goldsmiths or priests, for example, though not bound to follow these occupations, could prevent others from practicing them. In India neither castes nor subcastes were usually ethnic or cultural minority groups in the usual sense of the word, but in Iran and other parts of the Near East, cultural minorities are sometimes components in small-scale caste systems. It is possible, too, that in some areas of the southern states of the United States society could be most usefully described and interpreted by using a similar model of a local caste system.

Estate systems

If one is content to say, as are many, that in both South Africa and India social stratification is based on caste,

then the concept of closed stratification need be examined no further. But writers on India do not usually agree that "colour-bar" societies are necessarily caste systems, and certainly in South Africa, although individuals are assigned positions in ranked strata at birth, the society is large scale and is not made up of many small, local hierarchies. Most important, as in most industrial societies, the ties between individuals are not multiple: the fact that a man employs another does not usually mean that he is also his landlord, his creditor, his lawyer, a member of the same church congregation, and attached to the same political party. Thus in many of its aspects the South African system of closed stratified ethnic groups is nearer to the model of a feudal estate system than to that of a caste system.

In Europe feudal estates were defined and ranked by law, and penalties for a breach of the rules of rank were legal, not ritual as in a caste system. Admission to the estates was regulated by birth and a few other carefully stated criteria, so that a man's social standing as a noble, a cleric, or a commoner was legally defined and carried with its rights and duties shared by all members of his estate and not by others. Estates were also political in nature in that representatives for assemblies such as Parliament in England and France were recruited according to their rank. An estate was a closed group whose members could not in theory move out of it into another.

The population included in an estate system might, unlike that in a local caste system, be so large that the people in it did not have the multiple ties and close knowledge of one another, characteristic of a caste system and perhaps necessary for the working of the intricate and close-knit network of ties and regulations that held it together.

In writing of ethnic minorities in contemporary societies or of slaves in older societies, authors often speak of "caste"; and though they are generally aware of the fact that they are using an analogy, nevertheless misleading implications are frequently carried by the metaphor and are not always discounted. In analyzing the place of highly segregated minority groups in a large-scale society, a model based on a caste system is seldom as useful as one based on an estate system.

Social class system

Open stratification. In some situations and places, as in parts of the Caribbean, ethnic or culturally distinct groups are often treated as, and in fact may be, social classes. Sociologists generally consider a social class to be an aggregate of people in roughly the same social position, which is different from that of other social classes, and which, unlike position in a caste or an estate system, allows movement from one stratum to another. Because mobility is permitted it is not always easy to draw a boundary between one class and the next, nor is it always easy to decide how far the members of a class are bound to one another by ties that make the whole aggregate an institutionalized group. Furthermore, in an open system of stratification the scarce resources valued by the society are not overtly and explicitly allocated to each ranked stratum, and relations between classes are therefore necessarily primarily competitive, not cooperative as they often are in a closed system, where, within limits, an individual's share of resources is prescribed by his rank.

When some of the qualifications for belonging to a social class are also those for belonging to a minority group the difficulties of analysis may become very great indeed. In the later stages of the Spanish Empire in Mexico the qualities of being Indian—qualities such as physical appearance, the use of an Indian language, and a peasant or servile status—were also the marks of low class position. At a later period, with the acquisition of wealth, Spanish culture, and political influence, individuals from the lower classes could and did move in relatively large numbers from one social class to another. In the process the boundaries both between the social classes and between the ethnic groups often became blurred.

When social mobility in a system of stratification becomes very general and rapid, the system may sometimes change quite quickly into one of ranked individuals and not of ranked groups. In the opinion of some students of society, this condition has already been reached in the United States and may be a contributory factor in the difficulty that is sometimes met in absorbing certain minority groups. For instance, other members of American society are more likely to regard Negroes as stereotyped representatives of a minority group with prescribed social roles to play than they are to regard them as individuals making their own social roles outside fixed social groups. To the degree that Negroes are seen in this way they are encouraged to conform with the general view and assume stereotyped roles. Because they lack the greater freedom that other citizens have in choosing roles and in moving from one position to another, they often become even more effectively segregated.

The traditional models of stratified societies, which have been discussed here, are of course merely "ideal types" that have been constructed for studying social processes, which in different combinations produce different types of society. Any real-life society may in fact never quite match any of the ideal types, whose functions are simply to isolate the social processes that, acting together, produce particular situations. The models are useful in helping to think about such situations and in allowing a minority group to be placed in the structure of the society of which it forms a part. Much of the thought about minority groups in the past has been confused by focussing attention on inessential attributes. It is not the blackness of the Bantu in South Africa that is socially significant: it is their position in the formal structure of the society.

Relatively unstratified societies. In all societies there is a division of social labour and some diversification, but not all are stratified. Most of the discussion of pluralism has been of racially and culturally diverse populations within the jurisdiction of a central government—in such countries as Switzerland, Malaysia, and Canada. But there are other types of plural situations in which separate ethnic or cultural groups live within a common and accepted system of relationships that does not amount to either an effective central administration or a system of ranked groups. Systems of independent but allied groups of this kind were not uncommon, especially in tribal areas, before the expansion of European colonial regimes. In the eastern woodlands of North America, for instance, a group of Indian tribes known as the Iroquois had confederated for purposes of making decisions in common, even though central administrative control was lacking.

Unstratified societies with effective central administrative systems. In Switzerland, as has been noted, cultural and linguistic groupings are sharply differentiated and maintained. Within itself each group may be separately and differently stratified, but access to power in the wider society does not seem to be limited by local or cultural origins, nor does it seem to be limited by any ranking of the separate cultural groupings.

In western Malaysia the population is evenly divided between Malays and Chinese. There is also a small minority of Indians and even smaller ones of aboriginals, Thai-speaking Buddhists, and a few other small groups. The aboriginal population is almost wholly separated from the rest, and the majority of the Malays live in rural districts, whereas the Chinese are mostly urban dwellers. The ethnic groups are not segregated by law. The political constitution provides that the men who direct the machinery of government and hold legislative powers shall be impartially selected by voting. In practice electoral procedures and the staffing of the army and the civil service have put most political power in the hands of Malays. The greatest economic strength is in the hands of the Chinese. This distribution makes a permanent balance of power difficult. The fortunes of the Chinese could be threatened by political action, and they might use their wealth to upset the political balance and so threaten the Malays. Under the colonial rule of the British, the Malays, the Chinese, and the Indians were unorganized ethnic categories; but since independence the categories have for political purposes become more and more organized as corporate ethnic groups. Neither the Malay nor Chinese group is completely dominant, and in the strict sense neither is a minority group.

The situation in Malaysia, where no ethnic group is completely able to dominate the central government, is

exceptional. The situation in Kenya, Tanzania, or Uganda is more usual. Since independence the Africans have been the dominant group in all those countries, which also contain Arab, European, and Indian minorities. Like the Chinese in Malaysia, the Indians in East Africa are economically strong, but their numbers have always been far too few for them ever to think of making a bid to take over the central administration and run the country in their own interest. Indians, for various reasons, are now regarded by the Africans as an objectionable minority group, and for ideological reasons they are being driven out of Kenya and Uganda.

The effects of the rise of the nation-state

Toleration of minority groups depends in part, as noted earlier, on such factors, as size, wealth, access to power, and ideology. In modern nation-states wide differences of ideology are not as a rule very willingly permitted, possibly because the idea of a nation implies that if the nation and the dominant group controlling it are to survive, its citizens ought to hold strong, common, national values. The difficulties in deciding how far divergent minorities and ideas can be tolerated is not new, nor are they confined to new nation-states with ethnic and cultural minority groups held together only by the political power of a dominant group. In the 16th century France was divided between Roman Catholic and Protestant Christians, and each group, for its own safety, felt that it had to capture the central government. The victorious Roman Catholics, for their own sake and the safety of the French nation, as they saw it, did not cease their war until the Protestant minority was virtually eliminated.

The idea of a nation-state and the idea of tolerating a minority are to some extent in opposition to one another. The very long-standing solution to the problem reached by a country like Switzerland is unusual. The balance reached in such countries as Malaysia or Canada sometimes gives the impression of a temporary solution, which may at any time give way to a unitary state dominated by one group or to the secession of the minority to form a new nation. The dilemma is important in many parts of the world, especially in such areas as the Caribbean, the United States, and southern and southeastern Asia, where cultural distinctions have often fused with political hostility. It is this widespread fusion of ethnic and cultural hostilities as one of the main elements of political programs that has undoubtedly drawn many to the concept of a plural society. The concept does not give a satisfactory account of the structural characteristics of pluralistic societies, but what its advocates may be attempting to describe and analyze is a political situation in which ethnic and cultural loyalties are felt to be more important than the interests of class or party or any other group in society. This transformation of ethnic and cultural hostilities into political programs is closely related to the problems of minority groups and their place in society.

Unstratified societies without effective central administrative systems. In the past, pluralism has usually been discussed as if it were a feature only of societies with a central administration. It is, however, not uncommon to find ethnic and cultural groups, whether adjacent or intermingled, that live within a common system of relations constituting an alliance of some kind but not an effective central administration.

"Tribal pluralism"

In northern Thailand, Burma, Borneo, New Guinea, and other parts of the world small independent groups of ethnically and culturally distinct peoples were dispersed and mingled over wide areas without being subject in any effective sense to a central government. Among the Kachin of highland Burma, peoples of different languages and cultures, sometimes resident in the same villages, were united in a common network of marital, political, and trading alliances. On the Baram and Rajang rivers of Sarawak in Borneo, villages having different cultures and speaking different languages were interspersed along the banks. Each village was politically independent with its own characteristic social structure. In most villages the inhabitants were rigidly ranked: those of the lower ranks married and lived, for the most part, within their native villages; but those of the upper ranks, both men and women, married and often moved into other villages and other cultures for political and economic advantage. In most of these societies ranking was a fundamental principle of organization, but the different villages and cultures were not ranked against one another. Each was different from and none was considered better than the other. The system of alliances that held together a considerable population, culturally and linguistically divided, had certain similarities to the system by which modern nation-states maintain their independence and diversity.

THE FUNCTIONING OF MINORITIES AND ETHNIC GROUPS

Minority groups have performed almost every type of function in almost all known types of social system. Minorities have played economic roles as specialist bankers, traders, or craftsmen; more often they have been manual labourers, whether slave, serf, or free. As revolutionary groups they have acted as ideological leaders in political life; sometimes their function has been to hold the political balance; more often they have been the subject of exploitation. Some, because they have special skills, may be valued for their artistic contributions to social life and yet not be altogether socially acceptable to the rest of society. Earlier it was noted that the role played by a minority group within a society varies with such factors as its size, its skills, and its ideology and also with the roles of the dominant and other groups in the society. The amounts of wealth that the members of the group can accumulate and the political influence that they can exercise vary with the ways in which these factors arrange themselves. Factors such as these obviously work differently in different types of society, and even in the same type the balance is never likely to be identical in any two societies.

Their economic roles. *As specialists.* A division of labour is a feature of all societies, even the simplest, but unless the population is large enough to free specialists from the task of finding sufficient food and shelter for themselves and their dependents, the society is unlikely to be able to support wholly specialized groups. Without a sufficiently large population, too, a system of social stratification cannot develop; for a graded distribution of resources among the social strata is possible only if there is a sufficiency of basic goods for everybody.

In a system of closed stratification, in which movement from one stratum to another is forbidden, and in which resources are differentially allocated to the strata, functional specialization is likely to be a feature of the system. Specialized groups, handling resources that are the property, so to speak, of their members and of their stratum, are likely to resist change and defend their specialization and the resources that are needed for it. In an open system of stratification, on the other hand, where social mobility is the reward for talent of various kinds, the temptation to form perpetual, specialized minority groups is probably not so great.

In caste and estate systems

In a Hindu caste system the allocation of economic, political, and ritual roles is probably more formally and rigidly arranged than in almost any other kind of society; but it is not usual to consider the specialized subcastes, which make up the system, as minority groups. On the other hand, the outcaste groups, who were said by the Hindu to be altogether outside the system and whose menial occupations and servile structural position separated them, often physically, from the rest of society, are probably best understood when regarded as minority groups. Similarly, in a feudal system the roles of the estates were strictly prescribed: the nobles fought for all, the clergy prayed for all, and the commoners worked for all. Insofar as the nobles dominated society, both the clergy and the commoners could be considered minority groups, though it is not usual to call them so, for medieval European society had in it a number of smaller groups that were unambiguously minorities in the usual sense. The structure of a feudal system allows little place in it for trade, which indeed was not an approved activity in medieval Europe, though of course it was necessary and was carried on by specialized minority groups such as the Jews or by anomalous minorities such as the merchants in semi-independent towns.

In the colonial and postcolonial societies of the modern world, the population is more often than not ethnically and culturally diverse and is often also rigidly stratified. In these societies specialized minority groups are common. When the British established a colonial regime in East Africa at the end of the last century, the African population lacked most of the technical and trading skills that were needed to develop the area and produce a revenue for the administration. Immigrants from the west coast of India were therefore encouraged to move into the interior regions and set themselves up as a trading minority. They differed from the Africans and the Europeans in physique, language, religion, and culture. They were absorbed into the hierarchy of colonial society as a specialized minority group, with a virtual monopoly of retail trade and a position of wealth, influence, and prestige greater than that of the Africans but lower than that of the Europeans.

After independence the hierarchies in the East African countries were rearranged. The position of a non-African minority group with a major share of the country's retail trade was ideologically unacceptable to the new African rulers. The countries they ruled were ethnically and culturally highly diversified, and their main concern was to encourage the growth of a unified nation, whose citizens shared common, patriotic values, even, if need be, at the expense of equality of opportunity. The rulers therefore adopted policies designed to eliminate non-African minorities such as the Indians. With local variations the same situation can be found in all parts of Southeast Asia, where rulers of dominant groups have attempted either to absorb or to drive away Chinese and Indian minority groups that have specialized in trade or crafts.

As labourers. Minority groups have occupied servile or labouring positions in every type of society. The Bushmen of Botswana in southern Africa are a minority group loosely attached as herdsmen and servants to Tswana society. Pagan tribes, immigrants, and conquered groups have on many occasions been brought into Hindu society as servile minorities, usually as outcasts outside the caste system, but sometimes (though more rarely) they have been incorporated as subcastes on the lower rungs of the hierarchy. In England after the Norman conquest the servile roles of the feudal system were mostly occupied by the previously free Saxons, the Welsh, and other native minority groups. Ethnic and cultural minority groups of African slaves and Amerindians filled the lower strata of the rigidly organized class systems of the 18th century in the Caribbean area. In the less rigid class system of modern Britain immigrant minorities from the Caribbean, Pakistan, and India for the most part are labourers.

The labour of ethnic groups

Because the distinguishing marks of an ethnic group are physically carried by its members, it is harder for them to be assimilated. And if racial attributes are important in the ideology of the dominant group of the society, it may be easier for a cultural minority to improve its status than it is for an ethnic group. The Irish who settled in the United States as labourers and various eastern Europeans who followed them have found it possible (at least in the second or third generation) to be socially mobile by changing their cultural identities in a way that has not yet been really open to blacks. It is not entirely accidental that ethnic groups tend to be more segregated and restricted and therefore regularly found among the poorer classes of a society. As long as the ethnic groups reside in the lower ranks of the social system, the privileged classes —unthreatened—can talk easily and abstractly of the desirability of equality of opportunity and the undesirability of segregation and restriction of opportunity. If an ethnic minority does acquire wealth or influence, however, either through its own efforts within society or through circumstances outside it, then the dominant group is usually compelled to make the necessary changes in its practices as well as ideology to accommodate if not assimilate the minority. But the process is never easy and is seldom willingly undertaken.

Their political roles. The political roles open to a minority depend on many of the same factors as limit its economic opportunities. Its numbers may be too small, as were those of the Indian immigrants in East Africa, for it ever to contemplate making a bid for control of the central political machinery, whereas in Malaysia the Chinese were sufficiently numerous to threaten a takeover of the government and thus could protect themselves. In fact, they accomplished the takeover in Singapore. Many minorities, however—as, for instance, merchants whose main interest in a political system is the preservation of law and order needed for peaceful trading—do not possess an active political ideology. It is only if circumstances force them and the structure of the society permits it that they think of protecting themselves by direct action. A minority group by definition is subordinate, and its usual position is near the bottom of the power structure, especially if its values diverge markedly from those of the main body politic.

Minorities acquiring power and leadership

The opportunities for acquiring power depend not only on a minority's own characteristics but also on the nature of the larger society. In a Hindu caste system the hierarchy is one of ritual prestige. The priestly groups are endowed with the greatest sanctity and honour. But ideally secular power ought not to be in their hands. It is the prerogative of the administrative or warrior castes. In practice, political power under Hinduism has often been held by dominant priestly or merchant castes, and even by labouring castes, without wrecking the system. In theory, too, a society of feudal estates ought to be ruled only by nobles; but if, as happened at the Norman invasion of England, the conquerors were of mixed origins, they were soon attributed noble status and the system survived.

In an open class system, if a minority group achieves power in a modern society and displaces a dominant group, the distribution of roles among most of the ruled classes is unlikely to be fundamentally altered, whatever the fate of individuals. This situation can be documented from many parts of the Communist and the postcolonial worlds. In Ghana under the British colonial regime, the African elite was a minority group that was drawn from several indigenous linguistic and cultural groups and whose members were often foreign educated and English speaking. When this group took over the colonial administration from the British as the dominant group, the society in its broad essentials altered little, except that the Ghanian elite ceased to be a minority group.

As ideological leaders. A favourably placed minority may occupy a strategically important position in a political system without being the dominant group. After their defeat in the Boer War at the beginning of this century, the Afrikaners were sociologically speaking a minority group though numerically a majority of the white population of South Africa. Their ideology differed from that of the dominant English group, though both had much in common concerning ideas of race, stratification, and the political superiority of whites. These ideas were directly opposed to the idea of equality of opportunity, which was paradoxically held concurrently by the English. The ideology of the Afrikaners was less ambiguous, and in a real sense they were the ideological leaders of the whites in South Africa. When they eventually gained control of the central administrative system and, as the dominant group, began to put their ideas on race and minorities into practice, there was little real resistance.

In South Africa a linguistic and cultural minority was able, as in other parts of the postcolonial world, to become dominant without a military conquest or revolution. In the Soviet Union, in Mainland China, and in other Communist countries, the Communist Party has usually gained power by force. In all countries the party began as a minority group based on a political philosophy in which differences of race, language, and culture were not considered to be relevant criteria of membership. In the period before it gained control of the central administrative machinery, the party was undoubtedly, especially in many of the undeveloped tropical areas, a minority group that acted as an ideological leader and an instrument of social change.

As arbiters in balances of power. A minority group may be in an important political position in a society and yet not be able to take over the administration and rule it

in its own interests. In parts of the European colonial empires, for example, trading immigrants were disenfranchised minorities; but because their economic power was great in the countries in which they had settled, they could frequently persuade, and sometimes force, colonial rulers to modify policies unfavourable to trade, though possibly advantageous to the rulers or other groups in the society. In Hawaii in the 1890s the white men who dominated the economy even persuaded the United States to annex the islands after a new U.S. tariff had threatened their sugar exports.

In the latter part of the 19th century the Irish formed a minority group in Great Britain, and their parliamentary representatives in London were sufficiently numerous to hold the balance for a time in the political struggles of the rest of the United Kingdom. A not dissimilar position was held for a few years by the Cape Coloureds of the Cape Province of South Africa before their vote was removed.

The risks of arbitration

The role of arbiter, of choosing whether to give or refuse support, can be a dangerous one for a minority group. Even that of a "go-between" who mediates between the rulers and the ruled members of society may bring hatred and disaster in its train, as the Anglo-Indian community discovered when British rule ended in India; and as Chinese and Indian groups who had specialized in working with colonial civil services also discovered when former colonial territories gained independence. In situations where a minority holds a balance of power, it may feel obliged for the protection of its own interests, as were the Afrikaners, to become dominant in the society. On the other hand, it may be safer for the minority, if it can, to secede altogether and become independent before it is stripped, as was the Cape Coloured population of South Africa, of all political power. The Irish solved the dilemma by secession from the United Kingdom, and the Chinese of Singapore by secession from Malaysia.

As exploited groups. In the discussion of the economic functions of minority groups it was noted that most often they were servile or exploited and that both the form and function of the group largely depended on the form and ideology of the society of which they were a part. It was also seen that the degree of subordination or exploitation and the type of relation that a minority can have with a dominant group is variable. Thus, for instance, Brāhmaṇ priests in a caste system are the ritual servants of the other subcaste groups and may be exploited politically and economically by a lower ranked but dominant group, and yet the ritual relationship is not a servile one. Similarly in feudal Europe the nobles and the clergy exploited and segregated minorities such as the Jewish trading community, but the relationship differed in its degree of subordination and exploitation from the relationship of the nobles to serfs and slaves. Nor was the relationship of the Indian trading community in East Africa with the British rulers a wholly subordinate one, although the British exploited and used the Indians' commercial skills and political weakness. In British Guiana or Fiji the situation was different. There Indian coolies were imported as indentured plantation labourers when the suppression of the slave trade put an end to that source of cheap African labour, and as a result both the freed African and the indentured Indians, though not technically slaves, occupied an almost servile status.

Other roles. By its existence a minority group presents a challenge to the rest of society and its values. The challenge may be generated from within the society itself, as happened in England before the Reformation when Lollards formed themselves into minority groups for religious purposes without the approval of the church. On the other hand, the ideas through which a minority incorporates itself may come from outside the society. In the 19th and 20th centuries Christian missionaries have gone to countries in Africa and Asia with the deliberate intention of forming minority groups to transform the whole society. But whatever the sources from which a minority derives its existence, in coming to terms with it, the rest of society is challenged; and whether the society assimilates, tolerates, or eradicates the minority, the society is likely to be changed in the process.

A missionary group, whether religious or political, exists for the purpose of changing something in a society, but relatively few minority groups are missionary by intention. More often than not the changes that they bring about are, so to speak, inadvertent. In the late Middle Ages a minority group of Flemish weavers settled in England because it was possible to earn a good living there. The technical changes they caused helped transform the economic system of English society and its shape in the 16th century. In assimilating Jewish immigrants from Europe, the dominant Protestant culture of the United States has been modified in its business and artistic habits in more ways than is always apparent. In the United States, moreover, although the blacks have been segregated and barely tolerated by other groups in society, they have transmitted much of their music and many of their aesthetic ideas to the dominant group and from them to the rest of the world. In parts of Europe Gypsies have sometimes played a similar role.

The problems of cultural diversity and therefore of minority groups are inherent in the nature of social life. The structural adjustment of groups within a social system can, as has been noted, be attained in many different ways, but the fundamental problems of diversity remain, and in any foreseeable future probably must remain much the same as they have been in the past. With the growth of good communications and facilities for travel, racial and cultural diversity in societies all over the world is likely to increase rather than diminish. It is only by a clear understanding of the processes that make for diversity and by an ability to weigh the advantages and disadvantages of different answers to the problems of diversity that the people of the world, generation by generation, can learn to tolerate one another as well as, if not better than, their forebears.

BIBLIOGRAPHY. H. KOHN, *The Idea of Nationalism* (1944); L. WIRTH, "The Problem of Minority Groups" in R. LINTON (ed.), *The Science of Man in the World Crisis* (1945); and A.M. and C.B. ROSE (eds.), *Minority Problems* (1965), give a good background to the problems implicit in the ideas of nationalism and minorities, especially with reference to the West. L. KUPER and M.G. SMITH (eds.), *Pluralism in Africa* (1969), and H.S. MORRIS, *Indians in Uganda* (1968), examine the problem of ethnic and minority groups in non-Western societies and consider the subject of social and cultural pluralism in terms first raised by J.S. FURNIVALL in *Colonial Policy and Practice: A Comparative Study of Burma and Netherlands India* (1948). G.A. DE VOS and H. WAGATSUMA (eds.), *Japan's Invisible Race* (1966), discuss the same problems in a psychological rather than a sociological idiom. P. MASON, *Patterns of Dominance* (1970), treats a selected number of societies in different parts of the world and at different historical periods and evaluates various solutions for the problems of ethnic and cultural minorities. The discussion of ethnic and minority groups in terms of race relations is probably best looked at in A.W. LIND (ed.), *Race Relations in World Perspective* (1955). Of special interest for minority problems in the Western world are M. FREEDMAN (ed.), *A Minority in Britain: Social Studies of the Anglo-Jewish Community* (1955); E.J.B. ROSE *et al.*, *Colour and Citizenship* (1969); and C. WAGLEY and M. HARRIS, *Minorities in the New World* (1958). B. BENEDICT, *Indians in a Plural Society* (1961); F. and L.O. DOTSON, *The Indian Minority Zambia, Rhodesia and Malawi* (1968); S. PATTERSON, *Colour and Culture in South Africa* (1953); and D.E. WILLMOTT, *The Chinese of Semarang* (1960), give useful and illuminating descriptive accounts of ethnic groups and minorities outside Europe and America. The essays in F. BARTH (ed.), *Ethnic Groups and Boundaries: The Social Organization of Culture Difference* (1969), re-examine the traditional problems of maintaining ethnic or minority identities.

(H.S.M.)

Mirabeau, Comte de

Honoré-Gabriel Riqueti, comte de Mirabeau, was the greatest orator and probably the greatest political mind in the National Assembly, which governed France in the early phases of the French Revolution. Turning to politics after a dissipated youth, he advocated the course of constitutional monarchy, but his middle way was trusted by neither the court nor the radicals. His death in 1791 perhaps ended any possibility of moderation.

Mirabeau, portrait by Joseph Boze, 1789. In the Musée National de Versailles et des Trianons.
Giraudon

Troubled youth. Mirabeau was born at Bignon, near Nemours, on March 9, 1749. He was the elder son of the noted economist Victor Riqueti, marquis de Mirabeau, by his unhappy marriage to Marie-Geneviève de Vassan.

Ugly from birth and further disfigured by smallpox at the age of three, the precocious Honoré-Gabriel was soon on bad terms with his formidable father. At the age of 15 he was sent as a pupil to the strict Abbé Choquard in Paris, and at 18 he went as a volunteer to serve in a cavalry regiment at Saintes, where his father hoped that military discipline would curb him. His misbehaviour, however, led to his imprisonment on the Île de Ré, under a *lettre de cachet*, a written order permitting imprisonment without trial. Released to serve in Corsica with the rank of sublieutenant in the army, he distinguished himself there in 1769. Reconciled with his father, he married a rich Provençal heiress, Émilie de Marignane, in 1772, but his heavy spending and further misconduct led his father to have him imprisoned under another *lettre de cachet* in order to put him out of reach of his creditors. He was detained first at the Château d'If (1774), then at the Fort de Joux, near Pontarlier. Having obtained permission to visit the town of Pontarlier, he there met his "Sophie"—who, in fact, was the marquise de Monnier, Marie-Thérèse-Richard de Ruffey, the young wife of a very old man. He eventually escaped to Switzerland, where Sophie joined him; the couple then made their way to Holland, where Mirabeau was arrested in 1777.

Abduction of "Sophie"

The tribunal at Pontarlier had meanwhile sentenced him to death for seduction and abduction, but Mirabeau escaped execution by submitting to further imprisonment under a *lettre de cachet*. In the château of Vincennes he composed the *Lettres à Sophie*, some erotic works, and his essay *Des lettres de cachet et des prisons d'état* ("Of *Lettres de Cachet* and of State Prisons"). Released in December 1780, he finally had to surrender himself to arrest at Pontarlier in order to have the death sentence revoked; but by August 1782 he was entirely free. He now became involved in a lawsuit against his wife, who wanted a judicial separation. Pleading on his own behalf, he gained the sympathy of the public but lost his case (1783). Rejected by his wife and by his father, he had to renounce the aristocratic society into which he had been born.

For the next five years Mirabeau lived the life of an adventurer. He was employed sometimes as a hired pamphleteer, sometimes as a secret agent. He came into contact with Louis XVI's ministers Charles-Alexandre de Calonne; Charles Gravier, comte de Vergennes; and Armand-Marc, comte de Montmorin-Saint-Hérem. He also made an enemy of the Swiss banker Jacques Necker, at that time director of the finances, and engaged the playwright Pierre-Augustin Caron de Beaumarchais in controversy. His activities necessitated much travelling. In London he was introduced into the best Whig society by Gilbert Elliot (later 1st earl of Minto), who had been his fellow pupil under the Abbé Choquard; he had to take refuge in Liège when his *Dénonciation de l'agiotage* (against stockjobbing) annoyed Calonne; and he undertook a secret mission to Berlin in 1786. With the active assistance of a Brunswick friend, Jakob Mauvillon, he wrote *De la monarchie prussienne sous Frédéric le Grand* (1788; "The Prussian Monarchy Under Frederick the Great"), which he dedicated to his father; but *Histoire secrète de la cour de Berlin* ("Secret History of the Court of Berlin"), in which he made unscrupulous use of material derived from his mission in Germany, created a scandal in 1789.

Election to the States General. Within France, affairs were moving toward a crisis. The country, bankrupted by its 18th-century wars, was burdened with an archaic system of taxation and social privilege. The States General, an assembly of the three estates of the realm—the clergy, the nobility, and the commons Third Estate—was summoned to meet in Paris in May 1789 in an attempt to implement the necessary reforms. It was that meeting that set in motion the great French Revolution of 1789. (See FRANCE, HISTORY OF.) When the States General were summoned, Mirabeau hoped to be elected as a deputy for the nobility of Provence. For this he needed his father's support. Pleased by the book dedicated to him, the Marquis had summoned Mirabeau to Argenteuil in the autumn of 1788 but had not given him any real help. Mirabeau presented himself in the chamber of the nobility in the estates of Provence in January 1789 and uttered violent diatribes against the privileged classes but was not elected deputy, as he held no fief. Turning reluctantly to the Third Estate he was elected to represent both Marseilles and Aix-en-Provence, choosing to represent Aix.

Outbreak of the Revolution

Mirabeau came to the States General without any precise constitutional doctrine. An avowed enemy of despotism (he had written *Essai sur le despotisme* ["Essay on Despotism"] before he was 25), he was, nevertheless, a firm supporter of the monarchy and of the executive power. Without expressly adhering to the English system, he wanted representative government. A nobleman rejected by his class, he opposed the idea of an aristocratic second chamber. Like most of his contemporaries, he had no political experience, but his intelligence and his knowledge of men made him supremely capable of acquiring such experience rapidly. Lack of money, however, exposed him to pressure and to temptation.

From May to October 1789 Mirabeau played a decisive part in the battle between the Third Estate and the privileged orders. His aim was to become the spokesman of the nation to the King and at the same time to moderate the expression of the nation's wishes. Thus, on June 15 and 16 he was careful not to suggest the name National Assembly, which was the rallying cry of the Third Estate in its revolutionary debate of June 17, when it set itself up as representative of the whole nation. Yet, at the ending of the "royal session" of June 23, when Henri Évrard, marquis de Dreux-Brézé, in the King's name, ordered the assembled estates to return each to its separate chamber, Mirabeau's answer did much to confirm the deputies in their resolution to disobey and establish the National Assembly; and, in the feverish atmosphere of the early days of July, his speeches inspired the Assembly to demand the dispersal of the troops concentrated around Paris. After the fall of the Bastille (July 14), he urged the Assembly to demand the dismissal of the ministers who were to blame for the disorders. His popularity in Paris was then considerable. On the other hand, he disapproved of the Assembly's precipitate action in abolishing feudalism (on the night of August 4) and of the abstract Declaration of Rights; and, while he was openly against a second chamber, he yet wanted the king to have an absolute veto. In October, when the Parisians marched on Versailles and took Louis XVI back to Paris, Mirabeau's attitude was ambiguous and gave rise to the suspicion that he might be plotting against the King. To clear himself and to keep open the door to the court's favour, he addressed a memorandum to the King, advising him to leave Paris for Rouen, to secure the support of a small army, and to appeal to the provinces.

Contacts with the court

Mirabeau's prime concern, however, was to win "the battle of the ministry." Ostensibly a supporter of Necker, Mirabeau, in fact, did his utmost to destroy him: his brilliant speech on the bankruptcy of the nation was a masterstroke against this minister. Furthermore, he tried skillfully to induce the Assembly to grant to the king the option of choosing members of it to be his ministers; but the Assembly's decree of November 7, 1789, which precluded all deputies from the ministry for the duration of the session, frustrated his hopes of ministerial office for himself.

Intrigue with the court. From November 1789, notwithstanding his oratorical triumphs of January–April 1790 in the cause of the Revolution, Mirabeau was a prey to despondency and aimlessness until his friend Auguste, prince d'Arenberg, comte de La Marck, with the approval of Florimund, Graf Mercy d'Argenteau, Austrian ambassador to Paris and confidant of the queen, Marie Antoinette, approached him with the proposal from Louis XVI and the Queen that he should become their secret counsellor. Mirabeau accepted with delight: "I shall make it my chief business to see that the executive power has its place in the constitution" (letter of May 10). Part of the promised remuneration was to be the paying-off of his debts. In May 1790, when the Assembly was debating the king's right to make war and peace, Mirabeau successfully opposed the left-wing orator Antoine Barnave, whom he challenged with the words: "Tell us that there should be no king, do not tell us that there should only be a powerless, superfluous king." He impeded the progress of the Jacobins but risked his own popularity, and a pamphlet accusing him of treason was circulated ("Trahison découverte du comte de Mirabeau" ["The Uncovered Treason of the Comte de Mirabeau"]).

From June to October he had to work to recapture his prestige. This was the more necessary because the King and the Queen, despite their secret interview of July 3 with Mirabeau at Saint-Cloud, took little notice of his advice and continued to be influenced by his rival for court favour, the Marquis de Lafayette, who had scorned Mirabeau's offer of alliance. In October 1790 the Assembly further disappointed Mirabeau by refusing, after more discussion, to revoke the decree of November 1789 on the noneligibility of its members for the ministry.

While the court was displeased by some of Mirabeau's outbursts and by his "incurable mania of running after popularity," Mirabeau, for his part, was enraged to see a new ministry formed under the influence of his rivals Lafayette and Alexandre, comte de Lameth. By the end of November 1790 his relations with the court were severely strained. He restored them by submitting to the King's adviser Montmorin a "Plan" concocted for bringing pressure to bear by various means on the Assembly, on Paris, and on the provinces so as to coordinate "the means of reconciling public opinion with the sovereign's authority." The plan was perfect in theory but very difficult to put into practice. From January 1791 it was clear that Mirabeau had no intention of doing anything that might compromise his own popularity, though he was willing enough to sabotage the Assembly by getting it to adopt ill-considered measures of religious persecution and was eagerly and adroitly working to discredit Lameth's faction at court. His popularity rose to its zenith, and the eyes of all Europe were on him. As spokesman of the diplomatic committee, on January 28, 1791, he made a speech that bore the unmistakable stamp of statesmanship. Anxious to avoid anything that might compromise France's relations with neighbouring countries, particularly with England, he yet would not repudiate any of the Revolution's political victories or allow any necessary military precautions to be overlooked. On the following day he at last became president of the Assembly for a fortnight. In this office, from which he had been so long excluded, his control of the debates was masterly.

President of the Assembly

Mirabeau's problem was to know how and for how long his Machiavellian game could be continued before his intrigue with the court would be exposed. The people of Paris were restless, worried by rumours. Mirabeau's position was made difficult by his intervention on behalf of the King's aunts (who had fled from Paris), by his hostility to the law against the *émigrés*, and by his harsh words against the Lameths and their satellites in the Assembly ("Silence to the factious! Silence to the 33!"). On February 28 he was sorely pressed to justify himself to the Jacobins after a pitiless attack by Alexandre, comte de Lameth. The newspapers of the left redoubled their accusations of treason against him, and in March he experienced some notable reverses in the Assembly.

Death may have saved him from political defeat. Gravely ill since his presidency of the Assembly, he worsened his condition by excessive indulgence. He took to his bed on March 27 and died on April 2, 1791. The people's grief for him was boundless; he was given a magnificent funeral; and it was for him that the new church of Sainte-Geneviève was converted into the Panthéon, for the burial of great men. In the insurrection of August 10, 1792, however, papers proving Mirabeau's relations with the court were found in an iron chest in the Tuileries Palace; and on September 21, 1794, his remains were dislodged from the Panthéon by order of the Convention.

Assessment. As a statesman, Mirabeau failed in his main object, that of reconciling the monarchy with the Revolution and a strong executive with national liberty. He was too much of a monarchist for the Revolution, too revolutionary for the monarchy. As an orator he was unsurpassed. Even though his eloquence was fed by material gathered from every quarter and by a "workshop" of collaborators, it was Mirabeau who found the striking images and expressions that give to his speeches their brilliant individuality. Generally bad at extemporizing, Mirabeau could be moved by anger or by injured pride to an impassioned tone that would carry the Assembly with him.

BIBLIOGRAPHY. Still valuable are the 19th-century accounts of ETIENNE DUMONT, *Souvenirs sur Mirabeau* (1832, new ed. 1951; Eng. trans., *Recollections of Mirabeau*, 1832); M. LUCAS DE MONTIGNY, *Mémoires biographiques, littéraires et politiques de Mirabeau*, 8 vol. (1834–35; Eng. trans. of vol. 1–4, 1835), an uncritical work by Mirabeau's adopted son that should be read with LOUIS DE LOMENIE, *Les Mirabeau*, new ed., 5 vol. (1889–91); and ALFRED STERN, *Das Leben Mirabeaus*, 2 vol. (1889). Later accounts include JEAN-JACQUES CHEVALLIER, *Mirabeau, un grand destin manqué* (1947); O.J.G. WELCH, *Mirabeau: A Study of a Democratic Monarchist* (1951); ANTONINA VALLENTIN, *Mirabeau avant la révolution* and *Mirabeau dans la revolution* (1946–47; Eng. trans., 1948); RENE DE LA CROIX, *Mirabeau: ou, l'echec du destin* (1960); and J.J. CHEVALLIER, "The Failure of Mirabeau's Political Ideas," *Review of Politics*, 13:88–107 (1951).

(J.-J.C.)

Miracle

Miracles are extraordinary and astonishing happenings that are attributed to the presence and action of an ultimate or divine power. This article treats the nature, functions, and sources of miracles; it then attempts to discover the place of miracles in the religions of the world and the variety of interpretations given to them. It considers miracles mainly as seen, experienced, and interpreted in the religious context, irrespective of the scientific or philosophical judgment passed on them from the outside.

NATURE AND SIGNIFICANCE OF MIRACLES

Etymology and definition

A miracle is generally defined, according to the etymology of the word—it comes from the Greek *thaumasion* and the Latin *miraculum*—as that which causes wonder and astonishment, being extraordinary in itself and amazing or inexplicable by normal standards. Because that which is normal and usual is also considered as natural, miracles have occasionally been defined as supernatural events, but this definition presupposes a very specific conception of nature and natural laws and cannot, therefore, be generally applied. The significance of a miraculous event is frequently held to reside not in the event as such but in the reality to which it points (*e.g.*, the presence or activity of a divine power); thus, a miracle is also called a sign—from the Greek *sēmeion* (biblical Hebrew *ot*)—signifying and indicating something beyond itself. Extraordinary and astonishing occur-

rences become specifically religious phenomena when they express, reveal, or signify a religious reality, however defined.

Belief in miraculous happenings is a feature of practically all religions, and the incidence of miracles (*i.e.*, of belief in and reports regarding miracles) is universal, though their functions, nature, purpose, and explanations vary with the social and cultural—including theological and philosophical—context in which they appear. However inexplicable, all miracles have an explanation in the sense that they are accounted for in terms of the religious and cultural system that supports them and that, in turn, they are meant to support. Without such an accompanying—explicit or implicit—theory (*e.g.*, the presence, activity, and intervention of such realities as gods, spirits, or magical powers), there would be no miracles in the aforementioned sense but only unexplained phenomena.

TYPES AND FUNCTIONS OF MIRACLES

There is no general rule determining the types of occurrences that can be classified as miracles; they vary according to the cultural matrix of beliefs and assumptions. The mythological accounts of the origins of the gods and their activities in the primeval past, as well as accounts of the activities of other primeval beings, such as first ancestors and culture heroes, should, perhaps, not be classed as miracles, and the term is better reserved for outer, objective events—as distinct from such phenomena as inner experiences and visions—regarded as divine interventions or as manifestations of divine or supernatural powers. In many cultures, primitive as well as some more highly developed, such as the ancient classical and Oriental civilizations, the operation of extraordinary forces was taken for granted and was integrated into the total world picture and into the procedures and modes of action—*e.g.*, magic, oracles, divination, and shamanism—of ordinary life. Certain kinds of divine or spirit action and of cosmic operation were considered part of the normal order of things, even though it was generally admitted that priests and shamans would often resort to deception in their diverse activities, which included prophecy, oracles, healing, magic, and judgment by ordeal.

Purposes of miracles

Revelation and signification. The purpose of a miracle may be in the direct and immediate result of the event—*e.g.*, deliverance from imminent danger (thus, the passage of the children of Israel through the Red Sea in the Old Testament book of Exodus, chapter 14), cure of illness, or provision of plenty to the needy. Nevertheless, the ultimate purpose frequently is the demonstration of the power of the god or of the saint, the "man of God" through whom the god works, to whom the miracle is attributed. Thus, the crossing of the Red Sea by the Israelites is described not solely in terms of salvation from great danger but as a revelation of the saving presence of God and of the consequent obligation to serve and obey him; according to the account in Exodus: "and Israel saw the great work which the Lord did against the Egyptians, and the people feared the Lord; and they believed in the Lord and in his servant Moses." The purpose of a miraculous occurrence is thus often to reveal a divine reality or numinous dimension. The occurrence may be an event concerned with natural needs or situations, such as illness, hunger, or distress, or a specifically religious event that effects some form of salvation or revelation, such as the theophany on Mt. Sinai in which God gave to Moses the Ten Commandments, the Resurrection of Jesus Christ, or the revelation of the Qur'ān to Muḥammad. Even in these specifically religious events, the miraculous element is not necessarily of the essence but occurs as merely an accompanying circumstance designed to arrest the attention and to impress on everyone the unique character and significance of the occasion. Thus, theoretically at least, the theophany at Mt. Sinai could have taken place without thunder and lightning; Jesus need not have been born of a virgin; Muḥammad need not have made his miraculous journey to heaven. In actual fact, however, the very nature and quality of a religious event attracts miraculous elements, elaborations, and embellishments; and thus, for example, the founders of almost all religions are at the centre of great miracle cycles, and miracles occur as a rule in connection with persons and objects of religious significance, such as saints, sacraments, relics, holy images, and the like.

Authentication. In practice it is difficult to distinguish the revelatory or signifying miracles from miracles of authentication—*i.e.*, miraculous happenings that serve: (1) as credentials for claimants to religious authority in the form of leadership (*e.g.*, in Exodus, chapter 4, Moses convinces the Israelites of the authenticity of his mission by miraculous performances) or prophecy (*e.g.*, in Deuteronomy, chapter 18, it is said that a prophet is disqualified if the sign that he has predicted does not come to pass); (2) as demonstration of the superior power of a particular god (*e.g.*, in Exodus, chapter 7, Aaron's staff swallowed up the staffs of the Egyptian magicians, showing the superiority of the God of the Israelites); (3) as proof of the sanctity of a holy man, a holy site, or a holy object; or (4) more generally as evidence of the truth of a particular religion.

SOURCES OF MIRACLES

Spiritual sources. The source of miracles is always a divine, spiritual, supernatural, sacred, or numinous power that may be conceived in personal form (*e.g.*, God, gods, spirits) or impersonal form (*e.g.*, mana or magic).

Natural phenomena accompanied by special circumstances

The sacred may manifest itself in natural phenomena, such as thunderstorms or earthquakes, that evoke appropriate feelings of awe, but these are not usually considered miracles unless attended by special circumstances—*e.g.*, being predicted by a "man of God" or coinciding with an event of religious significance. As reported in the Gospel According to Matthew, chapter 27, at the moment of Jesus' death on the cross, "the curtain of the temple was torn in two, from top to bottom; and the earth shook, and the rocks were split; the tombs also were opened, and many bodies of the saints who had fallen asleep were raised, and coming out of the tombs after his resurrection they went into the holy city and appeared to many." The belief that thunder and lightning are manifestations of divine powers is very common, and many deities have been interpreted as personifying them or at least as being symbolized by them. Even in the Old Testament, thunderstorms and lightnings appear as manifestations or messengers of God. In this respect, the account of the theophany granted to the prophet Elijah in I Kings, chapter 19, marks a milestone in the history of religions, for "behold, the Lord passed by, and a great and strong wind rent the mountains, and broke in pieces the rocks before the Lord, but the Lord was not in the wind; and after the wind an earthquake, but the Lord was not in the earthquake; and after the earthquake a fire, but the Lord was not in the fire; and after the fire a still small voice" in which Elijah heard God.

In most cases theophanies and divine manifestations occur for a specific purpose: giving laws (*e.g.*, Moses and the theophany at Mt. Sinai; events in the lives of Numa Pompilius of Rome, Minos of Crete, and Lycurgus of Sparta, the ancient lawgivers in classical legend); saving interventions (*e.g.*, the voices resounding from the temple of Athena Pronaea in Delphi that caused the Persians to retreat); and the founding of cults (*e.g.*, the appearances of Mary, the mother of Jesus, at Lourdes, France, and Fatima, Portugal). Gods would appear to their devotees in visions and dreams, but these experiences should, perhaps, not be treated under the same general heading with other miracles. Immediate divine action was often perceived in omens preceding important undertakings, in apparently natural phenomena occurring providentially at critical moments or in miraculous—*i.e.*, sudden and seemingly impossible—cures. In most cases, however, such divine interventions took place through some form of mediation, human or inanimate: oracles are usually connected with certain sanctuaries, statues, or priestly functionaries; miraculous cures are effected through the intercession of particular saints, pilgrimages to certain holy sites, or rituals at certain sanctuaries (*e.g.*, at the temples of Asclepius, the Greek god of medicine, especially at Epidaurus).

Human and inanimate sources. Man can be the object of miracles, as when his disease is miraculously healed, or their subject, as when he performs miracles, such as healing others, in the name of whatever power is moving him. The two aspects cannot always be strictly distinguished, as is seen in the case of saints whose bodies are immune from corruption after death or founders of religions whose birth is attended by supernatural manifestations. Generally speaking, however, it is the role of holy personages—and of their tombs and relics—as sources of miracles that are of importance in the history of religions and more especially in the history of popular cults.

Subjects of legends

Founders of religions. The attitudes of the founders of the great religions to miracles vary considerably, but all have become the subject of legends of the most fantastic kind in popular belief, and much of this legendary material has been subsequently canonized in scripture and tradition. Moses is described as the man of God who performs miracles—occasionally with his miraculous staff—as the messenger and servant of the Lord; the fact that his place of burial is unknown rendered further miracles impossible. The ministry of Jesus was accompanied by many miracles and signs (*sēmeia* in Greek), and the sites and objects associated with his life and Passion have become centres of pilgrimages, objects of veneration, and sources of miracles. Muḥammad, though rejecting the demand to vindicate his ministry by signs, nevertheless became the subject of many miraculous accounts. The same applies to the Buddha, whose birth and passing away are said to have been accompanied by the most extraordinary phenomena: among other things, music resounded, and flowers descended from the heavens; jewel-laden trees sprouted, and perfumed rivers flowed from the Earth.

Holy persons. Much closer to the lives and devotion of ordinary folk than the superhuman figures of the founders are the saints, monks, ascetics, and diverse kinds of holy men and women. The attitude to saints and their miracles is very much the same on the popular levels of all religions, although the theoretical interpretations on the more theological level vary considerably. In Far Eastern religions it is often difficult to distinguish between saints and hero gods, because great men of renowned virtue can be deified and venerated and even receive officially approved state cults. Miracles occur as a matter of course at their tombs and relics. In Muslim as well as in Christian belief, the occurrence of miracles is part of the requirements for official recognition of sainthood and is interpreted as a special intervention by God, who thereby manifests his esteem for the saint or, more essentially, his salvific presence as realized concretely in the life and virtues of the saint. In Indian—Hindu and Buddhist—belief, miraculous powers are the "natural" result of ascetic practice and spiritual realization and can thus be considered as an almost natural manifestation of magical or spiritual causes. Among the many types of miracles ascribed to saints are not only those that are concerned with helping the needy or assisting persons in grave danger but also performances that demonstrate the powers or special status of the thaumaturgist, or miracle worker—*e.g.*, levitations and flying in the air. Thought or mind reading is another miraculous faculty ascribed to saints, but this is often made to serve a spiritual purpose—especially in Christian and Jewish legends—by being a means to turn sinners to repentance.

Relics, images, and amulets

Sacred objects. Because the life-span even of saints is limited, most of the miracles attributed to them occur through their inanimate remains at their tombs or through their relics. These relics may be parts of their bodies—often deliberately dismembered for wider distribution, so that a bone may be in one place, a hair in another, and the heart someplace else—or objects or parts of objects associated with their lives (*e.g.*, the shroud of Christ or fragments of the True Cross). The possession of relics was important for the reputation, prosperity, and safety of their place of preservation, and thus cities and churches would often compete for them or even steal them from each other (*e.g.*, there are legends concerning the transferral of the bodies of St. Mark to Venice or of St. Nicholas to Bari, Italy). Relics of the Buddha are preserved in several places (*e.g.*, the sacred tooth at Kandy in Sri Lanka) as are hairs of the Prophet Muḥammad. Very often these relics are simply objects of pious veneration rather than sources of constant miracles; nevertheless, popular belief and devotion frequently ascribe to them quasi-magical virtues.

Not all miracle-working objects of veneration are relics. Statues and icons can work miracles, and in many Christian countries images and icons of the Virgin Mary are especially famed for their miraculous virtues. In the Christian Middle Ages the veneration of the sacrament of the Eucharist brought about a proliferation of miracles. Here, as in the case of images, a distinction can be made between the magical character of folk beliefs and the diverse theological doctrines concerning these religious objects; only rarely have religious authorities opposed the cult of saints, images, and relics and the concomitant belief in miracles—an exception is classical Protestantism, which in the 16th century rejected such cults.

Although they are not strictly sources of miracles, talismans and amulets—*i.e.*, objects believed to possess magical virtues as good-luck charms or as protecting the bearer or owner from all kinds of danger—should be mentioned in this connection. They are found in diverse forms and sizes and in all kinds of material: as natural objects or artifacts; as relics; or as pieces of paper, parchment, or metal inscribed with verses from Holy Scriptures, religious invocations, or magic formulas. In many primitive religions such objects form part of the recognized cultic paraphernalia—*e.g.*, African fetishes and the medicine bundles of many American Indian tribes. In some religions what seems to be objectionable magic practice may be integrated with the normative religious system by the blessing of medals and similar objects.

Sacred places. Miracles are often connected with special sacred places. Normally these are natural shrines, such as sacred groves, or temples and sanctuaries in which a god or spirit lives or has manifested himself or in which his statue, his symbol, his holy objects, or his relics are enshrined. Holy places, such as Mecca and the Kaʿbah in Islām or the Buddhist stupas, are centres of pilgrimages and veneration because of their religious significance and the religious values that they symbolize and not necessarily because miracles are wrought there; yet, popular devotion associates miracles with many of these holy sites.

MIRACLES IN THE RELIGIONS OF THE WORLD

It has already been suggested that the mythologies of primitive and ancient religions should not be designated miraculous insofar as they deal with mythical origins and ages; frequently they attempt to explain how certain regularities and what is now considered the normal course of things have come into being. The crucial distinction lies between religion on the popular primitive level and the more highly developed forms of religious belief. The tendency of the former is to relate to a concrete, magical presence of the sacred and to envisage the possibility of using this presence for the achievement of such desired ends as healing, blessing, or success in an undertaking. The higher forms of religion—though recognizing miracles or even demanding dogmatic affirmation of belief in them—exhibit a far more differentiated and complex attitude.

Hellenistic religion

Hellenistic religion presents one of the best examples of a civilization in which miracles play a major part. The intervention of the gods in the affairs of the Homeric heroes takes place in a cosmos in which the divine and human spheres still interact. Later Hellenistic syncretism conceived of the sublunar world as a distinct sphere, though higher powers could miraculously irrupt into it. Miraculous cures (*e.g.*, at the sanctuary of Asclepius at Epidaurus), divine manifestations of various kinds (*e.g.*, voices, dreams, and theophanies), and even virgin births and resurrections were widely reported. Some statues of gods were said to have descended from above and to move when they wished to communicate an oracle. Miraculous cures, resurrections of the dead, and other ex-

traordinary feats were attributed to great thaumaturgists, among whom Simon Magus, mentioned in the New Testament Acts of the Apostles, and especially Apollonius of Tyana, a Neo-Pythagorean sage of the 1st century AD, are the best known. In the world of Hellenistic syncretism, belief in miracles and in magic merge and can hardly be distinguished.

Religions of the East. In the great religions of the East the belief in miracles is closely connected with the theory that ascetic practices and the knowledge of mystical formulas, such as the Sanskrit *mantras*, can give the practitioner unlimited magical powers.

Religions of India. India has become the classic land of wonders not because of the accounts of fantastic actions of divine beings or semidivine heroes and avatars (incarnations of Hindu gods) related in Indian mythology but because both popular religion and philosophical theory set no bounds to the magical powers that can be attained by great ascetics and yogins (adherents of Yoga, the Hindu philosophy teaching the suppression of all activity of mind, body, and will in order that the self may realize its distinction from them and attain liberation). Even if these magical powers are considered insignificant in higher religion and spiritually negligible, their reality is never doubted. The *Upaniṣad* and the *Brāhmaṇa*—ancient Sanskrit writings of the Vedic period—may consider the heights of religious insight and mystical experience as man's supreme aim, but neither the later classical sources nor contemporary Hindu belief ever question the miraculous powers of a holy man. The same attitude is shared by the other religions of Indian origin: Jainism and Buddhism. According to the earliest Buddhist accounts, Gautama Buddha deprecated the miraculous powers that he possessed as being devoid of spiritual significance; nevertheless, miraculous accounts of the circumstances attending his conception, birth, life, and death proliferated as time went on. Miracles also occurred in connection with his relics, which became centres of veneration. Miraculous and magical elements are especially prominent in the *Mahāyāna*, or Greater Vehicle, forms of Buddhism.

The Buddha

The Buddha himself refused to spread his teaching by impressing his audience with miracles. According to the Aṅguttara Nikāya, one of the collections of the Buddha's sayings, there are three kinds of miracles—the miracle of magic, the miracle of thought reading, and the miracle of instruction—and of these the last is the most wonderful and excellent, whereas the other two are not much better than a conjuror's tricks. Yet the same text also describes what is implied by the miracle of magic: "there is one who, . . . having been one becomes many, . . . appears and vanishes, unhindered he goes through walls. . . . He dives in and out of the earth as if it were water. Without sinking he walks on water as if on earth. Seated cross-legged he travels through the sky like a winged bird. With his hand he touches and strokes the sun and the moon. . . ." The same text also asserts that not only was Gautama endowed with these powers but so also were hundreds of monks of his order. The tendency to denigrate the spiritual value of miraculous powers is illustrated by the story of the Buddha's meeting, at the bank of a river, with an ascetic who had practiced austerities for 25 years and who claimed that he was at last capable of crossing the river by walking on the water. The Buddha's reaction was to express regret at the waste of so much time and effort, because for "one penny" the ferry would take him across the river.

Religions of China. In China, although Confucianism in the strict sense has little room for miraculous elements, Taoism has produced a rich crop of thaumaturgy and magic on all levels of folk religion. No doubt the teaching of the Tao (literally, the Way) can be interpreted in terms of a sublime moral and perhaps even mystical doctrine. In actual fact it was one of the main sources of Chinese magic in all its forms, including the quest for the elixir of life. Religious Taoism, with its theory of a balance and interaction of cosmic forces, lent itself to elaboration and expression on all levels—from philosophy to pseudo-science to magic.

Religions of the West. In Western monotheistic religions it is necessary to distinguish between the role of miracles on the level of popular beliefs and practices and the theory of miracles propounded by the theologians. Belief in a personal, omnipotent Creator who exercises his providence over his creatures implies a concept of miracle as a deliberate intervention in the course of events by the same sovereign God who also assures their normal regularity.

Judaism. Miracles are taken for granted throughout the Old Testament. God does "wondrous things" according to Psalms, chapter 72, and "great things and unsearchable, marvellous things without number" according to the Book of Job, chapter 5; these things are done in his creation in general and in the history of his people in particular (*e.g.*, the ten plagues of Egypt and the events of the Exodus). A list of the great wonders done by God is given in Psalms, chapter 136; their purpose is to make his creatures praise him, acknowledge his rule, and "know that I am the Lord." God's wondrous deeds range from the normal regularities of creation to extraordinary interventions that run counter to ordinary experience and thus serve as signs of his greatness and providence in wreaking vengeance on the wicked and giving salvation to his elect. Very few miracles—apart from prophetic experiences and visions—are told of the great literary prophets; the cycles of legends the heroes of which are the early prophets Elijah and Elisha, however, are closer to the traditions of folk religion.

The Old Testament and rabbinic Judaism

Later rabbinic Judaism took the occurrence of miracles for granted. It assumed a natural order in which things worked and within which man was supposed to discharge his duties; thus, to rely on miracles was nothing short of sinful. In special circumstances, however, or in connection with persons of extraordinary saintliness, God would intervene or spectacularly answer their petitionary prayers. It was not so much a matter of suspending as of relativizing nature, the normal course of which was just one possible expression of the divine will; *e.g.*, in the legend of the lamp into which vinegar had been poured by mistake instead of oil: "He who has commanded the oil to burn will also command the vinegar to burn—and the lamp burned for a whole day." It was only in the Middle Ages and under the influence of Greco-Arabic philosophy that the problem of miracles was systematically discussed on a philosophical and theological level. Normative, rabbinic Judaism, being mainly concerned with doing God's will as revealed in his Law, had little interest in miracles, though it accepted, as a matter of course, the veracity of the miracles recorded in scripture and in the Talmud (the collection of Jewish lore, legend, and law). On the level of popular piety both magic and the belief in miracles always flourished, especially under the influence of Kabbala, the esoteric, mystical movement within Judaism; the Ḥasidic movement (a pietist movement that arose in eastern Europe in the 18th century) in particular produced a rich crop of beliefs and legends concerning the miraculous virtue—through prayer, intercession, or magical power—of the great Ḥasidic saints and rabbis.

Christianity. New Testament accounts of the advent, birth, life, Passion, and Resurrection of Christ include many miracles. Jesus is reported in the Gospels to have performed miracles of diverse kinds: raising the dead, healing the sick, casting out demons, and causing nature miracles, such as the multiplication of loaves and the turning of water into wine at the town of Cana. New Testament scholars are still conducting detailed research as to which accounts are early and which developed later, which grew on original Jewish soil and which evolved under the influence of the Hellenistic genre of miracle legends. There seems little doubt, however, that Jesus impressed many of his contemporaries more by his power to work miracles than by his preaching. Unlike the Buddha and Muḥammad, Jesus had an ambiguous attitude toward miracles: on the one hand he performed them as a sign of his mission and of the impending coming of the Kingdom; on the other hand he reproved the desire for wonders and repeatedly forbade the disciples to publicize his miracles, insisting that it was faith alone that worked

Jesus Christ

miracles. In fact, because miracles could also be explained by attributing them to demonic agency, it was ultimately faith that determined the quality and function of the miracle.

Early Christianity developed in the atmosphere of Hellenistic, Greco-Roman culture, which was full of miraculous accounts and legends. These no doubt influenced Christian traditions and forms of devotion, especially as popular religion always hankered after miracles, and—at the conclusion of the Gospel According to Mark—Christ himself had promised the continuance of miracles in his church. In a world in which only a few critical minds doubted the reality of miracles, the similarity of the Christian signs to those reported in pagan legend was attributed to demonic imitation and counterfeit; St. Augustine, the great theologian of the early Western Church, held that "the miracles of the [pagan] gods are vanquished by those of the [Christian] martyrs." The problem of distinguishing between the two sources of miracles—because the devil often disguises himself as an angel of light—frequently solicited the attention of theologians and mystics. Whereas for the theologians a miracle was a sign of God's saving presence and design, for the mass of believers it was the manifestation of a sacred power inherent in individual persons, places, and objects. Spectacular miracles were reported occurring in connection with the sacraments of Baptism and the Eucharist, with the tombs of martyrs and saints, with relics and holy images; certain saints acquired a special reputation as thaumaturgists.

Medieval theologians—and specifically St. Thomas Aquinas—taught that as all knowledge was derived from sensible facts, so also "a certain degree of supernatural knowledge of the objects of faith" could be brought about "by certain supernatural effects that are called miracles." This doctrine already assumes a system of natural causality that God—though he normally works through the natural law of which he is the author and Creator—can temporarily set aside. It also assumes that—at least in theory, if not always in practice—natural and supernatural effects can be distinguished. Thus, in 1870 the first Vatican Council declared: "If anyone should say that no miracles can be performed, . . . or that they can never be known with certainty, or that by them the divine origin of the Christian religion cannot be rightly proved—let him be anathema." (Anathema is a solemn ban or curse pronounced by ecclesiastical authority.) Belief in miracles is thus obligatory in the Roman Catholic Church, although belief in any specific miracle is not necessarily so. In fact, miracles form part of the career, in life as well as after death, of Christian saints and their occurrence continues; proof of miracles is among the requirements for the process of canonization, in which a person is formally proclaimed a saint. Classical Protestantism, however, has confined its belief in miracles to those recorded in scripture.

Islām. Muslim religion assumes, as a matter of course, that Allāh works miracles and has done so in the past; *e.g.*, through Moses, Solomon, and Jesus but significantly not through the Prophet Muḥammad. According to the Qur'ān, Muḥammad explicitly rejected the idea of proving his vocation by signs and miracles: the Qur'ān itself was the greatest miracle, and he was but a human messenger and preacher of repentance. Nevertheless, subsequent narratives invested his birth and life with superlatively miraculous details.

The Prophet Muḥammad

Muslim popular religion, particularly under Ṣūfī (Islāmic mysticism) influence, abounds in miracles, pilgrimages to the tombs of wonder-working saints, and the like. Dogmatic theology, too, recognizes miracles as facts. The peculiar feature of Muslim theology is that, unlike Christian theology, it did not accept the idea of nature as an entity operating according to fixed laws ordained by the Creator. Because the universe is constantly being re-created by Allāh in successive time atoms, natural regularity is nothing but the regularity of Allāh's habit in re-creating the universe. Thus, a miracle is the omnipotent God's departure from his habit but no different, in principle, from the latter. Muslim dogmatics distinguishes between miracles (*karāmāt*), with which Allāh surrounds his saints (*awliyā'*) as a mark of distinction, and signs (*āyāt*, also *mu'jizāt;* literally, "acts of an overwhelming nature"). The latter are wrought by Allāh to prove the genuineness of his messengers and to overwhelm and reduce to silence their opponents. Such miracles, which deviate from the usual course of things and are of such nature that others cannot produce their like, are Allāh's testimony to the sincerity of his apostles. The problem is nevertheless complicated by the fact that Satan too can perform miracles. Generally speaking, miracles do not play a role in the continued life of orthodox Islām, though they loom large in popular belief and piety, especially in the cult of saints who—in life as well as after death at their tombs—have the power to impart *barakah* ("blessing"). Orthodox zealots, such as the Wahhābīs, have not hesitated on occasion to destroy the tombs of saints because they considered their cult as incompatible with true Islām.

INTERPRETATION OF MIRACLES

All of the more fully developed theologies have formulated a doctrine of miracles in the context of their beliefs regarding God, the world, the operations of nature, and causality. The emergence of the concept of nature as a closed system functioning in accordance with strict causal laws created problems more than once, but medieval Christian and Jewish thought had no difficulty in maintaining that the order created by God could also be suspended by him.

In classical antiquity. Miracles were denied even in classical antiquity. Thus, Cicero asserted that "nothing happens without a cause, and nothing happens unless it can happen. When that which can happen does in fact happen, it cannot be considered a miracle. Hence, there are no miracles." Cicero qualified this statement, however, by saying that miracle stories may be necessary for the piety of ignorant folk. The 2nd-century pagan philosopher Celsus is less dogmatic in his attacks on Christianity: the Christian miracles are insufficiently attested and most improbable, but, even if they were genuine, they could hardly offset the miracles of the pagan world—*e.g.*, the healings of Asclepius. This was the standard pattern of many religious polemics: miracles as such were not necessarily denied; only those claimed by the adversary were denied. When these could not be denied, they were ascribed to diabolic agency or to the fraudulent practices of priests or occasionally to a misinterpretation of essentially natural phenomena.

In the 18th and early 19th centuries. Rationalist criticism, although not completely absent in the Middle Ages, became a major factor in the 18th and 19th centuries. David Hume, a British Empiricist and a Skeptic, in the chapter "On Miracles" in his *Enquiry Concerning Human Understanding* argued that, given the general experience of the uniformity of nature, miracles were highly improbable and that the evidence in their favour was far from convincing. It should be emphasized that Hume, whose criticism led him to a denial of causality, did not dismiss miracles because they were inconsistent with causal law—as many other thinkers did, notably the Deists (those, especially British, who advocated a natural religion). Instead Hume insisted on the probability factor and thus on the importance of assessing historical evidence. Because all Christians agreed that biblical religion and the scheme of salvation set forth in it could be maintained only by stressing prophecy and miracle, there developed a vast body of literature, especially among Protestants, proving the authenticity of the Christian faith on the basis of the miracles recorded in the Bible. For many 18th-century thinkers, however, the vastness and complexity of the order of nature were even more impressive than any alleged exceptions to it. Thus, belief in miracles, although remaining an essential element of faith to good Christians, appeared as sheer superstition to the eyes of the torchbearers of Rationalist enlightenment. Deism and subsequently liberal Protestantism, each in its own way, defined religion or reinterpreted biblical religion in ways that dispensed with miracles. Sometimes the facts alleged

Rationalist criticisms

in the Bible were partly admitted, but a natural explanation was substituted for the miraculous one.

Criticism of the concept of miracle was articulated in more than one way. There was philosophical and scientific criticism to the effect that miracles were impossible and that even epistemologically (*i.e.*, within the limits of knowledge) the occurrence of a miracle could never be established; at most, these critics maintained, there were merely as yet unexplained natural phenomena. (This view comes close to the religious assertion that faith precedes the experience of a miracle and that the factuality of a miracle can never precede faith.) There was historical and philological criticism, arguing that the actual occurrence of miracles is unsubstantiated and analyzing the growth and evolution of the legends and texts reporting miracles. There was psychological criticism, arguing that some people want to believe in miracles and so produce imaginative creations answering their psychological needs. There was also a type of religious criticism implying that the truly spiritual has no need of miraculous supports. One suggested solution to the problem was the assertion that the term miracle does not describe an objective event but a subjective mode of experience: the more religious one is, the more one sees miracles everywhere. This view of Friedrich Schleiermacher, an early-19th-century Protestant theologian and philosopher, identified miracle with a religious understanding of any aspect of the world.

Liberal and Neo-orthodox Protestantism

In the late 19th and 20th centuries. Later 19th- and 20th-century liberal Protestant thinkers, such as Rudolf Bultmann, a German New Testament scholar, discarded the traditional notion of miracle together with other elements of what they termed the mythological apparatus of the Bible. Many of these liberal theologians sought the evidence for Christianity in the moral and religious transformation it brought to people's lives or interpreted the doctrine of salvation in Existential terms. The early decades of the 20th century, however, also witnessed a return to a more orthodox theological climate—as, for example, in the thought of Karl Barth, a Swiss Protestant theologian—and a new readiness to accept miracles as meaningful signs of God's salvific activity. This change of climate coincided with certain developments in science that appeared to question a too rigid and mechanical concept of causal determinism. It was argued that the discoveries of physics on the atomic level and particularly of Werner Heisenberg's Principle of Uncertainty (*i.e.*, that it is impossible to specify or determine simultaneously both the position and velocity of a particle with full accuracy) allowed for divine action in the form of noncausal miracles, though this argument was also judged by some as a misinterpretation both of science and of the religious meaning of miracle. The problem was complicated for modern theological apologetics by the compulsion of many people to believe in miracles—although their actual philosophical, intellectual, and even religious sensibility inclines them to dispense with this belief—simply because miracles are recounted in their hallowed scriptures and traditions; not everyone is prepared to abandon or "explain away" parts of his hallowed tradition.

Orthodox Jews, Christians, and Muslims still believe in the literal occurrence of the miracles recorded in their scriptures and traditions; Catholics, furthermore, believe in the continued occurrence of miracles, defining them as a direct divine effect upon nature. The liberal attitude—whatever the variations in detail and in sophistication of the explanation—is essentially similar to that propounded by Schleiermacher.

Miracles hold a prominent place in the history of religions; the encounter with the reality posited by religion (God, the gods, transcendence, the holy, sacred, or numinous) is essentially wonderful and transcending normal reality. Thus, religious experience tends to crystallize in miraculous accounts. The religious pattern of experience easily combines with the primitive yearning after the miraculous, born of the human desire to master ordinary needs, uncertainities, fears, and sufferings with a magical short-cut to greater powers. Whereas some theological systems assert that miracles both produce and confirm faith, many students of religion agree with Goethe's description of the miracle as "the dearest child of faith."

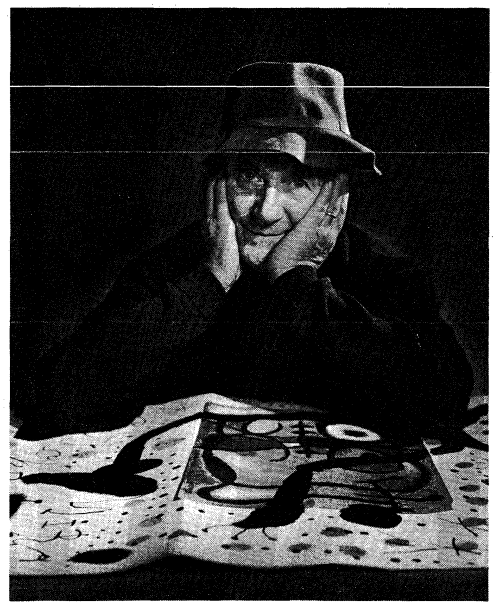
BIBLIOGRAPHY. GUSTAV MENSCHING, *Das Wunder im Glauben und Aberglauben der Völker* (1957), the best and most complete treatment of the subject, with good bibliography; RICHARD REITZENSTEIN, *Hellenistische Wundererzählungen* (1906) and OTTO WEINREICH, *Antike Heilungswunder* (1909), treatments of miracle cures and the literary genre of miraculous accounts in late classical antiquity; ROBERT M. GRANT, *Miracle and Natural Law in Graeco-Roman and Early Christian Thought* (1952), a description of Hellenistic attitudes and beliefs about miracles as the context in which the concept evolved in the early church; C.S. LEWIS, *Miracles* (1947; paperback ed., 1963), an example of a modern Christian apologetic; GUY DE BROGLIE, *Les Signes de crédibilité de la Révélation chrétienne* (1964; Eng. trans., *Revelation and Reason*, 1965), see pt. 3 for a contemporary statement of Roman Catholic doctrine.

(R.J.Z.W.)

Miró, Joan

A Spanish artist whose paintings, sculptures, and drawings are continually surprising in their playful simplicity, their wit, and their richness of imagination, Joan Miró creates works that haunt the spectator like a dream. Miró's ultimate artistic style was the outcome of many stages of development that were, from the outset, directed toward discovering in the hackneyed appearance of the visible world the pristine condition of myth. Through all the phases of modern art through which he has lived, he has never become outmoded. The latest trends in art and anti-art always seemed to have been anticipated in his works. Miró's work disowns all tradition and penetrates to the origins of art.

© Karsh—Rapho Guillumette



Miró, photograph by Yousuf Karsh, 1966.

Youth in Barcelona

Joan Miró was born on April 20, 1893, in the old town of Barcelona. His father was a watchmaker and goldsmith, and his forefathers had also been artisans from the Tarragona region of Spain and the island of Majorca. Both the artisan tradition and the austere Catalan landscape were to prove of the greatest importance to his art. From his father he learned a love of astronomy and a sense of the infinity of the cosmos.

Miró's artistic development did not progress with the triumphant directness of his countryman Picasso, who could draw like a master while still a boy. Instead of being allowed to go to an art school, Miró was expected to complete high school, though he failed to do so. He then attended a commercial college and worked for two years as a clerk in an office, until he had a mental and

physical breakdown. His parents took him for convalescence to an estate, Montroig, near Tarragona, which they bought especially for this purpose, and finally allowed him to attend an art school in Barcelona. His teacher at this school, Francisco Galí, showed great understanding for his 18-year-old pupil, advising him first to touch the objects he was about to draw, a procedure that strengthened Miró's feeling for the spatial quality of objects. Galí introduced his pupil to examples of the latest schools of modern art from Paris and showed him the churches and chapels of the region, with their Byzantine-influenced frescoes. Finally, he introduced him to the fantastic world of the buildings of Antonio Gaudí, Barcelona's famous Art Nouveau architect.

Miró's development as an artist differed from that of other modern artists: whereas artists of the contemporary Fauve and Cubist schools deliberately attempted to destroy the canons of tradition in order to attain a new kind of pristine vision, Miró possessed such a vision naturally. In his paintings and drawings he sought above all to establish means of metaphorical expression—that is, to discover signs that stand for concepts of nature in a transcendent, poetic sense. He wanted to depict nature as it would be depicted by a primitive man or a child equipped with the intelligence of a 20th-century adult; in this respect, he had much in common with the Surrealists and Dadaists, two other schools of modern artists, who were striving to achieve similar aims by more intellectual means than Miró used. Miró's art developed slowly from his first clumsy attempts at expression to the apparently playful masterpieces of his later period. His fanatical honesty and his conscientious craftsmanship compelled him to work on many of his pictures for years.

From 1915 to 1919 Miró worked in Barcelona, at Montroig, and on Majorca, painting landscapes, portraits, and nudes in which his interest centres on the rhythmic interplay of volumes and areas of colour. His colours were still dark and heavy, though he delineated details with a frenzied concentration as if superimposing on a heavy earthly ground a filigree of luminous leaves and blossoms. His manner was the same in landscapes, portraits, and nudes.

Miró was one of many artists who made their way from abroad to Paris during the first two decades of the 20th century and enriched French painting, which was to influence the art of the whole world. Most of these foreign artists elected to become Frenchmen after coming into contact with the French artistic metropolis, but Miró remained attached to his Catalan homeland, in his choice both of dwelling places and of subjects for his pictures.

From 1919 onward Miró lived alternately in Spain and Paris. In the works he produced in the period between World Wars I and II—the great still lifes, landscapes, and phantasmagorias set free from both space and time—he gradually removed the objects he portrayed from their natural context and reassembled them as if in accordance with a new, mysterious grammar, creating a ghostly, eerie impression.

Dream pictures and imaginary landscapes

From 1925 to 1928, under the influence of the Dadaists, Surrealists, and Paul Klee, he painted "dream pictures" and "imaginary landscapes" in which the linear configurations and patches of colour look almost as though they were set down randomly. After a trip to The Netherlands, where he studied the 17th-century Dutch realist painters in the museums, the figurative elements in his pictures once more assumed a firmer shape. But, when a tendency toward beautiful, tasteful forms emerged in his works, he countered it with more brutal signs, collages, and objects made up of the waste products of industry.

By the 1930s, Miró's artistic horizons were expanding. He designed decor for ballets. His paintings began to be exhibited regularly in French and U.S. galleries. In 1934 he designed tapestries, and that led to an interest in the monumental and in murals.

At the time of the Spanish political turmoil and Civil War in the late 1930s, Miró was living in Paris. In their demonic expressiveness, his pictures of this period mirrored the fears and horrors of those years. At the Paris World Exhibition of 1937, he painted for the pavilion of the Spanish Republic a mural, "The Reaper," containing a strong element of social criticism.

During World War II, Miró returned to Spain, where he painted his "Constellations," symbols of the happy collaboration of everything creative, of the elements and the cosmos. They represent a challenge to the anonymous powers of corruption in social and political life, the cause of misery and wars. During the last year of the war, Miró together with his potter friend Artigas, produced ceramics that revealed a new impetuosity of expression.

From 1948 onward he once more divided his time between Spain and Paris. That year saw the start of the series of very poetic works the symbols of which were based on the theme of woman, bird, and star. Pictures wildly spontaneous in character came into being alongside others whose forms were executed with punctilious craftsmanship. Both approaches were also combined in Miró's sculptures; in them all his earlier figurations were happily amalgamated to form erotic fetishes or signals towering into space.

Postwar fame

In the years following World War II, Miró became world famous; his sculptures, drawings, and paintings, as well as books he had illustrated, were exhibited in many countries. In 1950 he painted a wall for Harvard University. His ceramic experiments were crowned by the great ceramic wall in the UNESCO building in Paris (1958), for which he received the Great International Prize of the Solomon R. Guggenheim Foundation. In 1962 Paris honoured him with the first major exhibition of his collected works in the Musée National d'Art Moderne. The architect José Luis Sert built for him on Majorca the large studio of which he had dreamed all his life.

In spite of his fame, however, Miró continued to devote himself exclusively to looking and creating. A taciturn, introverted man of short stature, Miró had not found it easy to attain the wisdom permeated by irony that ultimately characterized his work. From his youth on, he had felt compelled by an almost anarchic obstinacy to keep his eyes firmly fixed on his goal despite the resistance offered by society or by the prevailing artistic theories. His late works manifest an even greater simplification of figure and background than his early ones. In order to realize his inner visions, it was enough now for Miró to set down a dot and a sensitive line on a sea-blue surface; in this the spectator could still find himself in a state of total enchantment. His early work anticipates the representational techniques of his later art style; thus, although the former playful or aggressive irony gives way to a quasi-religious meditation, all of his works form a coherent whole manifested by a common quality of rejuvenation and deepening.

MAJOR WORKS

"Self-Portrait" (1919; Pablo Picasso Collection, France); "The Farm" (1921–22; Ernest Hemingway Collection); "Carnival of Harlequin" (1924–25; Albright Knox Art Gallery, Buffalo, New York); "Dog Barking at the Moon" (1926; Philadelphia Museum of Art); "Dutch Interior" (1928; Museum of Modern Art, New York); "Woman in the Night" (1940; Mme Duchamp Collection, New York); "The Red Disc in Pursuit of the Lark" (1953; Galerie Maeght, Paris); "Blue II" (1961; Galerie Maeght, Paris); "Personage and Bird" (1965; private collection); "The Ski Lesson" (1966; Galerie Maeght, Paris); "Red Accent in the Quiet" (1968; private collection); Mural for the "Labyrinthe" of the Fondation Maeght, Saint-Paul, France (1968).

BIBLIOGRAPHY. J.J. SWEENEY, *Joan Miró* (1941), the first comprehensive presentation of Miró's life and work, written for a retrospective exhibition at the Museum of Modern Art in New York; WALTER ERBEN, *Joan Miró* (Eng. trans. 1959), a monograph that reproduces many different examples of the artist's work at his home and studio on Mallorca; JACQUES DUPIN, *Miró* (1961; Eng. trans., 1962), the standard work, written by an art historian and poet who has been a friend of the painter for many years—contains a detailed bibliography of writings on Miró as well as a catalog of his pictorial and sculptural works; MICHEL TAPIE, *Joan Miró* (1970), French text with many illustrations, especially of his later works, and an enlarged bibliography up to 1968.

(W.Er.)

Mississippi

General character of the state

Originally part of America's Old Southwest, Mississippi became the 20th state of the Union in 1817. Its name has long been symbolic of many of the characteristics attributed, correctly or incorrectly, to the Deep South. As it entered the last quarter of the 20th century, the state was engaged in efforts to alter economic and social patterns of the past and present. Through the inevitable confrontation of the forces of tradition and change, Mississippians had become conscious of their image as portrayed on the national scene.

For decades, an unusually large dependent population, above or below employment age, as well as a predominantly agricultural economy, helped to keep per capita income low and to create inadequate levels of living for many families. Though more than half of the 2,216,912 Mississippians counted in the 1970 census were still rural —but not necessarily on farms—the state ranked low in many economic indexes. During the 1960s, the industrial income surpassed the agricultural income for the first time.

Throughout most of its 47,716 square miles (123,584 square kilometres)—from Tennessee in the north to Louisiana and the Gulf of Mexico in the south, from Alabama on the east to Louisiana and Arkansas on the west —much of Mississippi's soil is rich and deep, and its low-lying landscape is laced with many rivers. Almost inevitably it became an agricultural state. The long dominance of a rural, unhurried way of life has contributed much to the problems of present-day Mississippi. This way of life has also left a sense of history among some Mississippians, whose ancestors created a culture of gentility that is still evident in the many historic mansions located in such old towns as Columbus, Biloxi, Natchez, and Holly Springs. (For information on related topics, see the articles UNITED STATES, HISTORY OF THE; CIVIL WAR, U.S.; MISSISSIPPI RIVER; MEXICO, GULF OF; and NORTH AMERICA.)

THE HISTORY OF MISSISSIPPI

Pre-Civil War days. The original Mississippi Territory created by Congress in 1798 was a strip extending about 100 miles north to south and from the Mississippi River to the Chattahoochee, on the Georgia border. The territory was increased in 1804 and 1812 to reach from Tennessee to the Gulf. In 1817 the western part achieved statehood as Mississippi (the eastern part became the state of Alabama in 1819). Natchez, the first capital, was replaced by Jackson in 1822.

The 1820s and 1830s were marked by the decline of the Jeffersonian Republicans, the ascendancy of the Jacksonian Democrats, and the removal of the Indians to Oklahoma.Those were the days of the steamboats, land speculation, and the growth of a plantation-based cotton economy, with its concomitant slave population. Slave owning, however, was not common among the small landowners, who became more numerous than the large planters but had little influence on public affairs for many years.

Sectionalism in both North and South had been growing for some time. Its ill feelings, though fed by many issues, gradually became dominated in both North and South by slavery. In January 1861, a convention adopted an ordinance of secession, and within a year the state was in the midst of war. The people suffered much privation, and the land, great devastation; by 1865 the state was in economic ruin.

Aftermath of war. Many historians consider that the Reconstruction period that followed was as bad as the war. Political control passed to a group comprising Radical Republican carpetbaggers, black freedmen, and scalawag whites. The excesses and abuses of Reconstruction ended by 1876, but the seeds of new problems had been planted. Controversy over the sharecropping system, ideas of "white supremacy," and economic policies would plague the state for 100 years.

Mississippians hoped to find economic salvation in the coming of industry and the railroads, but the hope was only partially realized. Emancipation had made the former slave free to go where he would, but most remained and eventually were absorbed into the tenant-farming system. The continued economic interdependence of the two races kept many of the customs and social systems intact. The constitution of 1890 effectively disfranchised most of the Negro population.

Hardening of rural dominance and social traditions

World War I hastened the end of Mississippi's physical and psychological isolation, and most of the bitterness remaining from the Civil War was lost in a surge of patriotism. Between the wars, the state was affected by the agricultural depression of the 1920s, the devastating 1927 flood, the Great Depression of the 1930s, the coming of farm-production controls, and the beginnings of new industrialization. After World War II, government farm programs and mechanization on a broad scale created another agricultural revolution.

Civil rights movement. After the war, a series of events developed that may be characterized as a revolution in race relations. Many long-accepted practices, customary throughout much of the South, received their first major jolt in 1954 with the United States Supreme Court decision declaring racially segregated schools to be unconstitutional. That was followed by years of increasing protest against other aspects of segregation and by large-scale registration of black voters. Within the black community, leaders emerged who sometimes were aided by whites, often from outside the state.

Tensions between tradition and change

The white Mississippians felt the entire fabric of their lives threatened by demands for massive and immediate change. They reacted strongly to what they believed was the duress imposed upon them by outside governmental forces and private reform groups alike. Pressures, publicity, and recrimination were accompanied by bitter words, confrontations, name-calling, sit-ins and boycotts by blacks, and even acts of violence. Among these were the murder of civil rights workers and the bombing of black churches. These acts of militance by both sides made international headlines. Yet the slow process of historical change saw the people of the state adjusting to many new laws and to new interpretations of long-standing laws. By the early 1970s a much changed, more tranquil atmosphere existed. Many observers found Mississippi's schools to be the most integrated in the nation, and blacks had been elected to a number of local offices.

In politics, disaffection developed between the state and national Democratic parties. Traditionally Democratic, Mississippians broke the trend and voted for other presidential candidates who appeared to be more conservative or more favourably disposed toward the region. Some feeling existed that increasing urbanization of the state, economic changes, and the growing role of the blacks might move the state toward a more conventional two-party system.

THE NATURAL AND HUMAN LANDSCAPE

Mississippi is a low-lying state, its highest point (in the northeast) reaching only 806 feet (246 metres) above sea level. Its major soil areas encompass hills, plains, prairie, river lowlands, and pine woods. The traditional regions relate generally to physiographic regions. Discernible differences exist, however, in life-styles, political persuasions, economic characteristics, and general cultural variations.

The Delta

Physiographic regions. In the northwest, the great fertile crescent called the Delta is the old floodplain of the Yazoo and Mississippi rivers, comprising 4,000,000 acres of black alluvial soil many feet in depth. Once subject to disastrous floods, the land is now protected by levee and reservoir systems. Though the Delta was only sparsely settled in antebellum days, it is today a region of highly mechanized farming.

East of the Delta looms the high wall of loess bluffs, marking the beginning of the highlands, or hills. One first encounters a brown loam belt of varying width extending from Tennessee to Louisiana. Most of southern Mississippi, from the brown loam belt to Alabama, lies in the high and rolling Piney Woods. Though settled early, the area did not prosper until the early 20th century, when the great virgin pine forests were exploited

and heavily decimated. Today it is a prosperous area based on diverse forest industries, cattle, and some specialty farming.

The coastal area, sometimes called the Coastal Meadows, or Terrace, borders the Gulf of Mexico. The soil is sandy and not well suited to crops, but its location, climate, and industrialization have made the region important.

Along the northern edge of the Piney Woods lies the narrow Central Prairie, separated from the Black Prairie by a section of hills and woods. The two prairies, with fertile black soil that is excellent for many types of agriculture, were once the site of cotton plantations. East of the Black Prairie, in the extreme northeast, are the Tennessee Hills. Arching between Tennessee and Alabama, these hills form the only area in Mississippi in which terrain and people are reminiscent of the southeastern mountains.

West of the Black Prairie another highland area, the Pontotoc Ridge, ranges south from the Tennessee border. This great ridge, averaging 400 to 600 feet above sea level, is one of the state's most distinctive topographic features. Its fertile sandy loam soil is excellent for orchards. A low-lying region called Flatwoods skirts the western edges of the Pontotoc Ridge and the Black Prairie. Its heavy clay soils drain poorly, and the area has never developed a prosperous economy. The North Central Hills range through northern and central Mississippi and eastward to Alabama. Their red clay soil, though not fertile, supports small farms. Before scientific farming methods were widely adopted, erosion ruined thousands of acres in these hills.

The natural environment. *Rivers.* Mississippi has many rivers, creeks, bayous, and other drainage. The state has five principal river systems, in addition to some important smaller ones. Except for a small area in the northeast that drains into the Tennessee River, these streams empty into the Gulf, either directly, or through the Mississippi and other rivers.

Climate. Mississippi's location endows it with a very favourable climatic range. The growing season is long (virtually year-round on the coast), rainfall is abundant, and extreme temperatures are unusual. Summers are warm, with temperatures sometimes exceeding 95° F (35° C). Autumn's bright, crisp days have the least rain and are usually the most agreeable of the year. January temperatures average from 42° F (6° C) to 50° F (10° C). Snowfall is rare but does occur. Supplies of water are abundant, and rainfall is usually adequate for replacement. The state's annual average is more than 50 inches (1,270 millimetres), varying by region. The coastal area is subject to hurricanes from June to October.

Vegetation and animal life. The mild climate, long growing season, and abundant rainfall provide Mississippi with a remarkable variety of plant and animal life. Live oaks and palms vary the landscape of the southern counties, and fruit trees and hardwoods thrive in the north. The magnolia and pecan trees are favourites throughout the state. More than half the land area is in forests, and both natural and cultivated floral displays are diverse and abundant.

Opening land to farming and heavy hunting reduced the once-abundant wildlife to near extinction. The wolf and panther (cougar) are gone, the bobcat is rare, and the bear even rarer. Yet deer are once more abundant and wild turkeys are increasing. The state has a great variety of resident and migratory birds. Some game fish can be taken throughout the year, with catfish, bream, bass, and perch the leading freshwater species. The Gulf is rich in shrimp, oysters, and fish, which are the mainstays of extensive commercial fishery.

Traditional regions. Almost every Mississippian knows and uses such regional designations as the Delta and the Hills; South Mississippi and the Coast; the Prairie or the Black Belt; and North Mississippi and the Northeast. The landscape of Mississippi is mainly one of forest, fields, and towns. Cities and factories take a lesser place. Cotton is king no more, and soils unsuited to row crops may support tree farms, pastures, or orchards. The two Standard Metropolitan Statistical Areas are Jackson, with nearly 260,000 inhabitants, and coastal Biloxi–Gulfport, with nearly 135,000.

Regional distinctions

Delta and hills. Geographers may speak of the Yazoo Basin, but to the people of the state it is the Delta. The Deltans pay homage to aristocratic plantation traditions, tend to be politically conservative, and usually display more solidarity than most other regions. The leading families are patrons of the arts and letters and tend to be well-read and travelled. The Hill people, however, do not defer to the Deltans, many of whose families originally came from the Hills.

South Mississippi and the Coast. What historians call the Piney Woods, covering most of the state south of Jackson, the people call South Mississippi. Its population and prosperity are growing. Change seems to come easier here than in most other regions. The farmer is fast being replaced by the worker in commerce and industry.

The coastal area is atypical of the state as a whole. More Catholic and Mediterranean than upstate, it blends French, Spanish, Latin American, Dalmatian, and British heritages in the most heterogeneous of Mississippi's regions.

Prairie and northeast. The Prairie region has some of the ways and style of both the Delta and the Hills. Its lands, however, are more productive than the Hills, its people are somewhat more expansive, and some traditions from the antebellum era remain. Northeast Mississippi developed as an area of small family farms and few plantations, and it has the lowest nonwhite population in the state.

THE PEOPLE OF MISSISSIPPI

Ethnic and religious characteristics. The white population of Mississippi is remarkably homogeneous. More than 98 percent native-born of native stock, they are predominantly of British, Irish, and northern European ancestry. The black population is of old stock and almost entirely native-born.

Until about 1940, blacks were in the majority, but by 1970 they made up about 37 percent of the population. About 4,000 Choctaw Indians remained, living on or near their reservation. Other groups are represented but in such small numbers that they are not conspicuous.

Various Protestant denominations dominate Mississippi's religious life, most notably the Baptists and Methodists. The Catholic population tends to be in the urban centres and the southernmost areas, especially the coastal counties. The Jewish congregation is almost entirely urban.

Demography. Like many predominantly rural areas, Mississippi has relatively low death rates and high birth rates. Birth rates, which, in the 1960s, averaged 22 per 1,000 inhabitants—17.4 for whites, 28.2 for blacks—were higher in only five other states. The overall fertility ratio—the number of children under five per 1,000 women of childbearing age—was 468 in 1970, a significant drop from 670 in 1960. The death rate was 9.6 per 1,000 in the 1960s: 9.2 for whites, 10.4 for blacks.

Mississippi has no great extremes of population concentration, nor are there extensive uninhabited areas in the state. Since 1950 there has been a slow but steady loss of farm population and a decline of smaller towns, and most of the centres of over 10,000 inhabitants have experienced significant growth. The average population density is 50 persons per square mile, ranging from 245 in Hinds County to seven in Issaquena County.

Patterns of migration

Mississippians, who inherited the frontier tradition of "moving on," have become as mobile as most Americans. Frequent movement by sharecroppers and tenant farmers from one farmstead to another was commonplace before 1920, at about which time the economic focus (and the locus of emigration) shifted to the cities and towns. About three-fourths of the white emigrants moved to one of the other Southern states, whereas the same proportion of black emigrants left the South entirely. The net loss by emigration has largely offset Mississippi's high rate of natural increase. If Mississippi is able to retain more of its youth, it is possible that the growth, com-

position, and distribution of the population will depart from the past experience and the current projections for the future.

THE STATE'S ECONOMY

In many economic indexes, Mississippi ranks low in its region and in the nation. During most of the 1950s and 1960s, personal income ranged from only 50 to 60 percent of the national average. This can be explained in part by the smallness of industrial employment, by the great numbers of dependent persons under 15 and over 65, and by poor re-employment opportunities for an increasing surplus farm population. Where agricultural employment continues, it is often seasonal or on a daily basis.

Personal income

There has been some improvement in employment, wages, and personal income; but the proportionate national and regional growth has been even greater, and the relative economic disadvantage continues. Nearly 90 percent of personal income is derived from eight sources. Manufacturing is the largest single source, though it falls short of the national average. The other main sources are the federal government, property, farms, state and local governments, wholesale and retail trade, operation of nonfarm commercial enterprises, and personal and business services.

Natural resources. Mineral fuels account for over four-fifths of the value of all minerals. In 1971 Mississippi ranked ninth and 16th, respectively, in petroleum and natural-gas production among the states. Important nonmetallic minerals include sand and gravel, clays, and magnesia. Iron has been mined intermittently since 1887. Aluminum ores are low in quality and little exploited.

Approximately 55 percent of the land area is in commercial forests that produce lumber, paper pulp, naval stores, and other forest products. Seafoods from Gulf waters are processed in coastal plants.

Electrical power produced within the state is from steam-generating plants. Hydroelectric power is brought in from Tennessee Valley Authority dams or through interconnections with power companies in other states. Two private companies, numerous rural cooperatives, and several municipal generating systems are in operation. Several large transmission facilities bring gas into and through the state.

Economic management. State agencies administer regulatory functions in the area of utilities, transportation, oil, gas, insurance, and pollution. The Agricultural and Industrial Board both seeks new businesses and aids in expanding existing ones by such means as loans to train and recruit workers. It is aided by the Mississippi Research and Development Center through research in economic development. The several economic-development districts promote activities in their constituent counties. Private enterprise also advertises the state's advantages.

Labour-union membership

Labour-union membership is relatively small, although widely dispersed. In 1968 unions reported 76,000 members, accounting for 13.8 percent of workers in nonagricultural employment, compared to the national level of 28.4 percent. Most large employers have a union membership, though Mississippi has a "right-to-work" law that prohibits compulsory union membership.

Transportation. The declining fortunes of rail transportation and the existence of obsolete segments of the state's highway system represent problems for transportation. Relief is hoped for, however, in the four new highways of the interstate system and in the program of highway improvement voted by the legislature. The heaviest volume of traffic is along the Gulf Coast, where it merges into the flow from the numerous upstate north–south and east–west patterns. The 200-mile Natchez Trace Parkway, part of the National Park Service, is protected from commercialization and truck traffic. Peaceful and bucolic, it preserves the natural surroundings and encompasses many sites of historic and other interest. Eventually, it will extend from Natchez to Nashville, Tennessee.

Commercial transportation, in addition to highway trucking, is very diversified. Of the 16 railways in the state, seven are entirely intrastate. Three scheduled airlines serve Mississippi, and 13 airports are certified for commercial service. Several interstate bus companies maintain service through the state, and almost all major cities and towns are served by one or more bus lines. Gulfport and Pascagoula can accommodate oceangoing ships, and low-draft oceangoing vessels can travel up the Mississippi to Natchez, Vicksburg, and Greenville. Barge traffic moves on the Mississippi, Pearl, and Yazoo rivers. The Gulf Intracoastal Waterway passes just off the coast across the Mississippi Sound.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of state government. The state government has the traditional executive, legislative, and judicial branches, but it differs from some states in that most heads of executive departments are elected rather than appointed. The bicameral legislature, which meets in limited annual sessions, comprises a 122-member House of Representatives and a 52-member Senate.

Executive, legislature, and judiciary

Persons elected to statewide offices serve four-year terms, and the governor and treasurer cannot immediately succeed themselves. In addition to commanding the National Guard, the governor may call upon a volunteer state guard and an unorganized militia, the latter potentially comprising all able-bodied men of military age.

Justice of the peace courts have original jurisdiction in misdemeanor matters and may have concurrent jurisdiction with county or circuit courts where fines, sentences, and judgments do not exceed prescribed limits. Chancery courts have jurisdiction over matters of equity, probate, juvenile delinquency (where county courts do not exist), divorce, and mental competence. The Circuit Court is the main trial court for major suits, criminal cases, and appeals from justice and county courts. The Supreme Court is the court of appeals; its nine justices are elected, from three judicial districts, for staggered terms of eight years.

Local government. Each county's governing body is a board of supervisors. The sheriff formerly functioned also as the tax collector and was compensated on a fee basis. Beginning in 1972, tax assessment and collection were combined on a salaried basis. Municipalities may be incorporated as villages, towns, or cities, though no new villages may be chartered. Governments of these units are of the mayor–council type with aldermen, the commission type with three commissioners (including the mayor), or the city-manager type, in which the manager is appointed by the council.

Politics. State and county officials are elected in the November general election. Under a system adopted by the legislature, all candidates must stand in a preferential primary in October. If no person receives a majority, the two candidates with the highest number of votes are placed on the ballot in the November election.

By 1875 Democrats had won most of the offices from the carpetbag rule that had prevailed for about a decade. Until the late 1940s, Mississippi was a one-party state, in which the Democrats usually won. Disaffection with the national party broke the pattern in following decades, when, except for 1952 and 1956, states'-rights candidates and other conservative presidential candidates received the state's electoral ballots. For offices below the national level, major consideration is focussed on personality and political philosophy on local issues.

Race and politics

In effect, Mississippi now has two Democratic parties. The traditional state party structure is conservative and has official recognition in the state. The Loyalist Democratic Party, allied with the more liberal national party, symbolizes the increased importance of the black voter. An increasing impact has been felt from drives to register black voters, court decisions banning literacy and other tests, and redistricting under court order. Beginning in the 1960s, blacks won election to an increasing number of municipal and county posts throughout Mississippi.

Finances. Sales taxes are the state's major source of revenue, followed by personal and corporate income taxes and gasoline taxes. Local governments derive their greatest income from property taxes. Considerable

amounts of federal monies are provided through numerous federal and state agencies.

Social services. Education, health, welfare, and other measurements of the quality of life in Mississippi necessarily must be considered in the perspective of the state's long history of segregation and the nearly equal number of people of each race to which that segregation was, in part, a response. Though segregation is by no means limited to the South, let alone Mississippi, the nearly equal numbers of whites and blacks in the state have intensified the actions and reactions on both sides and made the state a national scapegoat in the eyes of some and a symbol of the problem in the eyes of others. Certain researchers have pointed out that although the level of educational attainment of white Mississippians is among the highest in the nation, the inclusion of the black population in this index relegates Mississippi virtually to the bottom of the national scale. Analogous conditions prevailed in other areas.

Education. In 1957 a process of enlarging and consolidating public school districts was nearly completed. In addition, the years of confusion and emotion since the reversal of the separate but equal doctrine in 1954 have altered the Mississippi schools in a drastic manner. The year-by-year changes brought on by integration have produced a condition of semipermanent flux. The very low per pupil expenditure by the state is in large measure the result of the low tax bases that exist and the high proportion of school-age persons in the population. In terms of money spent for instruction, as distinct from buildings and utilities, the state ranks much higher than the comparative national figures indicate.

Despite its present problems, Mississippi has a distinguished history in education. In 1884 it established what was probably the first U.S. state-supported institution granting diplomas to women. The University of Mississippi, in Oxford, was chartered in 1844 and opened in 1848. In addition, the state supports 18 junior colleges, five four-year colleges, three universities, a medical centre, the Gulf Coast Research Laboratory, and the Universities Marine Center.

Health and welfare. Cases of dietary imbalance, hunger, and malnutrition found in Mississippi are associated with an overall cultural, economic, and social picture, as is the rate of infant mortality—the lowest in the state's history in the early 1970s but still nearly 60 percent above the national average. Almost all counties have some form of relief or welfare programs, many of which involve federal funds. In addition, welfare services include aid to the blind and disabled, the elderly, and dependent children. Health programs are administered by several state agencies, including the Board of Health, which dates from 1877. Historically, the state has been in the vanguard of various public health services. The causes of pellagra were discovered in 1915, through experiments at the penal farm. A model mosquito-control program eradicated malaria and yellow fever, and the state tuberculosis sanitorium is recognized nationally. Pioneer work was done and continues in educating blind and deaf children.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu

Before the Civil War, the "Planter Society" and those who identified with it had a highly developed sense of gentility. The life-style to which they aspired made patronage of the arts obligatory. They built their Georgian mansions and furnished them with art objects and fine furniture; their children were tutored in the social graces and the arts; and hospitality became an art in itself. This rural gentry, however, was only a part of the total society. The small landowner and slave alike fashioned simple, sturdy furniture, made their own oxbows and spinning wheels, or patiently crafted musical instruments. When they rested, they heard folk songs, ballads, or African lullabies. Their literature was myth, legend, and tall tale, and their dances were traditional or improvised.

From these varied sources came the present-day cultural and artistic heritage of Mississippians. In the 20th century, technology, mobility, and mass communication have created an adherence to a homogenized national culture. A freshening sense of history is evident, however, in the efforts to preserve historic landmarks and in the intensity of the collectors of the artifacts and furnishings of the past—of folk songs, implements and utensils, furniture, and manuscripts.

The arts. Mississippi has been a vital part of the flowering of Southern literature during the 20th century. The mythical county of Yoknapatawpha and the generations of its people were created by William Faulkner in a celebrated series of novels. The works of this Nobel Prize-winning Mississippian are often ranked among the highest attainments in American literature. Other natives of international literary renown include novelist Eudora Welty, novelist-critic Stark Young, and playwright Tennessee Williams.

In music, both white and black folk traditions are found: in English and Scottish ballads, and in blues, spirituals, and sacred harp singing (religious hymns or songs sung to the accompaniment of a harp). In the classical vein, the state has eight symphony orchestras, an opera guild in Jackson, and extensive musical activities at several colleges. The theatrical tradition dates from 1800, when a Natchez audience saw the first dramatic production to be presented west of the Alleghenies. Today, more than 20 community theatres and 30 colleges and universities offer dramatic fare, in addition to a semiprofessional company in Jackson.

Communications. Newspapers and broadcasting are and have long been a vigorous and important part of Mississippi's cultural life. Newspapers came on the scene in territorial days, with the appearance of *The Mississippi Gazette* at Natchez in 1799. In the early 1970s there were 19 daily and 114 weekly newspapers in the state. The *Woodville Republican*, founded in 1824, is the oldest paper continuing publication. *The Delta Democrat-Times*, edited and published in Greenville by Hodding Carter, gained national recognition through its liberal editorial policies. The state has more than 102 radio stations and 19 television stations, including the first colour educational television station in the country.

The press

Prospects. On balance, Mississippians are optimistic about the future. They are keenly aware of the state's needs, problems, and potentialities and of how these will affect their ability to retain the finer elements of their heritage and to attain a more prosperous and harmonious future. They are proud, too, of the many positive aspects of life in Mississippi: its clear air and water, its easy access to recreation and open spaces, and its low incidence of crime. In the early 1970s, Mississippi's economy was at its highest level in history, and its racial problems, though far from solved, were being confronted in an atmosphere of generally increasing goodwill.

BIBLIOGRAPHY. HODDING CARTER, *Lower Mississippi* (1942), is an interesting study by a major journalist. The classic definition of the Delta appears in DAVID COHN'S autobiographical *Where I Was Born and Raised* (1948), containing his earlier *God Shakes Creation* (1935). ARTHUR H. DE ROSIER, JR., *Removal of the Choctaw Indians* (1970), is a scholarly contribution. HORACE S. FAULKERSON, *Random Recollections of Early Days in Mississippi* (1885), contains stories from a time when vices and virtues were exaggerated. WILLIAM L. GILES, "Agricultural Revolution in the Delta," *Journal of Mississippi History*, 31:79–88 (1969), is essential for an understanding of contemporary plantation agriculture. LOUISE H. MOREHEAD, *Old Spanish Trail Along the Mississippi Gulf Coast* (1955), provides colourful material, as does W.A. PERCY, *Lanterns on the Levee* (1941). WILLIE MORRIS, *North Toward Home* (1967), is the autobiography of the former editor of *Harper's* magazine; and his *Yazoo: Integration in a Deep-Southern Town* (1971), continues his brilliant reportage. Research and opinion includes MISSISSIPPI STATE COLLEGE, AGRICULTURAL EXPERIMENT STATION, *Mississippi's Counties: Some Social and Economic Aspects*, ed. by WILLIAM A. STACEY, TOMMY W. ROGERS, and CARLTON R. SOLLIE, 3rd ed. (1966); VERNON L. WHARTON, *The Negro in Mississippi, 1865–1890* (1947); and JAMES W. SILVER, *Mississippi: The Closed Society*, new enl. ed. (1966), an unfriendly account by a noted scholar.

(J.N.B.)

Mississippi River

Paramount among North American rivers, the Mississippi combines with its major tributaries to form the world's fifth largest drainage system in area (approximately 1,244,000 square miles, or 3,221,000 square kilometres). As the central river artery of a highly industrialized nation, the Mississippi has become one of the busiest commercial waterways in the world; and as the unruly neighbour of some of the continent's richest farmland it has been subjected to a remarkable degree of man-made control. Furthermore, the river's unique contribution to the history and literature of the United States has woven it like a bright thread through the folklore and national consciousness of North America, linking the names of two U.S. Presidents—Abraham Lincoln and U.S. Grant—with that of the celebrated author Mark Twain. Draining all or part of 31 U.S. states and two Canadian provinces, the outline of the Mississippi system resembles a sturdy tree with two main lateral branches, the Ohio to the east and the Missouri to the west. Only a thin line carries the name of the parent trunk northward to the river's source at Lake Itasca, Minnesota.

Although the Mississippi can be ranked as third longest river in the world by adding the length of the Missouri (2,466 miles or 3,969 kilometres) to the Mississippi downstream of the Missouri–Mississippi confluence, the 2,348-mile (3,779-kilometre) length of the Mississippi proper is comfortably exceeded by at least a dozen other rivers. In volume of discharge, however, the Mississippi's 350,000,000,000 gallons per day is eighth greatest in the world (for further information see NORTH AMERICA).

The character and physical appearance of the Mississippi proper divides into three stages. From the source to the head of navigation at St. Paul, Minnesota, it is a clear, fresh stream winding its unassuming way through low countryside dotted with lakes and marshes. From St. Paul to the mouth of the Missouri it grows into the powerful dominating river immortalized in Mark Twain's *Tom Sawyer* and *Huckleberry Finn*. Flowing past steep limestone bluffs, it is in these reaches, where it gathers in the streams and rivers of Minnesota, Wisconsin, Illinois, and Iowa, that the river assumes the character that led Algonkian-speaking Indians to name it the "Father of Waters" (literally *Misi*, "big"; *Sipi*, "water"). Its union with the Missouri changes the Mississippi completely. The turbulent, flotsam-laden Missouri, especially when in flood, adds impetus as well as enormous quantities of silt to the clearer Mississippi. Beyond the junction with the Ohio at Cairo, Illinois, the Mississippi swells to its full grandeur. Often a mile and a half from bank to bank, it becomes a brown flood, descending with deceptive quiet toward the Gulf of Mexico.

Regional divisions

To geographers, the lower Mississippi has long been a classic example of a river in "old age"; that is, it loops and curls extravagantly along its channel, leaving behind meander scars, cutoffs, oxbow lakes, and swampy backwaters. More poetically, Mark Twain compared its shape to "a long, pliant apple-paring." Today the sunlight glittering on the twisted ribbon of water remains one of the most distinctive landmarks of a transcontinental flight. Now largely curbed by an elaborate system of embankments (levees) and spillways, this lower section of the Mississippi was the golden, sometimes treacherous, highway for the renowned Mississippi steamboats, those "palaces on paddlewheels" that so fired public imagination. From the explosive master pilot Horace Bixby portrayed by Mark Twain, down to the nostalgic lyrics of Oscar Hammerstein's "Ol' Man River," the creations of that era on the Mississippi have added colour to America's heritage.

THE PHYSICAL ENVIRONMENT

The geology and physical geography of the Mississippi drainage area are essentially those of the Interior Lowlands and Great Plains of North America. Fringes also touch upon the Rocky and Appalachian mountains and upon the rim of the Laurentian Shield to the north. The focus of the system, the floodplain of the lower Mississippi, is of particular interest in that the geology and physical geography of the region are of the river's own making. Like a huge funnel, the river has taken silt and debris from contributory areas near the lip of the funnel and deposited much of the product in the alluvial plain of the funnel's spout, illustrating the interdependence of the entire Mississippi system.

The most significant contributory area in recent times has been to the west of the river. Rising in western uplands, notably in the foothills of the Rockies, rivers like the Red, Arkansas, Kansas, Platte, and Missouri remove considerable silt loads from the rolling expanses of the Great Plains. A wide, gently sloping mantle of unconsolidated materials, laid down over rock beds of the Cretaceous Period (100,000,000 years ago), allows for each river to pick up an independent course toward the Father of Waters. Rainfall in these western areas is meagre, usually less than 25 inches a year; but because at least 70 percent of this precipitation falls between April and September, the erosive capability of the rivers is enhanced. The prairie soils, moreover, offer little resistance to erosion, so that the rivers are not only incised and braided in their courses but their silt load is high in relation to their volume. The Red River, for example, takes its name from the distinctive stain given to its waters by the red-brown soils of its upper course.

Contributory areas

The Mississippi's eastern contributory rivers drain the well-watered Appalachian Plateau. Most of this group, including the Kentucky, Green, Cumberland, and Tennessee rivers, flow via well-defined valleys into the Ohio and thence into the Mississippi. The erosive capacity of these rivers varies in relation to the geological structure of their basins. These consist of harder rocks in the higher plateau and a softer sill of limestone of the Carboniferous Period, 300,000,000 years ago, lying below the 1,000-foot line between the Ohio and Tennessee rivers and the glaciated area of the Ohio's right bank tributaries.

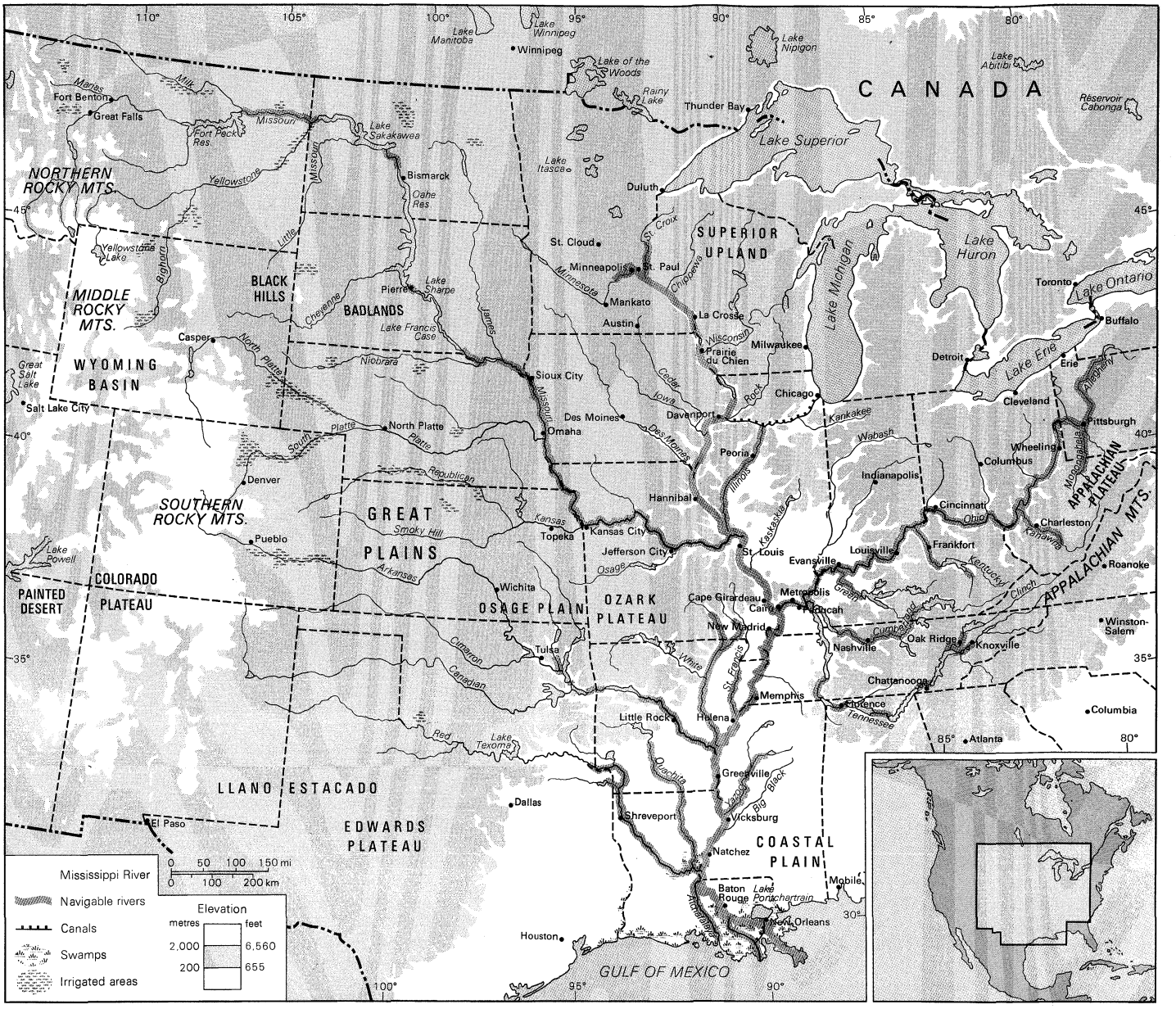
The third contributory area of the Mississippi differs, once again, from the other two. The upper Mississippi gathers its strength in a region marked and scarred by glacial action. After the great ice sheets had put down layers of glacial debris across much of Minnesota, Wisconsin, Illinois, and upper Iowa, huge quantities of meltwater flowed south, washing distinct channels through this debris. Today the upper Mississippi and its tributaries, the Wisconsin, St. Croix, Rock, and Illinois rivers, all trace the lines of these former glacial sluiceways.

Pouring southward the glacial meltwaters were joined by the proto-Missouri and Ohio rivers. The combined flood then gouged out the great north–south trough along which the lower Mississippi now flows. Six hundred miles long, this trough is from 25 to 200 miles wide and bounded by escarpments rising up to 200 feet above the valley floor. Geological studies have revealed that the floor of the glacial trough was later buried by a deep layer of material washed out from the ice sheet, and dumped to a thickness of 100–200 feet in the central section.

The delta

The Mississippi's delta is an even more striking monument to the river's constructive work. Here, at the tip of the drainage funnel, millions of years of sedimentation have spilled out across the floor of the Gulf of Mexico a great cone of sediment 300 miles in radius and 30,000 square miles in area. The surface expression of this cone (slightly to the east of its apex) is the Mississippi's present delta with an area of 10,100 square miles. Stretching into the Gulf the "fingers" of its levee-building distributaries, the Mississippi piles up 495,000,000 tons of sediment each year and inexorably advances its delta shoreline six miles seaward every century.

Understandably, the hydrology of so powerful a river as the Mississippi has been the subject of intense study. In the 19th century, Mark Twain described with considerable wit how the pilots of the Mississippi paddle-wheelers banded together to run a common information service about changing conditions along the channel. Today the Mississippi River Commission is responsible for river work and considers it worthwhile to maintain a working scale-model of the river so that its engineers can test new plans in miniature before embarking on expensive, full-scale projects. Indeed, by the 1920s it was generally held



The Mississippi River Basin and its drainage network.

that enough was known about the river's hydrology and enough control structures had been built to have tamed the river. Then, in 1927, came the most disastrous flood in the recorded history of the lower Mississippi Valley. More than 26,000 square miles of land were flooded. Communications, including roads, rail, and telephone services, were cut in many places. Farms, factories, and whole towns went temporarily under water. In modern values, something like $1,000,000,000 worth of damage was done, and 214 people lost their lives. The river engineers took another look at the hydrology of the Mississippi.

Mean water discharge data

Since the freak conditions of 1927, the mean discharge of water into the lower Mississippi by its major tributaries has been monitored as follows (in cubic feet per second): Arkansas 41,500; Illinois 20,700; Upper Mississippi 92,600; Missouri 70,600; Ohio 259,300; Red 49,300; White 30,200. The mean discharge of the main river past Vicksburg, Mississippi, is calculated at 548,160 cubic feet per second. These statistics, however, conceal all-important variations in river flow linked with the fluctuating state of the Mississippi's larger tributaries.

Broadly speaking, the western tributaries are the most irregular. They reach a spring or early summer peak that is up to three or four times as great as their winter contribution. The upper Mississippi and its tributaries reach their maximum flow about the same time (March–June), when melting snows are followed by early summer rains. The winter runoff from this area is, however, substantial. The crest of the Ohio's flow occurs slightly earlier. At Metropolis, Illinois, just short of the confluence with the Mississippi, the greatest monthly discharge is usually recorded in March, at which time the Ohio may be providing over 60 percent of the water being monitored past Vicksburg in the lower river. From these figures it is clear that the Ohio is chiefly responsible for the lower Mississippi flood situations, which may be aggravated by such factors as early rains in the Great Plains, a sudden hot spell melting the northern snows, and heavy downpours throughout the lower valley. Under such conditions the lower river will rise over its banks and put pressure against the levees. Tributaries will back up and form lakes on the far side of these same levees. The current, which normally runs no more than two to 3½ knots (2½ to four miles per hour), may then double at constricted points along the main channel. Thus the monitoring station at Vicksburg, which, at low water 1936, recorded as little as 93,800 cubic feet per second, measured 2,060,000 cubic feet per second at high water stage the following year, a daily flow of about 1,300,000,000 gallons.

Although the natural vegetation of the Mississippi's immediate valley is the product of climate and soil, rather

than of the river, the Mississippi's swamps and backwaters are ecologically noteworthy. Threaded along the river, from the wild rice marshes of Minnesota to the coastal wetlands of the delta, are pockets of thriving plant–animal associations. Here the abundance of natural cover, the comparative isolation, and the food provided by such plants as sedges, pondweeds, and millets encourages regular colonization by waterfowl. The path of these birds, as they move up and down the river with the seasons, has been called the "Mississippi Flyway," an appropriate name for the vast aerial highway that reaches from the delta to the distant summer nesting grounds in northern Canada. An estimated eight million ducks, geese, and swans winter in the lower part of the flyway, and many more birds use it en route to Latin America. Typical flyway migrants are Canada and lesser snow geese, large numbers of mallard and teal, black duck, widgeon, pintail and ring-necked ducks, and coots.

Ecology

The most important varieties of fish found in the river include several types of catfish (some of which grow to considerable size and are fished commercially by local concerns along the middle and lower river); walleyes and suckers, which thrive in the upper river and provide the basis for a sport fishing industry in Minnesota and Wisconsin; carp; and garfish. Alligators are now rare, being found only in the most isolated backwaters, while the shrimp and crab fisheries of the brackish waters are in decline.

HUMAN GEOGRAPHY

As its respectful Indian name indicates, the Mississippi played an important role in the lives of the aboriginal peoples settled on its banks. To the Indians of the river, the Mississippi was both highway and larder. On it they paddled their cottonwood dugouts and their bark canoes, and from it they took the fish that was a mainstay of their diet. Constant shifts of migration, local or large scale, interwove tribal languages and cultures. By the time the white man arrived, the Sioux who had originally lived on the upper river had withdrawn west to give place to Ojibwa, Winnebago, Fox, and Sauk. Lower down, the Illinois tribe had established prosperous agricultural communities. And in the lower valley itself lived clans of Choctaw, Koroa, Taensa, Chickasaw, Tunica, Yazoo, Pascagoula, Natchez, Biloxi and Alibamu.

The explorer Hernando de Soto, commander of the first white expedition to penetrate to the river, had high hopes of plundering the southern tribes. In May 1541 his Spanish raiding force reached the river at a point near Memphis, Tennessee. But the "Rio Grande," as the Spaniards called it, provided the newcomers with small profit and much grief. The river Indians launched repeated attacks; the Mississippi floods caught the Spanish unawares; and, ironically, de Soto, white discoverer of the river, was buried in its waters, after which the rest of his disappointed expedition retreated to the sea, their homemade boats under a running fire from the Indians.

Early exploration

The next white explorers of the river appeared in 1673 out of French Canada—two canoe loads of voyageurs commanded by Louis Jolliet, a French government agent, and Father Jacques Marquette, a Jesuit priest. Portaging from the Fox River to the Wisconsin, they paddled down the Mississippi as far as the mouth of the St. Francis River in present-day Arkansas. Nine years later the famous French explorer René Robert sieur de la Salle reached the delta itself, having opened up the even easier portage from the Great Lakes via the Illinois River. He grasped at once the strategic significance of the huge drainage system and promptly claimed the entire Mississippi Basin for France. Within a generation the Mississippi was a vital link between France's Gulf settlements and Canada, and La Salle's claim was vaguely designated as "Louisiana."

But France's grasp on the Mississippi was never firm. French traders settled the upper river, establishing towns like St. Louis and Prairie du Chien, whose names survive to this day. But the lower river passed into Spanish hands in 1769; the Treaty of Paris (1783) optimistically declared the river as the western boundary of the United States; and republican France re-acquired the much-bartered stream only long enough to sell it to the United States as part of the Louisiana Purchase (1803). This last move recognized what had been obvious for a quarter of a century—the growing domination of the river by the Americans. They came by raft, flatboat, and ark (a "raft with a rim"), built and loaded on the left bank tributaries that were in the forefront of the westward expansion of the United States. Unwieldy and expendable, these craft floated downstream to leave their cargoes and occupants as advance guards of American political and economic expansion. Only the long, slim keelboats made the return trip. They were worked upcurrent under pole, paddle, sail, or by the back-breaking "cordel," a system under which the crew went ashore with a long bow hawser and pulled the vessel upstream by brute strength.

The torch of exploration also passed to the Americans after the Purchase. In 1805–06 the pioneer expedition of Lieut. Zebulon M. Pike struggled to within 80 miles of the river's source, and in 1832 Henry Schoolcraft, an Indian agent for the U.S. government, identified and named Lake Itasca (Latin *veritas caput*, "true head") as the Mississippi's starting point.

Built in Pittsburgh in 1811, the "New Orleans" was the first steamboat to appear on the river. Like some fearful omen, her maiden voyage coincided with the series of earth tremors known as the New Madrid earthquake, the cause of much flooding and sudden relocation of sections of the main channel. But the "New Orleans" won through, and within a decade her successors had wrought a revolution on the Mississippi. In 1814 only 21 steamboats called at New Orleans, whereas 191 arrived during 1819, and 14 years later more than 1,200 cargoes were unloaded in the year. As the freight rate by steamer between Cincinnati and New Orleans plummetted from nine cents a pound to less than half a cent, it became cheaper to send freight to the east coast via the Mississippi and the long sea passage than to send it a tenth the distance over the Appalachians.

The coming of the steamboats

With the introduction of larger, high-pressure engines and more streamlined hulls, the steamboats extended their range, and the Mississippi became economic overlord to half the country. In 1820 the "Western Engineer" probed up the Missouri. In 1823, the "Virginia" churned her way up to Fort Snelling at the junction of the Mississippi with the Minnesota River. The steamboats brought an era of unprecedented prosperity to the river. Town after town grew up, dependent on the regular arrival of packet boats bringing mail and passengers or freight boats that took on local produce and left off manufactured goods. Riverbank plantations maintained their own landings so they could ship crops direct, and riverside towns vied with one another to provide services such as fuelling and warehousing. The waterfront at New Orleans, with its double line of twin-stacked steamers mingled with ocean-going ships, was among the busiest in the nation.

Then came the Civil War and, at once, a sharp struggle for control of this vital waterway, a struggle culminating in Grant's siege of Vicksburg helped by the fortified gunboats and armoured steamers of the Union. When Vicksburg and the river fell into northern hands, the South was dealt a heavy commercial and strategic blow. Lincoln, who himself had been a Mississippi flatboat man, could report "the Father of Waters flows unvexed to the sea."

The years immediately after the Civil War saw a brief but glorious revival in river traffic. New and faster steamboats were built and operated, often in rivalry to one another, a rivalry made famous by the three-day race, commencing June 30, 1870, between the "Natchez" and the "Robert E. Lee." The latter won by dint of stripping out all her unnecessary superstructure and taking on her extra fuel supplies from tenders while steaming at full speed upriver. Yet even as the river was at its most flamboyant, the same westward expansion that had brought its development now passed it over. With the construction of east–west railroads and canals, the Mississippi's north–south alignment came to be regarded as a nuisance. Towns that had once sought to be staging posts up and down the river now competed to become crossing

points. Commercial traffic dwindled, and the grand luxury paddle-wheelers gave way to sombre, more prosaic towboats with blocks of barges.

World War I produced a major resurgence in river trade. As other lines of transport became congested, the river was recognized as an increasingly valuable asset. With federal initiative, new barge lines were organized, and by 1931 the annual barge traffic moving along the river was twice the volume moved in any single year during the previous century. In 1917, for instance, the steamer "Sprague" established a new world record for size of tow. Her raft of 60 coal barges weighed 67,307 tons and covered an area of 6.5 acres.

"Sprague's" record is unlikely to be matched by any other paddle-wheeler, for the remaining steamboats are mostly showpiece examples, and the modern Mississippi towboat is of radically different design from her forebears. Screw-driven and diesel-engined, the modern towboat is made fast to the stern of her tow. Ahead stretches a rigid platform of barges as much as 1,500 feet long and with a designed draft of nine feet. Most barges are built for specific cargoes. If for dry cargo, they average 1,500 tons capacity and measure 195 feet long by 35 feet wide; for liquid cargo the proportions are about 2,500 tons capacity and 295 feet by 50 feet. To aid navigation, the towboat captain has at his disposal electronic depth finders, radar, contraguide rudders, a sophisticated system of river bank lights and markers, and radio telephone to warn other river users of his approach in narrow passages. Among the most recent cargoes to be carried along the river in this fashion are the booster rockets for space research, which are so bulky as to be unsuitable for any other mode of transport.

Responsibility for overseeing the maintenance and improvement of the Mississippi as a commercial waterway, as well as the related tasks of flood control and bank protection, falls upon the Mississippi River Commission. Created by act of Congress in 1879, the commission has supervised a massive program of river work that has profoundly reshaped the character of the Mississippi. The main stages of the navigation improvement program include a channel nine feet deep and 250 feet wide at low water between Cairo and Head of Passes (Gulf) authorized in 1896; a widening of this channel to 300 feet, approved in 1928; and its deepening to 12 feet put in hand in 1944 and still under way. Farther upriver, the outstanding achievement has been the construction of 26 locks and dams on the upper river to take the nine-foot channel up to Minneapolis-St. Paul. Connecting into the Mississippi is a vast complex of related waterways—from the Intracoastal Waterway in the south stretching between Florida and the Mexican border, to the nine-foot channels that lead up the Ohio to its tributaries and through the Illinois waterway to the Great Lakes and St. Lawrence Seaway.

Commercial use of the Mississippi waterway has shown sturdy growth. By the early 1970s traffic on the river between Minneapolis and the Gulf exceeded 230,000,000 tons a year, a 65 percent increase over the totals for a decade earlier. Between Baton Rouge and New Orleans it increased 153 percent to over 120,000,000 tons annually. Leading cargoes, by bulk, were petroleum and products, coal and coke, iron and steel, chemicals, sand and gravel, and sulfur. An increasing emphasis on bulk handling has inevitably meant the rapid growth of a few key port cities at the expense of their rivals. The chief winners in this changing relationship between river and population, with a comparison of tonnages, are given in the Table.

Commercial Tonnages Transported

	1959	1969
Minneapolis	849,865	1,687,949
St. Paul	3,592,913	4,474,863
St. Louis	9,254,125	10,463,032
Helena (Ark.)	1,970,158	2,087,259
Greenville (Miss.)	1,380,787	1,563,750
Vicksburg (Miss.)	881,120	1,569,311
Baton Rouge (La.)	25,672,105	40,845,124
New Orleans	50,441,941	113,426,557

Flood control

Flood control along the river dates back to the foundation of New Orleans by the French, who built a small levee to shelter their infant city. Over the next 200 years a complex array of riverbank structures was erected along the river to contain or divert floods. But it was not until after the catastrophic flood of 1927 that the federal government became committed to a definite program of flood control. Today the target is an integrated flood-control system able to master "project flood," the largest theoretical flood expected along the river. This program has altered the face of the river even more than the navigation program, with which it is linked, has altered its bed.

In principle, the Mississippi's floods are either constricted by levees, hastened out of danger zones by floodways and improved channels, dissipated down spillways and into reservoirs, or starved by the impounding of tributary floods. From Cape Girardeau to the sea, the river is virtually "walled in" by a vast line of main stem levees, now 1,870 miles long and planned to extend 2,192 miles. This stern concrete barrier has, incidentally, isolated the river from much of the surrounding countryside and many former riverbank towns are severed from their natural setting. In the event that the main levees are threatened, the excess floodwater drains away down spillways (*e.g.*, from north of New Orleans it is diverted past the city by spillways leading directly to the Gulf) or it bursts "fuse plugs," the specially weakened sections of levee leading to harmless floodways or reservoirs. A major example of this type of floodway occurs at New Madrid, just south of the Ohio confluence. Elsewhere, a massive program of bank strengthening using mattresses of concrete blocks has reduced lateral erosion and increased channel stability. The careful positioning of underwater dikes to deflect the current, the cutting through of oxbow bends at their necks, and a continuous dredging schedule have all helped to reduce flood levels. Tested by excessive high water stages in 1945 and 1950, the Mississippi flood-control system has noticeably reduced the risk of inundation.

PROBLEMS AND PROSPECTS

The Mississippi flood-control program is not without its lessons for the future. It has been increasingly necessary, now that the river is caught behind the main stem levees, to attend to the isolated tributaries. In some cases costly pumping stations must be installed to lift the impounded water over the levees and into the main river. More dramatic was the penalty for using the Atchafalaya River as a convenient spillway for floods on the lower river. The Mississippi threatened to divert permanently into this secondary channel, inundating the lower Atchafalaya, bypassing New Orleans, and rendering millions of dollars worth of flood-control works useless. Only at the cost of a vast and complex lock and barrage system has the danger been averted. For the future it is evident that now the Mississippi's delicate hydrology has been disturbed, the program of river work must continue for many years to come, readjusting the interplay of waters. Similarly, it is clear that no single stretch of the commercial waterway can be neglected in favour of other sections. Traffic will use the river only if the entire system is open to a common draft, and with the increasing trend to bulk transport the demands on the waterway are liable to increase.

The same situation applies to the problem of river pollution. Water samples at New Orleans show a relatively high dissolved oxygen content, mostly in the 6.0–10.0-milligrams-per-litre range, and low biochemical oxygen demand; so, by these traditional indices, the river cannot be considered polluted to any great extent. Moreover, federal water-pollution-control legislation should prevent further degradation of the water quality. On the other hand, it remains to be seen whether the current standards will be acceptable either to recreational users of the river in the future or to industries that must inevitably turn to the Mississippi as one of the nation's largest reserves of freshwater.

BIBLIOGRAPHY. MISSISSIPPI RIVER COMMISSION, *Mississippi River Navigation* (annual), a brief history, with emphasis on the development of navigation, commerce, federal par-

ticipation in waterways development, navigational details for part-time pilots, and a bibliography of charts and film material available, and *Flood Control in the Lower Mississippi Valley* (annual), information on flood-control plans, drainage statistics, and a summary of legislation on the Mississippi River and Tributaries Project; W.D. THORNBURY, *Regional Geomorphology of the United States* (1965), general data on the geology and geomorphology of the Mississippi system, with particular reference to the lower valley and marine delta formation; MARK TWAIN, *Life on the Mississippi* (1883), a classic humorous account of river conditions and of the life of a river pilot in the mid-19th century that is invaluable for background detail to the river's history; TIMOTHY SEVERIN, *Explorers of the Mississippi* (1967), a popular history of the exploration of the river.

(G.T.Se.)

Missouri

Located near the geographical centre of the conterminous United States, Missouri is the meeting place of the timberlands of the East and the prairies of the West, of the cotton fields of the South and the cornfields of the North. It has represented the political and social sentiments of a border state since its admission as the 24th member of the Union in 1821. The question of its admission as a slave or free state produced in Congress the Missouri Compromise (1820), regulating the spread of slavery in America's western territories.

Missouri was the westernmost state of the nation until the admission of Texas in 1845, and for decades it served as the eastern terminus of the Santa Fe and Oregon trails, over which tens of thousands of explorers and settlers plunged deeper into the frontier. For the West, St. Louis, Missouri's largest city, long was its closest contact with the more settled society and the culture of the East, and for the East the state had a reputation as the chief gateway to the West.

The Missouri River cuts across the state from Kansas City in the west, through Jefferson City (the state capital) in the centre, to just above St. Louis in the east, where the river joins the Mississippi. With the exception of Tennessee, Missouri has more neighbouring states than any other American state. To the north lies Iowa; across the Mississippi River to the east, Illinois, Kentucky, and Tennessee; to the south, Arkansas; and to the west, Oklahoma, Kansas, and Nebraska. According to the census of 1970, there were more than 4,600,000 Missourians inhabiting the state's 69,686 square miles (180,487 square kilometres), more than 60 percent of them living either in St. Louis or Kansas City and their surrounding counties and about another 10 percent in smaller urban areas.

Comparisons with neighbouring states

Among its neighbouring states, Missouri ranks second only to Illinois in its degree of urbanization and its amount of industrial and economic activity, though it maintains a vigorous and diversified agriculture and, in its rugged Ozark Plateau, a scenic beauty and many lively folk traditions. Missouri retains numerous conservative characteristics of the rural life that predominated prior to the 1930s. In Missouri, as in the southern states, the Democratic Party does not necessarily represent more liberal political philosophies than the Republicans. As in such other border states as Kentucky and Tennessee, Republicans have begun to make sharp inroads into the traditional Democratic orientation. Continuing low tax bases prevent the elaboration of social services, a problem felt most acutely in the two major cities, which have experienced an increasing loss of wealth to the suburbs, coincident with greatly expanding needs within the city itself. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; and MISSISSIPPI RIVER.)

THE HISTORY OF MISSOURI

Before the coming of white explorers, and for long thereafter, the land that was to become Missouri was the home of a diverse group of Indian tribes whose mounds and other remains dot the state. One of the tribes was called the Missouris.

Exploration and settlement. The recorded history of the region dates from the settlement of some French lead miners and hunters from Kaskaskia in the Illinois country at Ste. Genevieve, on the western bank of the Mississippi around 1735. At some distance from its original site, Ste. Genevieve remains the oldest continuously inhabited white settlement in Missouri. Some 30 years later, Pierre Laclede, a French fur trader from New Orleans, founded St. Louis. At the time of the Louisiana Purchase in 1803, most of the 10,000 settlers were Frenchmen from the Illinois country, but some Americans had come in from Kentucky and Tennessee, which, with Virginia, were the major immediate sources of settlers in following decades.

Statehood and growth. The "pull of the West" solidified Missouri's position as a land of passage after it achieved statehood as a slave state in 1821. Migrants bound for Texas outfitted in Missouri, and later, thousands poured through Saint Joseph, Independence, Westport Landing, and the City of Kansas (Kansas City) bound for Oregon or for California gold.

Role in the Civil War

The 1850s were years of increasing dissension, worsened by the Kansas-Nebraska Act of 1854 that set slave- and free-state advocates at one another's throats for control of those adjoining territories. Border skirmishes with Kansas became commonplace. Missouri already was moving toward a free-state economy, however, and a wave of immigrants from the northeast and from Europe in the 1850s was influential in keeping the state within the Union during the Civil War. Missourians in the Union Army outnumbered those in the Confederate Army by nearly four to one. Considerable conflict occurred in the state, much of it guerrilla warfare along the Kansas border. After the war, Confederate sympathizers were dealt with harshly; overall, however, Reconstruction was not as severe as in states in the Deep South.

The continued growth of Missouri in the late 19th and early 20th centuries was celebrated in the famous St. Louis Exposition in 1904. The state remained heavily rural and agricultural until the Great Depression of the 1930s and World War II brought about vast movements of people from the land to the cities.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Physiographic regions.* North of the Missouri River, Missouri's landscape is characterized by gently rolling hills; open, fertile plains; and well-watered prairie country. South of the Missouri, except in the extreme southeastern corner of the state and along the western boundary, the land is rough and hilly, with some deep, narrow valleys and clear, swift streams. This is a region abounding with caves and extraordinarily large natural springs. Much of the land is 1,000 to 1,400 feet (300 to 425 metres) above sea level. Near the western border, however, the elevation rarely exceeds 700 to 800 feet, and southeastern Missouri is a part of the alluvial plain of the Mississippi, less than 500 feet.

Soils and waters. Drainage and soil conditions permit farming in all of Missouri's counties, though thin soils make the Ozark Plateau a subsistence-farming region. Northern Missouri is generally well drained, much of it covered by rich soils. The alluvial soils in the bottomlands along the many rivers and streams, which some soil experts believe to be the most extensive in the nation, also add to the farming potentiality. Except for the rivers that flow southeasterly into the Mississippi, many through Arkansas, the Missouri drains most of the state. Missouri's many lakes have been created by damming rivers and streams. The Lake of the Ozarks is a large artificial lake, with an area of 65,000 acres (26,000 hectares) and a shoreline of 1,672 miles. Most have been built to furnish hydroelectric power and to prevent flooding, but they provide Missouri with excellent recreational resources. Many natural springs, some of them among the largest in the world, have a daily flow in excess of 1,000,000 gallons (about 4,000,000 litres) each.

Climate. Missouri's midcontinental location makes it susceptible to the influences of cold Canadian air; of warm, moist air from the Gulf of Mexico; and of drier air from the southwest. Though winds are variable throughout the year, summer winds generally blow from the south and southwest, whereas winter winds come from

the north and northwest. Rainfall, usually sufficient for crops, varies from around 35 inches (890 millimetres) in the north and northwest to 45 or more inches in the extreme southeast. About one-third of it falls in April, May, and June. Heavy snows are unusual, most often occurring in February and March. While Missouri does not have as many tornadoes as Oklahoma and Kansas, it does lie in "Tornado Alley," the zone of maximum occurrence, and experiences an average of 27 tornadoes annually. Maximum January temperatures range from about 35° F (2° C) in the north and northwest to 48° F (9° C) in the southeast. The extreme northwest usually experiences cooler summers than the southeast, but summer temperatures well above 100° F (38° C) may occur in any part of Missouri.

Vegetation and animal life. Originally, about two-thirds of Missouri was forested, and the remainder was covered with prairie grasses. The river bluffs and canyons of the Ozark Plateau offered, as they still do, a wide variety of unusual plants. Fameflower, royal catchfly, Trelease's larkspur, coneflower, gayfeather, and fringed poppy mallow are limited in their normal distribution to the Ozark Ridge. Trelease's larkspur and the coneflower are found only in Missouri. Elk, deer, bison, and bear once were plentiful, as were such smaller animals as beaver, otter, and mink. After settlement and the development of agriculture, most of the larger animals, with the exception of deer, disappeared, and the animals valuable for their fur were trapped out of existence.

Human settlement and exploitation. Missouri's traditional regions reflect accidents of geography, topography, and economic pursuits, or of the ethnic origin or religious or political persuasion of their inhabitants. The "Bootheel" in the extreme southeast was settled by planters from the South and was appended to Missouri at the time of statehood through the great influence of one planter; it remains today as the centre of Missouri's cotton culture, which is now declining due to cotton disease and is being replaced by the cultivation of soybeans and sweet corn. The Ozark Mountains, whose rugged terrain is unsuited to extensive agriculture and which lack roads and railroads, historically have been among the poorer regions of Missouri. Today the Ozarks constitute one of the great tourist attractions of the state. "Little Dixie" is a block of counties that lies generally north of the Missouri River and extends westward along its banks to the middle of the state. It was initially settled by persons sympathetic to the South, to slavery, and to Democratic politics. Some of the finest examples of antebellum residences are found here. South of Little Dixie, on the bluffs and uplands south of the Missouri River and west of St. Louis, is a concentration of German settlements, known traditionally as the "Missouri Rhineland." In the western part of the state, north and east of the Missouri River, is "Mormon Country." There, followers of Joseph Smith settled around 1831, first at Independence and subsequently in other counties, until they were driven out by hostile neighbours. In the centre of the state, around Boonville, Franklin, and Columbia, is the "Boone's Lick Country," where the frontiersman Daniel Boone and his sons moved from Kentucky to make salt.

The "Bootheel," the Ozarks, and other regions

Human settlement has altered Missouri's landscape significantly. Only one-third of the state remains forested, mostly on the hills and slopes of the Ozarks. Nearly all of the prairie land has been brought under cultivation. The damming of streams has produced numerous lakes and reservoirs, and in the southeast, drainage systems have converted former swamps into one of the state's richest agricultural regions. Traditionally, agriculture has been characterized by family-owned farms, but, as elsewhere in the nation, the number of farms has been decreasing for decades, while acreage and productivity have been increasing. Urbanization also has eaten into the amount of land available for agriculture. Although urban settlements are scattered throughout the state, Kansas City and St. Louis are Missouri's important centres of commerce and manufacturing and the nuclei of large metropolitan areas that extend into Kansas on the west and Illinois on the east.

THE PEOPLE OF MISSOURI

Ethnic and religious groups. After the early French settlers, immigration came largely from states to the east and northeast, as well as from the South, with the implantation of its type of economy and society in the Bootheel and in Little Dixie. The first immigrants from abroad—particularly Germans, Irish, and English—came in great numbers after 1820. By 1860, large groups of Germans had settled in St. Louis and immediately to the west, while many Irish had settled in the city. Between 1860 and 1870 the immigration from Ohio, Illinois, and Indiana exceeded that from the Upper South, while an increasing number of immigrants from Germany arrived, most of them settling in urban centres. Subsequently, both St. Louis and Kansas City attracted sizable colonies of Italians and Greeks as well as Poles and Jews.

Patterns of immigration

During the middle of the 20th century, Missouri has been affected by the northward migration of many Negroes from the rural South. From 1940 to 1960, the white population increased by 11 percent, whereas the nonwhite population increased 62 percent. By 1970, Negroes made up nearly 11 percent of the state's total population; but in St. Louis and Kansas City they comprised 41 and 22 percent, respectively. This growth has had enormous consequences, as it has had for urban centres throughout the nation.

It is not unusual that a crossroad state should exhibit great religious diversity. The Roman Catholic Church, dominant until the Louisiana Purchase, remains powerful, particularly in the two major cities. After 1803, the chief Protestant denominations as well as many smaller sects established themselves. Jewish communities have flourishing congregations in the larger cities.

Demography. Population density varies enormously, from around 10,000 per square mile in St. Louis proper to seven in Shannon County in the Ozarks. Slightly more than 70 percent of the Missourians live in urban areas. The birth rate for Missouri in 1970 was 18.0 per 1,000, as compared with the national rate of 18.2. Missouri's death rate stood at 11.3, as compared with a national rate of 9.4.

Though the trend of population movement from rural to urban areas undoubtedly will continue, the 1970 census revealed that St. Louis is losing population to its suburbs, and that Kansas City's modest gain is probably attributable to its large land area, the sixth-largest of any city in the nation. The older sections of both cities are losing population. Though Missouri has grown modestly over past decades, it nevertheless has been a population-exporting state, a pattern more characteristic of heavily rural or economically less developed states. Nothing in the present economic, political, and social conditions in Missouri suggests that this exodus of young Missourians will soon be reversed. The proportion of older inhabitants among the population will probably continue to rise, perhaps producing a simultaneous rise in conservatism and need for social services.

THE STATE'S ECONOMY

Although agriculture has remained important as an income-producing activity, manufacturing has forged ahead since World War II. Missouri has become the commercial and industrial leader among all its adjacent states, except Illinois. In some types of manufacture, particularly in the production of aerospace and transportation equipment—including automobile assembly—Missouri ranks second in the nation. Kansas City and St. Louis have always been important trading and commercial centres for large regions reaching into neighbouring states. They rank among the foremost grain and cattle markets of the nation.

Economic resources. The state's variety of resources includes lead and iron ore, limestone, timber, animal hides, vegetable fibres, hydroelectric power, natural transportation routes, and labour in abundance. Its chief sources of income are, in order, manufacturing, agriculture, and tourism. Manufacturing is led by the production of aerospace and transportation equipment, followed by the processing of food and kindred products and the

production of chemicals. Agricultural production is valued at over $1,000,000,000 annually, and tourism produces income in the millions of dollars.

Minerals

Mineral-rich Missouri leads the nation in lead production, and continuing new discoveries of lead, as well as iron ores, assure this position. About two-thirds of the forest stand lies in the Ozarks, and since the 1950s industry has made increasing use of these resources. The acreage of many marginal farms in the Ozarks is being given over to tourism, one benefit of which is their return to forest.

Finance and planning. Missouri ranks among the top ten states in banking and financial institutions of all kinds. Federal reserve banks are located in both St. Louis and Kansas City, and the Internal Revenue Service offices in Kansas City serve much of the Middle West.

Workers in Missouri have enjoyed the benefits of an expanding economy, and per capita income has kept pace with national averages. Unions have had great influence in increasing the salaries of teachers and clerical workers as well as those of workers in the trades and crafts. In spite of favourable comparisons with other states in income and revenue, Missouri has consistently ranked near the bottom in terms of state tax revenue per $1,000 of per capita income. As a result, it has suffered even more than many other states from a lack of sufficient tax revenue to meet the critical needs and services that modern governmental agencies are expected to provide.

A state Department of Business and Administration includes several divisions and commissions that have had a significant influence upon the state's economic development. Many industries have located new plants in Missouri, creating thousands of new jobs. Local chambers of commerce and private financial groups also have stimulated the state's economic growth.

Transportation. Though even the most remote towns and villages are not far from modern, hard-surfaced highways, the major flows of traffic within the state are from east to west along the Missouri Valley and southward along the Mississippi. Missouri has more than 30,000 miles of state-controlled highways and is served by seven interstate highways. Its some 7,000 miles of railroads link up with most of the nation's major trunk lines. St. Louis, Kansas City, and Jefferson City are served by Amtrak passenger service. The Mississippi and Missouri rivers, providing 1,000 miles of navigable waterways within the state, connect waterborne traffic with New Orleans. St. Louis is served by 14 airlines, Kansas City by eight. Public mass-transit facilities, however, have continued to decline in the face of the private automobile, further contributing to the decline of the city cores.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. Missouri is governed under its fourth constitution, ratified in 1945, but the basic structure of government has remained constant since the first constitution of 1820. The governor, elected for a four-year term and able to succeed himself, has the unusual power of "item veto," by which he may strike individual provisions from any appropriation bill, except those for public school support. He thus exercises enormous control over fiscal affairs. In the General Assembly, each senator represents equivalent population units, whereas each county has one representative, regardless of its population. In the early 1960s, voters in 42 rural counties had as many representatives as nearly ten times the number of citizens living in St. Louis, St. Louis County, and Kansas City and surrounding Jackson County. Missouri has yet to come forward with an acceptable plan for redistribution to meet the 1965 order of the U.S. Supreme Court for equality of representation.

The judicial system is similar to that of most other states, with a supreme court as the highest tribunal. Below it are three courts of appeals, 42 circuit courts, and a probate court in each county. An unusual feature of Missouri's judicial system is a method of selecting judges by merit. Known as the Missouri Plan, it virtually eliminates partisan political considerations, and it has been adopted by several other states. Under this plan, the governor fills a vacancy in the court by appointing one of a three-man panel selected by a nonpartisan judicial commission. When the appointee's term is about to expire, his name must appear on the next election ballot to be voted on. To remain a judge, he must first be appointed and then elected.

The city and the county are the most important units of local government. The state has 114 counties plus the city of St. Louis, which is independent of surrounding St. Louis County. Counties are administered by a county court of three elected judges, who have no judicial authority but are responsible for administering county affairs. Counties with a population of more than 85,000 may adopt their own charters. In the early 1970s, St. Louis and Jackson counties were the only home-rule counties. Missouri, the first state in the nation to permit cities to adopt their own governing charters, today has several cities with home-rule charters.

In common with other border states, Missouri usually votes Democratic, but in the 1940s the Republicans elected governors and returned Republican majorities to the General Assembly on a number of occasions. Both parties contain factions espousing liberal and conservative approaches to social and economic questions. The Democratic Party is somewhat stronger in the two large metropolitan centres, but the Republicans have considerable strength there as well.

Social conditions. In spite of its high degree of urbanization and its relatively high economic ranking, Missouri retains a strong conservative bent that is nowhere so evident as in its handling of social issues.

Conservatism in social issues

Missouri retained discrimination against Negroes long after World War II. In 1957, however, the General Assembly established the Commission on Human Rights; and in 1961 it enacted the Fair Employment Practices Act, which made discrimination on racial or religious grounds unlawful in firms employing 50 persons or more. Four years later it passed a public accommodations law, forbidding any business offering goods or services to the general public from practicing discrimination based on "race, color, creed, sex or ancestry." While some discrimination still exists, particularly in housing, Missourians generally have shown a willingness to conform to these laws. Blacks have become increasingly conscious of their rights, and in major cities there have been numerous minor confrontations and occasional serious riots. Kansas City was the scene of a serious disturbance in April, 1968. Fortunately, these disturbances have not reached the magnitude of similar outbreaks in other large cities.

Education. Since 1945, many small school districts have consolidated into fewer large ones, and school enrollments and revenues have increased. Despite the growth, however, Missouri has lagged behind other states in support of public education. In 1970, Missouri ranked 33rd nationally with respect to expenditure per pupil. After 1954, schools were desegregated rather smoothly, except in the Bootheel region, where the process lagged.

Higher education has expanded both in the public and private institutions. The University of Missouri has campuses in Kansas City, St. Louis, and Rolla, in addition to the main and oldest campus in Columbia. There are also seven state colleges plus Lincoln University, founded originally for Negroes but now enrolling whites. Among the more than 20 private institutions are Washington University and Jesuit-run St. Louis University, both nationally recognized universities in or near St. Louis. Financial support of higher education consistently has been less than its needs.

Health, welfare, and living standards. The state Department of Public Health and Welfare provides for the ill and the indigent, especially in mental-health services. Progress in the latter slowed in the early 1970s, however, because of revenue problems. In cooperation with federal agencies, St. Louis and Kansas City have undertaken massive urban-renewal programs to help relieve the problems of inner-city blight and congestion.

Though costs of living have risen, especially in the larger population centres, per capita income also has risen. Pockets of poverty exist in depressed rural areas and

in city slums, but Missouri generally has not experienced the severe poverty of states with more heavy industrialization or a greater amount of subsistence farming. The disparities between rich and poor are greatest in and around St. Louis and Kansas City. Missouri has five Standard Metropolitan Statistical Areas: St. Louis, Kansas City, Springfield, St. Joseph, and Columbia. Because those of St. Louis and Kansas City cut across state lines, their problems of metropolitan government and management are compounded.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. Diversity characterizes Missouri's cultural milieu, from the centres of fine art, music, and theatre along the St. Louis–Kansas City axis to the folk culture and native crafts of the Ozarks.

The state has furthered cultural opportunities through the Missouri State Library, established in 1946, and the Missouri State Council on the Arts, created in 1965, the second such body in the nation established with legislative support. The state library has been responsible for the rapid growth of county and regional libraries. The larger cities have their own library systems. The council on the arts has stimulated communities to expand their cultural resources.

The arts. From the state's beginnings, the arts have flourished in Missouri. In painting, George Caleb Bingham and Thomas Hart Benton have been pre-eminent. Samuel Langhorne Clemens, known as Mark Twain, remains Missouri's most distinguished literary figure, both for his world-renowned works and for his immortalization of mid-19th-century life in Hannibal and along the rivers. Schools of art and design exist in St. Louis and Kansas City, and music flourishes in both. St. Louis' Gateway Arch, designed by Eero Saarinen, is the most spectacular of the diverse architectural styles in evidence throughout the state.

Folk traditions and crafts

The Ozarks abound in folk traditions, tales, and ballads. The growth of tourism there has generated a revival of culture and native crafts. The Ozark Ridge was settled primarily by pioneers from the southern Appalachians, who brought with them traditional songs and ballads, some of which had been brought over from England and Scotland in the 17th century. Native crafts once practiced out of necessity by the pioneers have begun to flourish again in recent years in response to the interest of tourists and also because of recognition of the intrinsic merit of the objects made. Weaving, basketmaking, and pottery are among some of the most important crafts, and their development is encouraged through the Missouri Federation of Arts and Crafts.

Cultural institutions. Besides its universities and colleges, Missouri has outstanding cultural institutions, among them the St. Louis Symphony Orchestra, the second-oldest and one of the major musical ensembles in the nation. In Kansas City, the William Rockhill Nelson Gallery and Atkins Museum of Fine Arts owns one of the finest collections of Oriental art in the Western Hemisphere, and the Linda Hall Library of science and technology has an outstanding scientific collection. Independence is the home of the Harry S. Truman Library and Museum. Fulton, where the British leader Winston Churchill made his famous "Iron Curtain" speech, has a collection of Churchilliana in the St. Mary Aldermanbury Church, whose stones were reassembled there after its destruction in the World War II bombings of London.

Recreation. Kansas City and St. Louis are the homes of major-league baseball and football, along with other professional sports. Increased leisure time and increased mobility have stimulated an enormous interest in recreation in Missouri. The State has developed a superb system of state parks and historic shrines and memorials, which are attractive to residents and visitors alike. Numerous manmade lakes afford fishing and water skiing, while the clear cool rivers of the Ozark Plateau offer pleasure to canoeists, fishermen, and campers.

Press and broadcasting. The Missouri Press Association, established in 1867, has had an important effect upon the development of the press in the United States. It was responsible for establishing, at the University of Missouri, the world's first school of journalism and for founding the Missouri State Historical Society, with the largest membership in the nation. In addition to numerous local newspapers and journals, Missouri publishes newspapers of national distinction, including the *St. Louis Post-Dispatch*, made famous by Joseph Pulitzer, and the *Kansas City Star* and *Times*. There are more than 160 commercial radio stations and 24 television outlets, both commercial and educational.

Prospects. Since the end of World War II, social and economic conditions have been transformed by the forces associated with the technological revolution. Manufacturing has supplanted agriculture as the chief income-producing activity, population has become concentrated in urban areas, and important advances have been made in education, health, and welfare. In spite of this progress, in the early 1970s Missouri had lost some of its forward momentum, and its rate of growth was falling behind states of similar wealth and resources. One of the most critical problems is the conservatism of Missourians, who tend to dislike large legislative appropriations and tax increases to support their institutions. Reflecting the same conservatism is the steady refusal to call for a new constitution, more adequate to present conditions. Unless rapid changes come about, many observers believe that the future may see retreat rather than advance in Missouri's progress.

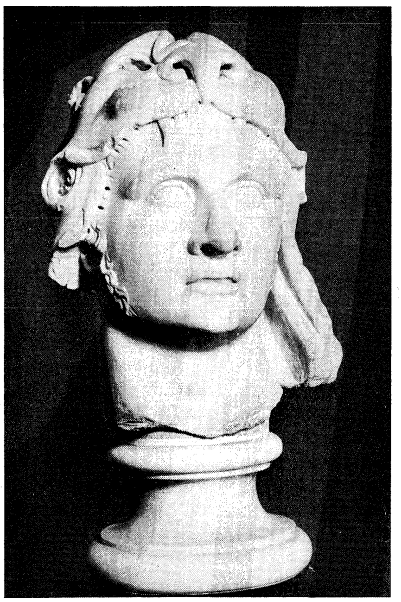
BIBLIOGRAPHY. LOUIS HOUCK, *A History of Missouri from the Earliest Explorations and Settlements Until the Admission of the State into the Union*, 3 vol. (1908), the definitive history of Missouri to statehood; DUANE G. MEYER, *The Heritage of Missouri: A History*, rev. ed. (1970), perhaps the best one-volume history of Missouri; EDWIN C. MCREYNOLDS, *Missouri: A History of the Crossroads State* (1962), a well-balanced survey, with special emphasis on political and economic factors; NOEL P. GIST (ed.), *Missouri: Its Resources, People and Institutions* (1950), essays and studies on every aspect of Missouri's development—perhaps the most useful single volume for the interested reader; FRANCES LEA MCCURDY, *Stump, Bar, and Pulpit: Speechmaking on the Missouri Frontier* (1969), a delightful study of speeches and speechmakers in early Missouri, with excellent bibliographical references; ROBERT F. KARSCH, *Political Parties, Elections, and the General Assembly in Missouri* (1964), a penetrating analysis of the problems and prospects confronting Missouri politics; MILTON D. RAFFERTY, RUSSELL L. GERLACH, and DENNIS J. HREBEC, *Atlas of Missouri* (1970), a useful atlas emphasizing economics and political and demographic resources, divisions, and concentrations; MACKINLAY KANTOR, *Missouri Bittersweet* (1969), delightful recollections of a trip through Missouri by a distinguished novelist.

(E.J.W.)

Mithradates VI Eupator

Mithradates VI Eupator, called the Great, ruled from 120 to 63 BC over the kingdom of Pontus, one of the small, superficially Hellenized states that grew up in Anatolia in the wake of the conquests of Alexander I the Great. Under Mithradates VI's energetic leadership, Pontus expanded to absorb several of its small neighbours and, for a brief while, contested Rome's hegemony in Asia Minor.

Mithradates the Great was the sixth—and last—Pontic ruler by that name. Mithradates (often misspelled Mithridates and meaning "gift of [the god] Mithra") was a common name among Anatolian rulers of the age. When Mithradates VI succeeded his father, Mithradates Euergetes, in 120 BC, he was then only a boy, and for a few years his mother ruled in his place. About 115 BC, she was deposed and thrown into prison by her son, who thereafter ruled alone. Mithradates began his long career of conquest by dispatching successful expeditions to the Crimea and to Colchis (on the eastern shore of the Black Sea). Both districts were added to the Pontic kingdom. To the Greeks of the Tauric Chersonese and the Cimmerian Bosporus (Crimea and Straits of Kerch) Mithradates was a deliverer from their Scythian enemies, and they gladly surrendered their independence in return for the protection given to them by his armies. In Anatolia, however, the royal dominions had been considerably dimin-



Mithradates, bust in the Louvre, Paris.
Cliche Musees Nationaux, Paris

ished after the death of Mithradates V: Paphlagonia had freed itself, and Phrygia (*c.* 116 BC) had been linked to the Roman province of Asia. Mithradates' first move there was to partition Paphlagonia and Galatia between himself and Nicomedes III of Bithynia; but next, he quarrelled with Nicomedes over Cappadocia. Successful at first on two occasions, he was on both deprived of his advantage by Roman intervention (*c.* 95 and 92). While appearing to acquiesce, he resolved to expel the Romans from Asia. A first attempt to depose Nicomedes IV of Bithynia, who was completely subservient to the Romans, was frustrated (*c.* 90). Then Nicomedes, instigated by Rome, attacked Pontic territory, and Mithradates, after protesting in vain to the Romans, finally declared war (88).

The First Mithradatic War

Nicomedes and the Roman armies were defeated and flung back to the coasts of the Propontis and the Aegean. The Roman province of Asia was occupied and most of the Greek cities in western Asia Minor allied themselves with Mithradates, though a few held out against him, such as Rhodes, which he besieged unsuccessfully. He also sent large armies into Greece, where Athens and other cities took his side. But the Roman generals, Sulla in Greece and Fimbria in Asia, defeated his forces in several battles during 86 and 85. In 88 he had arranged a general massacre of the Roman and Italian residents in Asia (80,000 are said to have perished), in order that the Greek cities, as his accessories in the crime, should feel irrevocably committed to the struggle against Rome. As the war turned against him his former leniency toward the Greeks changed to severity; every kind of intimidation was resorted to—deportations, murders, freeing of slaves. But this reign of terror could not prevent the cities from deserting to the victorious side. In 85, when the war was clearly lost, he made peace with Sulla in the Treaty of Dardanus, abandoning his conquests, surrendering his fleet, and paying a large fine.

In what is called the Second Mithradatic War, the Roman general Lucius Licinius Murena invaded Pontus without provocation in 83 but was defeated in 82. Hostilities were suspended, but disputes constantly occurred, and in 74 a general war broke out. Mithradates defeated Marius Aurelius Cotta, the Roman consul, at Chalcedon; but Lucullus worsted him outside Cyzicus (73) and drove him, in 72, to take refuge in Armenia with his son-in-law Tigranes. After scoring two great victories at Tigranocerta (69) and Artaxata (68), Lucullus was disconcerted by the defeat of his lieutenants and by mutiny among his troops. In 66 Lucullus was superseded by Pompey, who completely defeated both Mithradates and Tigranes. Mithradates established himself in 64 at Panticapaeum (Kerch) on the Cimmerian Bosporus and was planning an invasion of Italy by way of the Danube when his own troops, led by his son Pharnaces II, revolted. After vainly trying to poison himself, Mithradates ordered a Gallic mercenary to kill him (63 BC). His body was sent to Pompey, who buried it in the royal sepulchre at Sinope, the Pontic capital.

Assessment

Mithradates was a man of great stature and physical strength, a brave fighter, and a keen hunter. The ancient historians have invested him with a halo of romance, and many strange stories are told of his life. Particularly famous is his supposed self-immunization against poison, which he achieved by administering to himself increasingly large nonlethal doses. Whatever the truth of these stories, there can be no doubt that he had some of the worst traits often associated with the Oriental despot—cruelty and sensuality, in particular. On the other hand, it cannot be denied that Mithradates was a ruler of astonishing energy and determination, or that he possessed political skill of a high order. The fact that he was one of the few men to offer a serious challenge to the Roman Republic is sufficient testimony to his ability. He organized the forces at his disposal very effectively, and he had a good grasp of strategy. He was unlucky in having to face three exceptionally brilliant Roman generals; unlucky, too, in coming to power at a time when the Hellenistic world was in the final stage of its collapse. It is quite conceivable that had he been born a century earlier he could have constructed an enduring Greco-Asiatic empire. A cunning, brutal tyrant, he concerned himself solely with maintaining and strengthening his own power. He posed as the champion of Hellenism, but this was mainly to further his political ambitions; it is no proof that he was deeply imbued with Greek culture or that he felt a mission to promote its extension within his domains. Hellenism made advances in Pontus during his reign, as it had under his predecessors, but this was a natural process. He treated all alike; Greek, Roman, and Asiatic were welcome at his court provided that they could be of use to him (his military subordinates were mostly Greeks, though in later years he employed several Roman renegades); but he trusted no one. Just as it is impossible to speak of his favouring one religion or culture above another, so it is impossible to believe that he had any notion of bringing Greeks and Asiatics closer together in a new kind of political and social system. His posing as a liberator of the Greeks from Roman oppression and, later, his encouragement of social revolution in the Greek cities of the province of Asia can only be interpreted, in both cases, as the actions of an opportunist seeking immediate political advantages.

BIBLIOGRAPHY. T. REINACH, *Mithridate Eupator, roi de Pont* (1890), gives a full account and analysis of the historical sources for Mithradates' life (in French). M. ROSTOVTZEFF in the *Cambridge Ancient History*, vol. 9, ch. 5 (1932), describes the Pontic kingdom, its institutions, and its history; the available evidence derives mainly from Mithradates' own reign. Volume 1 of D. MAGIE, *Roman Rule in Asia Minor* (1950), provides a narrative account of Mithradates' reign; vol. 2 consists of critical footnotes—an examination of the literary and archaeological evidence. See especially chapters 9, 14, and 15.

(R.H.S.)

Mithraism

Mithraism was the worship of Mithra, the Iranian god of the sun, justice, contract, and war in pre-Zoroastrian Iran, known as Mithras in the Roman Empire during the 2nd and 3rd centuries.

HISTORY

The pre-Zoroastrian period. Before Zoroaster (6th century BC or earlier), the Iranians had a polytheistic religion and Mithra was the most important of their gods. First of all, he was the god of contract and mutual obligation. In a cuneiform tablet of the 15th century BC that contains a treaty between the Hittites and the Mitanni, Mithra is invoked as the god of oath. Furthermore, in some Indian Vedic texts the god Mitra (the Indian form of Mithra) appears both as "friend" and as "contract." The word

mitra may be translated in either way, because contracts and mutual obligation make friends. In short, Mithra may signify any kind of communication between men and whatever establishes good relations between them. Mithra was called the Mediator. Mithra was also the god of the sun, of the shining light that beholds everything, and hence was invoked in oaths. The Greeks and Romans considered Mithra as a sun god. He was probably also the god of kings. He was the god of mutual obligation between the king and his warriors, and hence the god of war. He was also the god of justice, which was guaranteed by the king. Whenever men observed justice and contract, they venerated Mithra.

Sacrifice of the bull

The most important Mithraic ceremony was the sacrifice of the bull. Opinion is divided as to whether this ceremony was pre-Zoroastrian or not. Zoroaster denounced the sacrifice of the bull, so it seems likely that the ceremony was a part of the old Iranian paganism. This inference is corroborated by an Indian text in which Mitra reluctantly participates in the sacrifice of a god named Soma, who often appears in the shape of a white bull or of the moon. On the Roman monuments, Mithra reluctantly sacrifices the white bull, who is then transformed into the moon. This detailed parallel seems to prove that the sacrifice must have been pre-Zoroastrian. Contract and sacrifice are connected, since treaties in ancient times were sanctioned by a common meal.

The Achaemenid period. Beginning with Darius (522–486), the Persian kings of the Achaemenid dynasty were Zoroastrians. But Darius and his successors did not intend to create political difficulties by attempting to eradicate the old beliefs still dear to the heart of many nobles. Thus, the religion of Zoroaster was gradually contaminated with elements of the old, polytheistic worship. Hymns (the Yashts) were composed in honour of the old gods. There is a Yasht dedicated to Mithra, in which the god is depicted as the all-observing god of heavenly light, the guardian of oaths, the protector of the righteous in this world and the next and, above all, as the archfoe of the powers of evil and darkness—hence, the god of battles and victory.

Decline of knowledge about Persian Mithraism

In the mixed religion of the later Achaemenid period, however, the Zoroastrian aspects clearly dominate the heathen aspects. The sacrifice of the bull, abhorred by every Zoroastrian, is never mentioned. When Alexander the Great conquered the Persian Empire about 330 BC, the old structure of society appears to have broken down completely and about the worship of Mithra in Persia no more is heard.

The Hellenistic period. Local aristocrats in the western part of the former Persian Empire retained their devotion to Mithra. The kings and nobles of the border region between the Greco-Roman and the Iranian world still worshipped him. When Tiridates of Armenia acknowledged the Roman emperor Nero as his supreme lord, he performed a Mithraic ceremony, indicating that the god of contract and of friendship established good relations between the Armenians and the mighty Romans. The kings of Commagene (southeast of Turkey) venerated Mithra. Mithradates VI of Pontus may have been a worshipper of the god, and his allies, the Cilician pirates, are known to have performed Mithraic ceremonies (67 BC). The worship of Mithra, however, never became popular in the Greek world, because the Greeks never forgot that Mithra had been the god of their enemies, the Persians.

The Roman period. There is little notice of the Persian god in the Roman world until the beginning of the 2nd century but, from the year AD 136 onward, there are hundreds of dedicatory inscriptions to Mithra. This renewal of interest is not easily explained. The most plausible hypothesis seems to be that Roman Mithraism was practically a new creation, wrought by a religious genius who may have lived as late as *c.* AD 100 and who gave the old traditional Persian ceremonies a new Platonic interpretation that enabled Mithraism to become acceptable to the Roman world.

Roman Mithraism

Roman Mithraism, like Iranian Mithraism, was a religion of loyalty toward the king. It seems to have been encouraged by the emperors, especially Commodus (180–192), Septimius Severus (193–211), and Caracalla (211–217). Most adherents of Mithra known to us from inscriptions are soldiers of both low and high rank, officials in the service of the emperor, imperial slaves, and freedmen (who quite often were very influential people)—persons who probably knew which god would lead them to quick promotion.

Mithraic sanctuaries and dedications to Mithra are numerous at Rome and Ostia, along the military frontier, in Britain, and on the Rhine, Danube, and Euphrates. Few dedications are found in peaceful provinces; when they do occur the dedicator is usually a provincial governor or an imperial official. Within a few generations, the Roman world had completely assimilated the Persian god. When Diocletian attempted a renewal of the Roman state and religion, he did not forget Mithra. In the year AD 307, in a dedication from Carnuntum (at the Danube, near Vienna), Diocletian and his colleagues dedicated an altar to Mithra, as the patron of their empire (*fautori imperii sui*). But in 312, Constantine won the battle at the Milvian Bridge under the sign of the cross. Instantaneously, the dedications to Mithra ceased, even though there was no immediate public interdiction of Mithraic ceremonies. The worship seems to have collapsed quite suddenly when imperial favour ceased to be with the Mithraists. Dedications to Mithra, it is true, appear again between about 357 and 387, but only at Rome. The dedicators all come from the old pagan aristocracy of the city of Rome, which in this period was in open opposition to the new Christian emperor at Constantinople. In these inscriptions, however, Mithra is only one of many traditional pagan gods who were worshipped by the Romans. The Mithraic mysteries had gradually faded long before. And when the Roman opposition was defeated, pagan worship was suppressed altogether.

MYTHOLOGY AND THEOLOGY

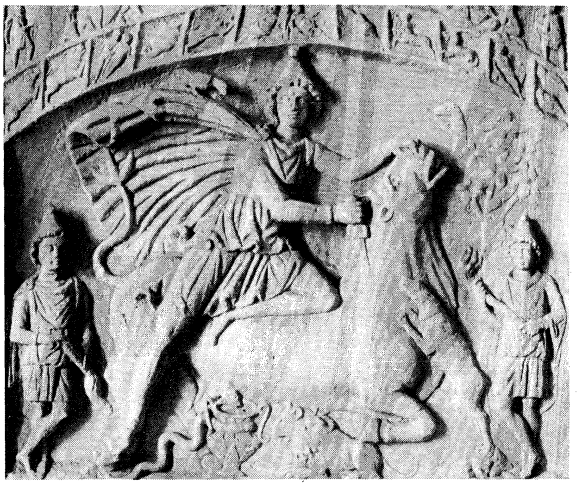
Creation myths

Mythology. The creation of the world is the central episode of Mithraic mythology. According to the myths, the sun god sent his messenger, the raven, to Mithra and ordered him to sacrifice the bull. Mithra executed the order reluctantly; in many reliefs he is seen turning aside his face in sorrow. But at the very moment of the death of the bull, a great miracle happened. The white bull was metamorphosed into the moon; the cloak of Mithra was transformed into the vault of the sky, with the shining planets and fixed stars; from the tail of the bull and from his blood sprang the first ears of grain and the grape; from the genitals of the animal ran the holy seed which was received by a mixing bowl. Every creature on earth was shaped with an admixture of the holy seed. One Mithraic hymn begins: "Thou hast redeemed us too by shedding the eternal blood." The plants and the trees were created. Day and Night began to alternate, the Moon started her monthly cycle, the Seasons took up their round dance through the Year, and thus Time was created. But, awakened by the sudden light, the creatures of the dark emerged from earth. A serpent licked the bull's blood. A scorpion tried to suck the holy seed from the genitals. On the reliefs, a lion often is also seen. With the bull's death and the creation of the world, the struggle between Good and Evil began: thus is the condition of man's life. The raven symbolizes Air, the lion Fire, the serpent Earth, and the mixing bowl Water. So the four elements (air, fire, earth, and water) came into being, and from them all things were created. After the sacrifice, Mithra and the sun god banqueted together, ate meat and bread, and drank wine. Then Mithra mounted the chariot of the sun god and drove with him across the ocean, through the air to the end of the world.

Theology. The myth was interpreted by the Roman Mithraists in terms of Platonic philosophy. The sacrifice took place in a cave, an image of the world, as in the simile of the cave in Plato's *Republic*. Mithra himself was equated with the creator (demiurge) of the *Timaeus:* He was called "demiurge and father of all things," like the Platonic demiurge. The four elements, the mixing bowl, the creation of Time, the attack of the wicked animals upon the newborn creature are well-known features of the

Timaeus. The Mithraic doctrine of the soul is intimately linked with the myth of creation and with Platonic philosophy. As in the *Timaeus*, the soul of man came down

Bavaria-Verlag



Mithra slaying the bull, flanked by two torchbearers. Below the bull is a group composed of a serpent, an urn, and a lion; the signs of the Zodiac are depicted on the arch. Bas-relief, 2nd century AD. In the Städtisches Museum, Wiesbaden.

from heaven. It crossed the seven spheres of the planets, taking on their vices (*e.g.*, those of Mars and of Venus) and was finally caught within the body. The task of man is to liberate his divine part (the soul) from the shackles of the body and to reascend through the seven spheres to the eternal, unchanging realm of the fixed stars. This ascension to the sky was prefigured by Mithra himself, when he left the earth in the chariot of the sun god.

WORSHIP, PRACTICES, AND INSTITUTIONS

The Mithraic sanctuaries were subterranean caverns, which presented obvious limitations of size. None of the many excavated shrines could receive more than a hundred persons, and most even fewer. All ceremonies were of necessity enacted in artificial light. The cavern always contained a well. Access to the cavern often consisted of a system of subterranean passages, which were used in the initiation ceremonies. Men only were admitted to this religion of soldiers, and no organizational hierarchy seems to have existed.

Initiation rites

The initiates were organized in seven grades: *corax*, Raven; *nymphus*, Bridegroom; *miles*, Soldier; *leo*, Lion; *Perses*, Persian; *heliodromus*, Courier of (and to) the Sun; *pater*, Father. To each rank belonged a particular mask (Raven, Persian, Lion) or dress (Bridegroom). The rising of the Mithraist in grade prefigured the ascent of the soul after death. The series of the seven initiations seems to have been enacted by passing through seven gates and climbing a ladder of seven steps. Each grade was attributed to one of the seven planetary gods. The zealous Mithraist gradually passed the spheres of these minor deities and finally reached the region of the fixed stars. Little is known about initiation ceremonies. Ancient texts refer to ablutions (baptism) and purifications and chastisements, to fetters and liberation, and to certain ceremonial passwords. Frescoes at Capua (Italy) show the initiates blindfolded, kneeling, and prostrated. A simulated death and resurrection was probably part of the ceremony. Tertullian, the 2nd century North African Christian theologian, describes the test of courage to which the *miles* (Soldier) was subjected. Weapon in hand, he had to force his way—probably by a sham duel—to a wreath. When he had succeeded, an officiant offered to crown him with the wreath. But the candidate had to decline, saying that Mithra alone was his wreath; and throughout the rest of his life, he never again bore a wreath.

The Mithraic caverns were decorated with frescoes, reliefs, and statues of minor deities and of the planetary gods. A narrow aisle was flanked on both sides by a broad, raised bench on which the worshippers kneeled or reclined. At one end of the aisle (often apselike) there was always a relief or fresco representing the sacrifice of the bull. Sometimes, the relief could be turned on a pivot; the back of the stone represented the repast of Mithra and the sun god. While it is unlikely that the ceremony of the bull's sacrifice was frequently performed, the common meal of the initiates was a regular feature of Mithraic worship.

MITHRAISM AND CHRISTIANITY

There are obvious similarities between Mithraism and Christianity; however, they should not be overestimated. Mithraism permitted only male devotees, and the social strata to which it appealed included mainly soldiers, officials, and men in imperial service. Christianity, on the other hand, was primarily the religion of the poor or of the lower classes. If there was competition between the two, it was for the favour of the emperor rather than for followers. The Christian Church triumphant ultimately destroyed pagan worship and sanctuaries everywhere. The Mithraic shrines were pillaged, and on the spot of the former caverns Christian basilicas were built.

BIBLIOGRAPHY. *The Avestan Hymn to Mithra* was edited and translated from the Iranian text by ILYA GERSHEVITCH (1959). For the Roman mysteries, the fundamental work is FRANZ CUMONT, *Textes et monuments figurés relatifs aux mystères de Mithra*, 2 vol. (1894–1900). His conclusions are available in English, *The Mysteries of Mithra* (1903), and *The Oriental Religions in Roman Paganism*, ch. 6 (1911). Cumont's work is now partly replaced by the monumental work (in English) of MAARTEN VERMASEREN, *Corpus inscriptionum et monumentorum religionis Mithriacae*, 2 vol. (1956–60). Vermaseren also published a popular account, *Mithras, de geheimzinnige God* (1959; Eng. trans., *Mithras, the Secret God*, 1963). For the monuments in Britain, see EVE and JOHN R. HARRIS, *The Oriental Cults in Roman Britain*, pp. 1–54 (1965). See also LEROY A. CAMPBELL, *Mithraic Iconography and Ideology* (1968). On the creation myth and the Platonizing theology, see R. MERKELBACH, "Die Kosmogonie der Mithrasmysterien," in *Eranos-Jahrbuch*, 34:219–257 (1965).

(R.Me.)

Modern Dance

The art form that came to be known as modern dance emerged in the early years of the 20th century as a protest against the prevalent forms of theatrical dancing. Led by innovators in central Europe and the United States, the movement was aimed at making dance a more viable, contemporary art. Leaders of the movement objected, basically, to two forms then prevalent. On the one hand there was the classical ballet, which, its detractors felt, had reached a low point in its artistic development and had become largely an empty display of technical virtuosity. On the other hand there was the so-called interpretive dance, which they recognized as being often full of feeling but found devoid of a basic concern for communicating with an audience.

In the words of the critic John Martin, modern dance aimed "to externalize personal, authentic experience." Modern dance evolved as a means of projecting feeling that was deeply, personally experienced but that also communicated a wider, more universal experience to the theatre audience.

PRINCIPLES UNDERLYING MODERN DANCE

Ballet and modern dance. In the beginning, there was a complete rejection of the vocabulary of the balletic tradition as it had formed in the late 17th and 18th centuries to embody ideals of aristocratic elegance and decorum. In the 19th century, ballet had become a vehicle for pure virtuosity, sacrificing even its original nobility for spectacle and display. By the start of the 20th century, dance innovators felt they could not communicate with techniques that expressed values they saw as alien to contemporary society. They sought to return to the most primitive, most basic, most human function, the immediate communication of significant feeling; and they believed that to be significant the feeling had to be contemporary.

In line with the tendencies of modern painting, sculpture, and music, choreographers stripped the dance to the essentials of movement, which could be inherently expressive when not obscured by artificial ornament.

New approaches to dance

For their themes, the modern dancers turned away from the fairy tales and romantic legends of the ballet to primitive ritual and classic myth and to the social and psychological problems that beset their own time. With themes that allowed the choreographer to comment on the most basic needs and aspirations of man, the subject of the dance was no longer incidental to the dance: it was the vital core and substance of the work.

For the modern dance, movement was motivated by message and strict economy was observed. The approach was functional; if a dancer leaped or turned, it had to be for an emotional, dramatic reason, never for mere display. Taking a serious theme and presenting it sparsely, choreographers of modern dance often met with charges of obscurity and drabness. But because their objective was to develop an art that the audience must make an effort to understand, they remained adamant and made no attempt to adorn their works with graceful but irrelevant movements to capture the fancy of a broad public.

In contrast to the pleasing symmetrical design of ballet, modern dance movements struck many as ugly, gross, distorted. New ideas of physical beauty, derived in part from Picasso and primitive artifacts, favoured expressiveness over charm. In terror the body curled inward on itself; in daring assault it reached outward precariously and fell heavily to the ground; anger made the dancer jerk in violent spasms to jagged rhythms.

Similarly, there was no compromise with the demands of spectacle in stage design. Costumes were most often of plain black jersey; the feet were bare, exposed. Scenery was sparse; the music was often dissonant when not dispensed with altogether. The aim of all appeared to be to startle the audience into an awareness that was frequently unpleasant. Modern dance thus understood has never been a form of popular art.

Intellectual and artistic influence. If ballet of the late 19th century was a negative influence on the modern dance, there were positive ones as well. Some of these came from other arts in which the revolt against romanticism had already established modernist tendencies by 1900. Others came from changes in the social and intellectual environment. Although dance was among the last of the arts to be affected by these changes, it may have been the one most drastically altered by them.

Impacts of psychology and anthropology

Of primary importance for the dance were developments in two areas: psychology and anthropology. In the former, the writings of Sigmund Freud were prompting people to look not so much to man's actions as to the motives that allegedly led him to those actions. Nineteenth-century thought held that intense emotions were best hidden from view and preferably not even admitted to exist. Fighting the denial of these basic human drives, Freud's followers urged frank and honest emotional revelation. Modern dance choreographers, in their turn, began to rediscover the power of movement in communicating subjective experience, and their new directions were reflected in both subject matter and structure. The use of naturally expressive gesture, known to the ancient Greeks, had long been lost to an art that had adopted a strict code of graciousness and good manners.

The themes of modern dance dealt with intimate human relations and passions. The action of a dance could centre not on a series of adventures culminating, for example, in a wedding but on an internal change, a recognition of a significance or truth. In structure, modern dance abandoned the series of decorative movements hung upon a feeble thread of a plot and created works in which every moment was both dance and drama, every movement served to express feeling. The anthropological revolution, stirred by the publication of Sir James George Frazer's *Golden Bough* (1890; expanded 1907–15), led the modern dance to investigate the forms of primitive ritual. For three centuries theatrical dancing had been derived from a refinement of patterns originating in folk dance, once seriously social in function but long since reduced to a more or less formal kind of play. Modern dance turned instead to the older concept of dance as an instrument of magic and religion, as a means for the survival of a community through the proper performance of sacred rites designed to banish evil spirits or to promote growth and fertility. This recalling of dance to its original seriousness was congenial to choreographers eager to make the public aware of the social significance inherent in their work. For their contemporary world, they saw the theatre as potentially a place of magic in which the souls of men could be purified and refreshed through the vigorous, direct qualities of movement found in primitive dance.

Impact of other arts

The modern dance found sources of technical inspiration in developments within the other arts. In the early 20th century, serious painters and sculptors had moved away from the concept of art as decoration or as an attempt to reproduce the literal forms of nature toward an increasing abstraction that centred on such movements as Cubism, Expressionism, and Fauvism. Like these visual artists, modern dancers sought the essence of form and emotion underlying the literal: they stripped away the superfluous, they used sharp body angles and off-balance positions for the direct expression of emotion. And as the painter found a new aesthetic in his materials—in the textures and colours of his paints, canvases, and multimedia collages—so the choreographer asserted the inherent beauty of the human body with its joints and its natural weight subject to the pull of gravity.

Like the musical composers who in the early 20th century were introducing new ideas of tonality, rhythm, and structure, modern dancers turned from the romanticism of the past with its rich, sweet harmonies and its even, pleasing rhythms. As Igor Stravinsky and Béla Bartók experimented with polytonal and polyrhythmic structures, so the choreographer looked for new principles by which to order his composition. His performers moved to dissonant music with uneven measures. Sometimes dispensing with musical accompaniment altogether or using only percussion for a skeletal score, he juxtaposed one violent rhythm against another. Like the musician who refused to confine himself to the traditional major and minor scales and who refused to end every phrase with a consonant chord, the choreographer sought freedom in structuring his sequences of movement.

Unrestricted by a limited vocabulary of steps and positions and thus able to draw instead on a whole range of possibilities in human movement, choreographers could use their freedom and personal temperament to select the range of movements most appealing to them as individuals. Thus some dealt with the sombre, introverted aspects of human nature, others with man's more outgoing, social nature; some were happiest identifying with past times and distant cultures, others felt only for the concerns of the present. The concept of what was personal and authentic in experience naturally differed with each artist. Modern dance became a multifaceted form, but a form with a cohesion rooted in a conviction about the function of dance as an art capable of expressing thoughts and feelings relevant to contemporary society.

HISTORICAL DEVELOPMENT

Modern dance, as here understood, began in two areas, Europe and the United States. The Europeans tended to abstract concepts of the body in relation to the space surrounding it, the Americans to turn for inspiration to cultures that they felt were superior to their own. Consequently two modern dance movements developed almost simultaneously but practically independently of one another. They held the belief in common that dance, freed of its restrictive vocabulary, could express a personal, relevant, contemporary view of life.

Pioneers. In Europe two men had experimented with new forms of body movement, although one was concerned only with singers and actors, the other only with musicians. François Delsarte (1811–71), disturbed by the artificial mannerisms he saw on the Parisian stage, worked out a system of gesture by which the performers' movements could more naturally express the meanings behind his words, not merely adorn them. In Switzerland,

Experiments of Delsarte, Dalcroze, and Laban

Émile Jaques-Dalcroze (1865–1950) pioneered the teaching of eurythmics, the representation of musical rhythms in movement. He found that musicians could improve their sense of rhythm if trained to respond to music through movement. His exercises evolved into forms also useful for the dancer and were presented in performance-demonstrations. Both men revealed the possibilities of a freer kind of theatrical movement, and dancers began to explore its significance for their own art.

Working in Switzerland during World War I, the Hungarian dancer Rudolf Laban (1879–1958) experimented with the spatial and directional possibilities of the body's movement. With the dancer Kurt Jooss (1901–), he codified principles of physical expression based on analyses of the coordination of mind, nerves, and muscles. He sought primarily to establish the full range of human movement, believing that once the dancer mastered this vast repertory he would be free to utilize any range within it to express whatever inner motivations impelled his dance.

The dance of Wigman, Duncan, and St. Denis

Laban's pupil Mary Wigman (1886–), who also had studied with Dalcroze, saw more forcefully and successfully the dramatic use to which his theories could be put. In her work, a dancer related his person to a surrounding space that was either hostile or amiable so that space became a character and a dramatic relationship was established. Wigman's dance was an objectified telling of universal, deep-rooted experience. "Beyond the personal interpretation underlying the dancer's gesture," she wrote, "there is always the responsibility and obligation to make clear the universal, super-personal meaning—which the dancer can neither change forcibly nor exchange arbitrarily without endangering the general validity of his message." (From Mary Wigman, *The Language of Dance*, translated by Walter Sorell, Wesleyan University Press, 1966.)

In America, where the experimentation had begun shortly after the turn of the century, the foundations of modern dance were different. Isadora Duncan (1878–1927) found in the culture of ancient Greece the counterpart of her self-proclaimed devotion to the good, the true, and the beautiful. In antique sculpture she found representations of the kind of dance she wanted to see and to perform: the beautiful, natural body clothed in loose robes and with bare feet, its movements echoing the swaying of trees, the ebb and flow of the waves of the sea. She was inspired too by music, especially that of Beethoven, Schubert, and Chopin. She sought a dance "that might be the divine expression of the human spirit through the medium of the body's movement." She would not teach technique as the ballet taught it, believing that it only disfigured the natural grace of the body. She told her pupils:

> Listen to the music with your soul. Now, while listening, do you not feel an inner self awakening deep within you—that it is by its strength that your head is lifted, that your arms are raised, that you are walking slowly toward the light? (From Isadora Duncan, *My Life*, Garden City Publishing Co., Inc., 1927.)

From her ideas came a basic characteristic of the American modern dance: movement emanating from individual feeling. As she said in the same source:

> My art is just an effort to express the truth of my Being in gesture and movement. . . . I have given [the public] the most secret impulses of my soul. From the first I have only danced my life.

Unfortunately the genius of Duncan fell into disrepute for a long period through the teachings of many of her followers, who tended to emphasize pure, unstructured feeling at the expense of carefully articulated communication through movement, an uncomfortably close approach to the earlier interpretive dance from which Duncan had raised the art.

Meanwhile another young American was longing "to translate into rhythmic patterns a spiritual significance." The motivation of Ruth St. Denis (1877–1968) was personal, an "inward impulse" she called it, but she felt impelled to represent her feeling more objectively than Duncan had. Dance, she said, "can be expressive of the whole range of thought and philosophy." The civilizations of the East provided her with concepts of spirituality that she found lacking in America; as well, they provided her with themes, characters, and episodes and with ideas for costumes and settings. She never pretended that her works were authentic reconstructions of Eastern styles, only that they were based on those styles. But the East provided her with vast resources for ideas of movement for which she could feel a personal affinity.

The late 1920s and after. Thus audiences were gradually exposed to new forms of movement. The negative aspects of the freedom opened by Delsarte, Dalcroze, and Laban were becoming focussed in positive directions: for each it was a personal search for a means to communicate significant values. In a sense, however, it was still escape. Seeking to ennoble the dance, the innovators saw little of dignity or stature in their immediate surroundings, and for all their claims to revolution, much of romanticism remained in their work. Duncan echoed the idealism of Walt Whitman and danced to the heroic strains of Tchaikovsky's *Marche Slave*; St. Denis portrayed goddesses in *Egypta* and *Radha*, while her partner-husband Ted Shawn told a tale of ancient Mexico in *Xochitl* and mourned the fate of the American Indian in *The Feather of the Dawn*. Wigman was not without her mystical strain: she had her demonic *Witch Dance* and her ecstatic *Seraphic Song*.

Where these early innovators had looked to the glory of Greece and the mysticism of the East, to the romanticism of Whitman and Schubert, the next generation was to become more fully aware of its own present, of the writings of Freud and T.S. Eliot, of the music of Bartók and the painting of Picasso. In addition, society was becoming aware of Hitler, who banished modern dance from Germany, and of the rising labour agitation that reached its peak in the depression of the 1930s. Such consciousness, in dancers and in audiences, would breed a choreography far different from that of Duncan and St. Denis.

Second generation: Graham, Humphrey, and Limón

Modern dance was still concerned with externalizing personal experience: Martha Graham (1893–) called it "making visible the interior landscape." But the exponents of the late 1920s and the 1930s had new ideas about what kind of experience was important. Rejecting not only the ballet but also Duncan and her own mentor St. Denis, Graham wrote:

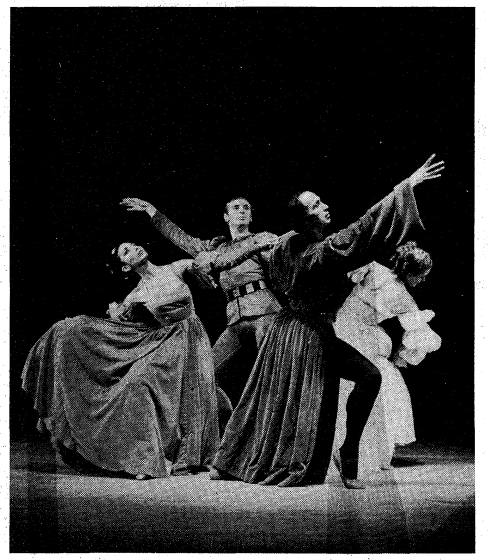
> I do not want to be a tree, a flower, or a wave. In a dancer's body, we as audience must see *ourselves*, not the imitated behavior of everyday actions, not the phenomena of nature, not exotic creatures from another planet, but something of the miracle that is a human being, motivated, disciplined, concentrated. (In Frederick Rand Rogers, ed., *Dance: A Basic Educational Technique*; The Macmillan Co., 1941.)

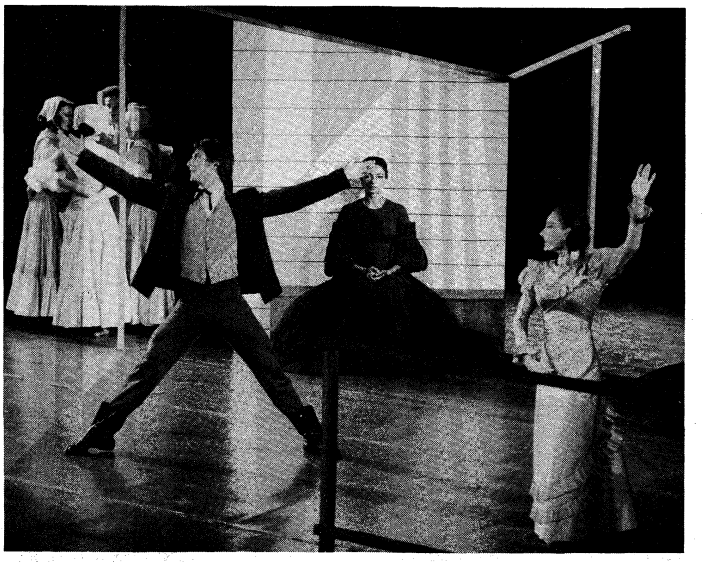
Doris Humphrey (1895–1958) chose another way to break away from the bases of her training in the company of St. Denis and Shawn. "I knew something about how the Japanese moved, how the Chinese or Spanish moved, but I didn't know how I moved or what the American heritage should be." (From Doris Humphrey, "New Dance: An Unfinished Autobiography," in *Dance Perspectives 25*, 1966.) In seeking her identity, she also found her own way of dance: It

> is composed as an expression of American life as I see it today. . . . I believe that the dancer belongs to his time and place and that he can only express that which passes through or close to his experience. The one indispensable quality in a work of art is a consistent point of view related to the times. (From Doris Humphrey in *Dance: A Basic Educational Technique*.)

The vocabulary of ballet was unsuited to the communication of ideas "related to the times" but so were the idioms of the first innovators. If form was to follow function, as modern dancers had proclaimed, still other kinds of movement had to be found to embody the new directions.

Graham asserted that she had found a guiding principle for her work in the act of breathing. The body contracted in inhalation, released in exhalation, and by intensifying the act, the dancer could use the breath as a force to send him into a turn, a leap, a fall. The movement could be sharp and spasmodic, signifying agony or terror or ecstasy, or it could be controlled and lyrical. For her, movement spoke always of basic passions: it was at once savagely of the ground but aspiring to the sky. "Your spine,"





Modern dance of the early 20th century.
(Left) *The Moor's Pavane* by José Limón; music by Henry Purcell, arranged by Simon Sadoff; costumes, Pauline Lawrence; José Limón Dance Company, (left to right) Pauline Koner, Lucas Hoving, José Limón, and Betty Jones. (Right) *Appalachian Spring* by Martha Graham; music, Aaron Copland; costumes, Edythe Gilford; decor, Isamu Noguchi; Martha Graham Dance Company.
(Left) Jack Mitchell, (right) Martha Swope

she told her students, "is the line connecting heaven and earth."

Where Graham saw the struggle within the person, Humphrey looked outward—to the environment and the society that surrounded the dancer. She saw movement as generated by efforts to resist the pull of gravity, of those forces that threaten man's security. The dancer fell, crushed by hostile antagonists, but always rose again, sustained by his will to persist and conquer. Her approach was inherently dramatic, for it was based on a fundamental human conflict: the longing for progress and change pitted against the desire for stability. She believed that without the daring, there is no action, no life, and "Dance," she said, "takes place within the arc between two deaths."

Changes in the approach to dance did not necessarily involve a shift to contemporary subject matters, and in fact such subjects were rarely used. (Notable exceptions were Graham's *Immediate Tragedy*, which dealt with the Spanish Civil War, and Charles Weidman's *Lynch Town*.) But the point of view was contemporary. Graham saw social cohesiveness founded on the act of ceremony in *Primitive Mysteries*. She approached the legends of Greece with Freudian insight in a great cycle that was climaxed with *Clytemnestra* in 1958. Humphrey commented on democracy through the complex relations of group and individual in *New Dance* and on social injustice through her treatment of an episode from 19th-century England in *Inquest*.

Grief for contemporary turmoil made some choreographers look to the past, not with nostalgia but with a view to its relevance for their own problems. Graham recalled the courage of pioneers in *Appalachian Spring*; Humphrey evoked the fervour of a religious sect in *The Shakers*. Such subject matter was not romanticized: the past was viewed both sympathetically and critically.

When modern choreographers found that they had made their point about the capacity of dance to express the darker, more ruthless part of man's nature, they felt able to deal with his nobler aspects without harming their own cause, but they never were quite able to return to the sanguine views of the earlier years. José Limón (1908–72) wrote that he tried to compose works that were involved with

> man's basic tragedy and the grandeur of his spirit. . . . I reach for demons, saints, martyrs, apostates, fools and other impassioned visions. (S.J. Cohen [ed.], *The Modern Dance*, pp. 23–24.)

He turned to Shakespeare's tormented lovers in *The Moor's Pavane* and to heroism in wartime in *Missa Brevis*. Even Graham, while continuing her concern with man's animal drives, created in 1948 *Diversion of Angels* about the joy and sadness of being in love. A contemporary and unsentimental lyricism had evolved.

Broadening subject and technical repertoires

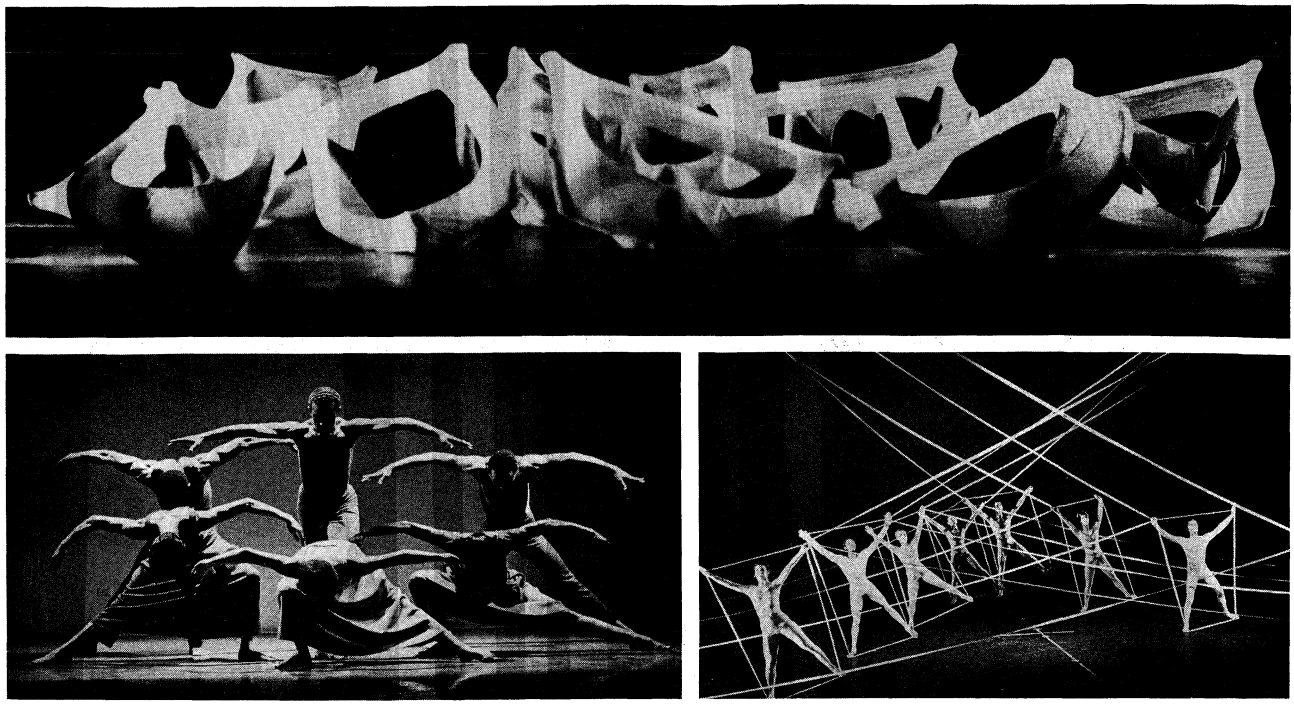
An expanded range of subject matter required a correspondingly larger repertoire of technical means in terms of movement and theatrical spectacle. The early modern dance had rejected the ballet vocabulary, but the ballet itself had entered an era of creative rejuvenation in the early decades of the century. Modern dance choreographers began to adapt certain of its elements as they found it possible to accommodate some of the ideas it represented. Technical skills, it was felt, were not necessarily inexpressive; they had only been made so by having been exploited for their own sake, and they could as well depict important aspects of personal experience—youth, joy, dreams. In time some choreographers reinstated costume and decor and even romantic music. These elements were used, however, in fresh ways: never decoratively, always for the sake of expression. Modern dance, while changing, retained its identity.

Renewed searches for new forms. But as modern dance seemed to become richer, some critics in fact began to find it poorer. Though the repertory of technique had been expanded, it was still limited by the dictum that the movement of the dancer must be meaningful. The dance, it was felt, was becoming too literary in its concerns, and bodily movement was a means to an end rather than an end in itself. At the same time, many dancers were undertaking fresh experiments with movement, not for its balletic qualities of brilliance but for its values of pure shape, rhythm, and texture. To develop dance as an independent art, free of drama, free of music, beholden to nothing but itself, became a new avenue for exploration.

Experiments of Cunningham and Nikolais

Choreographers who felt thus had to search for new movements and new ways of combining them. Because they rejected dramatic motivation, they had to look elsewhere for their initiating impulse. The acknowledged leader of the group, Merce Cunningham (1919–), wrote:

> I start with a step. . . . This is not beginning with an idea that concerns character or story, a fait accompli around which the actions are grouped for reference purpose. I start with the movement, even something moving rather than someone (pillows in the air). . . . But then out of this the action begins to assume its own proportions, and other possibilities appear as the dance proceeds. New situations present themselves—between the dancers and the space, the space and



Experimental modern dance of the later 20th century.
(Top) *Sanctum* and (Bottom right) *Tensile Involvement.* Choreography, sound score, costumes, decor, and lighting by Alwin Nikolais; Alwin Nikolais Dance Company. (Bottom left) *Revelations* by Alvin Ailey; music, traditional Negro music; costumes and decor, Ves Harper; the Alvin Ailey American Dance Theatre.
(Top) Susan Shiff Faludi—Three Lions, (bottom left) Jack Mitchell, (bottom right) Fred Fehl

the time. It can take a momentum of its own. . . . I have a feeling it produces some kind of atmosphere. (Merce Cunningham *et al.*, "Time To Walk in Space," in *Dance Perspectives*, 34, 1968.)

To ensure independence of personal choice or subjective motivation, Cunningham at times introduced chance elements into his work, including improvisation and dances performed in random sequences, to discover new patterns and combinations of movement; at other times he used elaborate mathematical charting techniques.

Other devices have been employed by Alwin Nikolais (1912–). Granting that movement is naturally expressive, he tried taking it out of the context with which it is normally associated. He would, for example, in the instance of a man reaching for a box, remove both the box and the man's purpose. This allows the spectator to look only at the quality of the action of reaching. Analyzing the movement's qualities, he would combine it with other movements quite unrelated to it emotionally. To assure that the audience saw the movements for their own sake without reference to any potentially dramatic situation, he characteristically clothed his dancers in a variety of unrevealing shapes or sacks, often attaching them to inanimate objects to alter further their configuration in space. He called the result not dehumanization but depersonalization. His works were not dances, he asserted, but rather "theatre pieces of light, color, motion, and sound." He is thus characteristic of the newer generation of choreographers who have readmitted scenic values and technological potentialities into modern dance.

Other choreographers, notably Ann Halprin and Yvonne Rainer, worked with the idea of what were called "task movements." Each performer was given a task—*e.g.*, carrying a load of objects along a specified path—and the motions resulting from the execution of the task constituted the dance. They and others, such as Meredith Monk and Twyla Tharp, let their dances develop as responses to the environments, utilizing whatever space happened to be available—an expanse of grass, a museum, a gymnasium.

Modern dance continued, however, to be a many-faceted art. Holding to the original concept of representing personal experience, some choreographers continued to create in the dramatic vein. Anna Sokolow and Pauline Koner persisted in the belief that dance is an embodiment of human values, a representation of human actions and passions. Others, like Paul Taylor, who began as a member of the Martha Graham company, turned to dramatic dance after experimenting with nonrepresentational forms. Still others, such as Alvin Ailey and Donald McKayle, borrowed elements of ethnic dance from Africa and the West Indies, combining them with modern dance techniques to develop individual styles of expressive jazz. Their work differed from that of Pearl Primus and Katherine Dunham, both of whom performed a theatricalized "ethnic dance" quite apart from the mainstream of modern dance.

Interactions with ballet and musical comedy

Since influences upon these modes of choreography frequently came from such other art forms as painting, music, and theatre, the ballet was open to them as well. Some of the works created by George Balanchine for the New York City Ballet were as abstract in structure as those of Cunningham. The original opposition of the classical ballet and modern dance broke down in this way, and distinctions between the forms blurred for other reasons as well. As modern dance incorporated technical elements from the ballet, the reverse also occurred. Some choreographers, such as Antony Tudor and Jerome Robbins, although trained within the ballet idiom and eager to use its repertory of physical skills, wished to make dances that were serious, expressive, and relevant to their own time. As their purpose approached that of the modern dance, they assimilated elements from its store of techniques.

Further influences resulted from the use of dance by the musical theatre and films that employed both ballet and modern dance choreographers. Among the most notable of these were Agnes deMille, whose choreography of *Oklahoma!* (1943) significantly altered the face of the musical comedy; Helen Tamiris, who choreographed the dances in *Annie Get Your Gun*; Hanya Holm, a student of Mary Wigman, who did *My Fair Lady*; and Robbins with *West Side Story*. The attitude of modern dance that allows character and situation to dictate movement was found most appropriate to the musical stage and film.

Modern dance today. By the middle of the 20th century modern dance was an almost exclusively American form. Wigman and others toured or worked in the United States following the German political suppressions

of the 1930s. After World War II the void in European dance was filled by the resurgence of ballet as an authentic form. In England modern dance was restricted to educational institutions, where adaptations of Laban's techniques were not of a serious artistic calibre. The work of the Martha Graham school in London—the London School of Contemporary Dance, founded in the late 1960s—indicated that American influences might initiate a fresh beginning.

Companies and festivals

Even in America, by the 1970s modern dance companies were existing precariously, many performing only a few months each year. There was no central organization for, or cooperation among, these companies, and only two dance festivals were well established. The Jacob's Pillow Dance Festival in Lee, Massachusetts, was guided by Ted Shawn from the early 1940s; it staged a repertory of ballet and ethnic and modern dance. The only festival devoted entirely to modern dance was begun in the late 1940s at Connecticut College in New London.

After the mid-1950s, despite the work of Taylor, Ailey, and others, the trend in modern dance was toward the abstract. At the beginning of this period of nonrepresentational dance, the concern was to draw attention to movement for its own sake. After the first decade or so some exponents began to draw on such other idioms as eurythmy, the interpretation in movement of a spoken or sung accompaniment, and the film, slide, and light projections characteristic of such contemporaneous multimedia forms as the rock-and-roll discotheque, the seemingly improvised "happenings" that were carefully prepared for presentation on city sidewalks or college campuses, and similar phenomena associated with "pop art" culture. Practitioners of other arts such as the painter Robert Rauschenberg began experiments with choreography that utilized their proved skills in combination with dance. Experiments in linking art and engineering drew electrical technicians into the dance theatre. New art forms combined with the new technology to represent an integral part of contemporary man's experience. By the early 1970s there were signs of another kind of development. Growing protest of social and political origin found its way into dance works and started a trend, once again, away from the abstract and toward a more emotionally, if not dramatically, oriented dance theatre. With actors like James Cunningham turning to choreography, at least one facet of the modern dance concerned itself with arousing the emotions and intellect of its audiences. The modern dance, attempting to relate to its time, appeared headed again toward growing social involvement.

BIBLIOGRAPHY. JOHN MARTIN, *Modern Dance* (1933), the most detailed analysis of theory up to that time by the critic most responsible for bringing modern dance to the public; and *Introduction to the Dance* (1939), an analysis of all forms of dance, but especially good on the theory of the early modern dance; M. LLOYD, *Borzoi Book of Modern Dance* (1949), the most complete survey of the early years; S.J. COHEN (ed.), *The Modern Dance: Seven Statements of Belief* (1966), essays by choreographers Erick Hawkins, Pauline Koner, José Limón, Donald McKayle, Alwin Nikolais, Anna Sokolow, and Paul Taylor; W. SORELL (ed.), *The Dance Has Many Faces*, 2nd ed. (1966), a collection of essays on many forms of dance, including some excellent ones on the various aspects of the modern dance; MERCE CUNNINGHAM, *Changes: Notes on Choreography* (1969), random but very revealing and significant thoughts on experimental compositions; and (with others), "Time to Walk in Space," *Dance Perspectives 34* (1968), essays by critics and by Carolyn Brown, leading dancer in the Cunningham company; M.B. SIEGEL, "Dancers' Notes," *Dance Perspectives 38* (1969), essays by modern dance choreographers Remy Charlip, Judith Dunn, Lucas Hoving, Katherine Litz, Murray Louis, Jack Moore, Daniel Nagrin, and Don Redlich; F.R. ROGERS (ed.), *Dance: A Basic Educational Technique* (1941), essays on ideas and techniques by modern dance choreographers; M. WIGMAN, *Die Sprache des Tanzes* (1963; Eng. trans. by W. SORELL, *The Language of Dance*, 1966), reflections on the creative process involved in some of the choreographer's major dance works; I. DUNCAN, *The Art of the Dance* (1928, reissued 1970), a collection of very important writings in the innovator's aesthetics.

Pertinent information on modern dance may also be found in the following autobiographical and biographical writings: ISADORA DUNCAN, *My Life* (1927; reissued as *Isadora*, 1968), not very accurate factually, but important for its statements of attitudes and beliefs; A.R. MACDOUGALL, *Isadora: A Revolutionary in Art and Love* (1960), the most complete and accurate biography; L. LEATHERMAN, *Martha Graham: Portrait of the Lady as an Artist* (1966), observations on the choreographer at work; M. ARMITAGE (ed.), *Martha Graham* (1937, reissued 1966), an excellent collection of early analytical essays by critics; DORIS HUMPHREY, *The Art of Making Dances* (1959), the author's own analysis of the craft of choreography; and "New Dance: An Unfinished Autobiography," *Dance Perspectives 25* (1966), traces her early career to 1928, chronology included; RUTH ST. DENIS, *An Unfinished Life* (1939), a detailed autobiographical account of events and reflections on art and life; WALTER TERRY, *Miss Ruth* (1969), a biography of Ruth St. Denis by a close and long-time friend; D. MCDONAGH, *The Rise and Fall and Rise of Modern Dance* (1970), includes some most informative interviews with contemporary avant-garde choreographers.

(S.J.C.)

Modes, Musical

In musical theory and analysis, the term mode has been employed in a variety of more or less correlated meanings. Mode designates particular aspects of melodic construction; its meaning ranges from strictly defined scales whose identity stems from the position of given intervals (distances between notes) in the sequence of tones (see SCALES, MUSICAL) to broadly delineated melody types and constellations of motivic figures (brief, identifiable melodic fragments) that may serve as bases for improvised or written compositions. Throughout the history of Western and non-Western art music, various modal systems have been formulated by theorists. The use of modal concepts in the analysis of the vast bodies of Western and non-Western folk music, including that of tribal cultures, substantially advances the understanding of this music.

This article covers the ancient Greek modes, Jewish liturgical modes and reciting formulas, Eastern and Western church modes, and it briefly touches upon non-Western modal systems (see also ISLAMIC PEOPLES, ARTS OF: *Music*; MUSIC, EAST ASIAN; SOUTH ASIAN PEOPLES, ARTS OF: *Music*; SOUTHEAST ASIAN PEOPLES, ARTS OF: *Music*). The term rhythmic mode refers to the 13th-century system of simple rhythmic patterns in three-beat, or ternary, metre (see MUSICAL RHYTHM) and, occasionally, to longer rhythmic patterns in non-Western music.

Ancient Greek modes. The modes of Greek antiquity were placed by theorists in orderly fashion within a larger context. Although the modes were a series of seven-note diatonic scales (*i.e.*, containing five whole tones and two semitones), the nucleus of the tone system was the tetrachord—a group of four consecutive notes (as, from C to F on the piano) comprising the interval of a fourth. Except in late antiquity, the notes were always arranged in a descending order, the basic tetrachord consisting of two whole tones and one semitone: E–D–C–B. Two such tetrachords, separated from one another by a whole tone, formed the so-called Greek Dorian mode: E–D–C–B A–G–F–E. The Dorian mode was taken as a basis for the construction of the larger system. Its single-octave range was extended by the addition of a third tetrachord, A–G–F–E, on top and of a fourth tetrachord, E–D–C–B, at the bottom. In contrast to the two inner tetrachords, which were separated by a whole tone (disjunct tetrachords), each outer tetrachord was linked with the neighbouring inner one by a shared note (conjunct tetrachords):

A G F E D C B A G F E D C B.

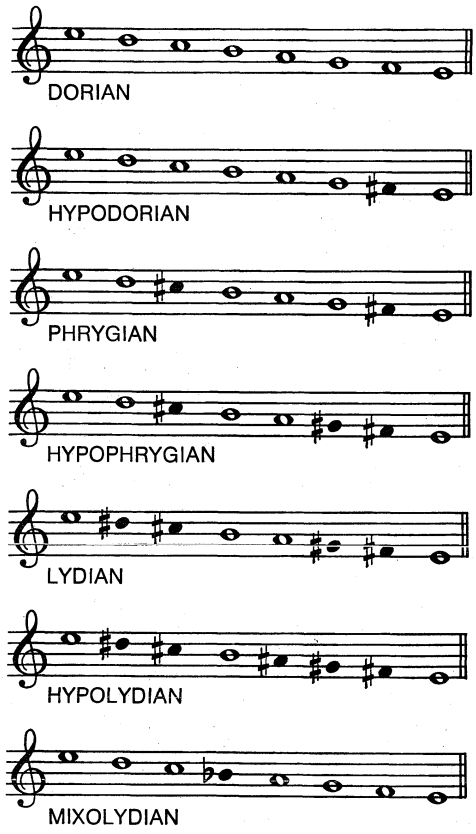
Because the combination of the four tetrachords yielded a range of two octaves minus one whole tone, a low A was added by theorists to achieve the following diatonic two-octave system: A G F E D C B A G F E D C B A. This two-octave row, or disdiapason, was called the Greater Perfect System. It was analyzed as consisting of seven overlapping scales, or octave species, called *harmoniai*, characterized by the different positions of their semitones. They were termed as follows (semitones shown by unspaced letters):

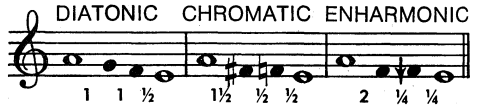
The Greater Perfect System

A G FE D CB A	Hypodorian
G FE D CB A G	Hypophrygian

FE D CB A G F	Hypolydian
E D CB A G FE	Dorian
D CB A G FE D	Phrygian
CB A G FE D C	Lydian
B A G FE D CB	Mixolydian

Although the names of the *harmoniai* were identical with those of the Greek modes, the *harmoniai* represented not the modes themselves but were instead projections of the interval patterns of the modes into the more extensive Greater Perfect System. The modes proper were termed *tonoi*, their essence being their interval pattern. On the kithara or *lyra* (the two basic plucked string instruments of ancient Greece) the *tonoi* were produced either by the basic tuning or by the raising or lowering of one or more of the strings by a semitone. Thus, the seven *tonoi* would sound within the octave E–E as follows (black notes indicate changes from the Dorian tuning):



Greek theory distinguished three different genera of tetrachords, producing an additional variety of modes. The previously described tetrachord (two descending whole tones plus one semitone) was called diatonic. There were also chromatic and enharmonic genera. The two tones bounding the tetrachord were fixed and always formed a perfect fourth; the two inner tones were movable. The chromatic tetrachord consisted of a minor third (encompassing 1½ whole tones) plus two semitones, the enharmonic tetrachord of a major third (encompassing two whole tones) plus two approximate quarter tones:

Also prominent in Greek music was the concept of ethos, which ascribed certain ethical characteristics to the different modes. The Dorian mode was preferred because of its strong and virile character; the Phrygian mode was ecstatic and emotional, the Lydian mode intimate and lascivious. In *The Republic* Plato stressed the educational values of the Dorian mode and warned against the softening influence of the Lydian mode.

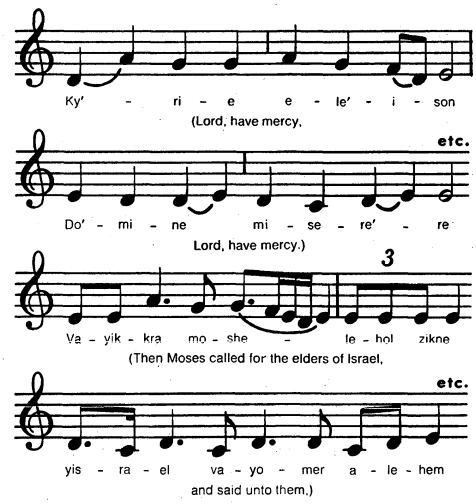
In early Greek antiquity a system of modal categories developed, referred to as *nomoi* (singular, *nomos*, "law"). The *nomoi* represented modes in that they were characterized by distinctive melodic formulas suited to different song types. The performers were free to improvise within the boundaries of these modal formulas.

Jewish and Eastern Christian chant. Ancient Hebrew music followed well-established modal patterns. According to Abraham Zevi Idelsohn, a musicologist whose comparative research in the early decades of this century established modern understanding of the Hebrew modes,

> A mode . . . is composed of a number of motives (*i.e.* short music figures or groups of tones) within a certain scale. The motives have different functions. There are beginning and concluding motives, and motives of conjunctive and disjunctive [*i.e.*, convergent and divergent] character. The composer operates with the material of these traditional folk motives within a certain mode for his creations. His composition is nothing but his arrangement and combination of this limited number of motives. His "freedom" of creation consists further in embellishments and in modulations from one mode to the other.

The modal Hebrew music strongly influenced early Christian chant. This correlation can be illustrated by comparing a Gregorian chant Kyrie, in the third mode, with a Babylonian Jewish melody for a phrase from Exodus:

Reprinted from *Music in the Middle Ages* by Gustave Reese. Copyright 1940 by W.W. Norton & Company, Inc. Copyright renewed 1968 by Gustave Reese. By permission of W.W. Norton & Company and J.W. Dent & Sons Ltd.



Eastern church chants

Syria played an important part in developing early Christian chant by integrating both Hellenistic and Hebrew elements. The Syrians devised a musical system called *oktōēchos*, a term suggesting a classification into eight *ēchoi*. The Syrian *ēchoi* are modes, although there is no consensus on whether they represented modes in a specifically technical sense, comparable to the Greek *tonoi*, or melodic formulas, comparable to the Greek *nomoi*.

Byzantine chant molded the features of early Christianity with Hellenic and Oriental traits, including the Syrian *oktōēchos*, and achieved a brilliant and distinctive style that served as a prototype for the chant of the Greek Orthodox Church. The eight *ēchoi* of the Byzantine *oktōēchos* were divided into four authentic and four plagal (derived) forms. The most common classification of the Byzantine modes was in terms of typical initial and final notes of melodies in a given mode, with the characteristic distinctions as follows (the orderly progression of notes in each series should be observed).

	Ēchos	Initial Note	Terminal Note
Authentic	I	a′	a′ or d′
	II	b′ or g′	e′ or b′
	III	c″ or a′	f′ or c″
	IV	d″ or g′	g′ or d″
Plagal	I	d′ or g′	d′
	II	e′ or g′	e′
	III	f′ or a′	f′
	IV	g′, a′, or c″	g′

The above classification reflects only two of various characteristics (not all completely clarified by modern scholars) that gave the modes their identity.

Even before the foundation of the Byzantine Empire, Armenia adopted Christianity as a state religion (AD 303). Although the early Armenian chant did not survive, the arrangement of the hymns of the Armenian Church in the comprehensive collection, *Sharakan*, indicates that the Armenians used an *oktōēchos* classification the modal characteristics of which seem to have been defined by melodic formulas rather than by scalar distinctions.

Another variety of the *oktōēchos* occurs in Russian Church chant. Although a concept of eight *ēchoi* points to the Byzantine system, the Russian *ēchoi* show a different structure. The melodic motives characteristic of the *ēchoi* are called *popievki*; but similar *popievki* could be employed in more than one *ēchos*. The use of some *popievki* is limited to the beginning, the middle, or the end of a chant. Occasionally, two *popievki* are merged into a compound *popievka*.

Gregorian chant. Gregorian chant, or plainsong, which forms the core of the musical repertoire of the Roman Catholic Church, consists of some 3,000 melodies collected and organized during the reigns of several 6th- and 7th-century popes. Most instrumental in codifying these chants was Pope Gregory I.

Melodically, Gregorian chants are based on eight different modes, often called church modes. Seven of them were given names identical with those used in the musical theory of ancient Greece: Dorian, Hypodorian, Phrygian, Hypophrygian, Lydian, Hypolydian, and Mixolydian, while the name of the eighth mode, Hypomixolydian, was adapted from the Greek. Each mode comprises a diatonic scale with the compass of one octave. The modes are classified by their *finalis*, the usual final note of a melody in that mode. Each of the four notes of the tetrachord D–E–F–G serves as the *finalis* of an "authentic" mode (see chart below).

Structure of Gregorian modes

An authentic mode consists of a pentachord (a succession of five diatonic notes) followed by a conjunct tetrachord, for example:

D E F G A B C D.

But the tetrachord may be added below rather than above the pentachord, thus generating a "plagal" mode:

A B C D E F G A.

In either case the *finalis* falls on the lowest note of its pentachord. Each authentic mode has a correlated plagal mode, which is identified by the prefix *Hypo*. In the following chart of the eight church modes the *finalis* is marked by a capital letter:

1. Dorian		D e f g a b c d
2. Hypodorian	a b c	D e f g a
3. Phrygian		E f g a b c d e
4. Hypophrygian	b c d	E f g a b
5. Lydian		F g a b c d e f
6. Hypolydian	c d e	F g a b c
7. Mixolydian		G a b c d e f g
8. Hypomixolydian	d e f	G a b c d

The tones of the Hypomixolydian mode are identical with those of the Dorian, but the two modes differ in the location of their *finalis*. The character of the church modes was further determined by a number of distinctive melodic formulas, and sometimes a particular ethos was attributed to the different modes.

Contrary to the Byzantine classification, which lists first the four authentic and then the four plagal modes, the Roman classification alternates the authentic and the plagal modes, so that modes with similar *finalis* follow each other. This principle underlies the medieval fourfold system of the so-called *maneriae* (Latin: "manners"), a division of the modes into four pairs. The first pair, or *protus maneria*, includes the Dorian and Hypodorian modes; the second, or *deuterus*, the Phrygian and Hypophrygian; the third, or *tritus*, the Lydian and Hypolydian; and the fourth, or *tetrardus*, the Mixolydian and Hypomixolydian modes.

Although Greek names were sometimes applied to the church modes and the principle of diatonic octave scales is found in both systems, certain significant discrepancies seem to belie any direct historical connection. Most conspicuous is the different meaning attributed to the names of the Greek octave species and of the church modes. Comparing the two systems provides a plausible explanation: medieval theorists apparently assumed wrongly that the Greek octave species were named in ascending rather than descending order. The Greek octave species Dorian (E–E), Phrygian (D–D), Lydian (C–C), and Mixolydian (B–B) thus appeared in the church modes as Dorian (D–D), Phrygian (E–E), Lydian (F–F), and Mixolydian (G–G).

Gradual emergence of major and minor tonality

The strict consistency of the system of church modes was gradually weakened by the appearance of B♭ in addition to B♮, although the two notes never occurred in succession. The main reason for the use of a tone not included in the basic scale pattern was that medieval musicians sought to avoid the tritone F–B. The tritone (so called because it includes three whole tones) was considered an undesirable interval sharply contrasting with the perfect fourth F–B♭. The substitution of B♭ for B♮ changed the character of a mode. For example, the Lydian mode with a flattened B was identical with the modern major mode, specifically, with the F-major scale (F G A B♭ C D E F); and the Dorian mode with a flattened B generated a minor mode corresponding to the natural D minor scale (D E F G A B♭ C D).

Nevertheless, for centuries medieval theorists considered these alterations as special forms of the Lydian or Dorian mode rather than as new modes. The reluctance to acknowledge the existence of additional modes is reflected in the so-called musica ficta. According to this practice, musical notation appears to conform strictly to the system of church modes but presupposes that the performer makes certain adjustments by raising or lowering a note through the insertion of a sharp or flat.

Two different developments occurring between the 12th and the 16th centuries resulted in a radical change in modal theory: an infiltration of folk music into the ecclesiastical and secular art forms and the steadily evolving fabric of harmony destined to unify the growing complexity of polyphonic (many-voiced) musical texture. Finally, a theorist, Heinrich Loris, commonly known by his assumed name Henricus Glareanus, sanctioned the coexistence between the old church modes and the emerging major and minor modes. In his *Dodecachordon* (1547; from Greek *dōdeka*, "twelve," and *chordē*, "string"), perhaps the most significant musical treatise of his time, Glareanus enlarged the system of the eight church modes by adding the following four:

9. Aeolian		A b c d e f g a
10. Hypoaeolian	e f g	A b c d e
11. Ionian		C d e f g a b c
12. Hypoionian	g a b	C d e f g

The Ionian and Hypoionian correspond to the major mode, the Aeolian and Hypoaeolian to the "natural" minor mode. The 12 modes of the *Dodecachordon* comprise authentic and plagal structures with tonal centres on C, D, E, F, G, and A, without recourse to sharps or flats. Glareanus mentions another two modes: the Locrian and the Hypolocrian, having B as their tonal centre. But because in these two modes B and the fifth degree above it, F, form a "false" (*i.e.*, diminished, or flattened) fifth (another form of the tritone), Glareanus states that for practical purposes only 12 modes are available.

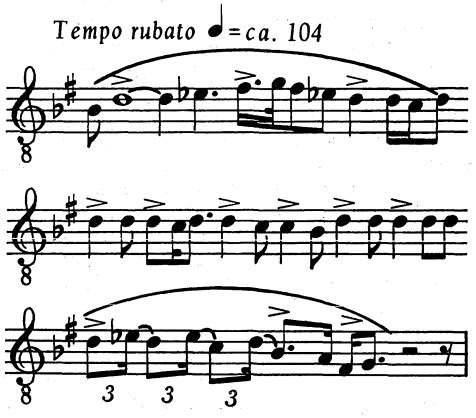
The growing complexity of polyphonic music caused the distinction between authentic and plagal modes to become more and more irrelevant, and as a result, the number of modes was virtually reduced to six. The further development of Western art music is characterized by the gradual abandonment of the old church modes in favour of the dual major–minor system that dominated 18th- and 19th-century harmony. This system is often termed tonal, in contradistinction to that of the church modes; in fact, some 20th-century works reviving the patterns of the old church modes, as well as folksongs that occasionally use them, are often termed modal. Nevertheless, major and minor have all essential characteristics of modes and should therefore be evaluated as such.

Non-Western modes and melody types. Modal concepts permeate all manifestations of Oriental art music. They represent a synthesis of well-established systems both of scalar construction and of a variety of melodic formulas and motivic configurations. The amalgamation of these elements yields specific melody types that are imbued with ethical, emotional, and cosmological connotations and form the basis of musical creation. Whereas the Western composer strives for originality and individual expression, his Eastern counterpart submits to the limitations of existing melody types, seeking to excel in an exploitation of their potentials.

The *rāga* and the *maqām*

The most elaborate modal concepts are those of the Indian *rāga* and of the Arabian–Persian *maqām*. They correspond to the Javanese *patet*, the Jewish *nigun*, and the aforementioned Greek *nomos*. The Oriental musician first exposes the characteristic features of a particular melody type in a free, improvisatory prelude leading to the more formalized section of the composition. These improvised introductions developed into a specific art form, called in India *ālāpa* ("conversation"), in Java *bebooka* ("guide"), in Iran *mukhtaṣarī* ("summary"), in Egypt and Turkey *taqsīm* ("dissection, analysis"), and in northwestern Africa *istahbar* ("information"). The following example (recorded by G. Schünemann, transcribed by R. Lachmann)—the first of four phrases constituting an introduction to a Tunisian art song composed in the *maqām raṣd ad-dīl*—illustrates this aspect of modal exposure.

From Robert Lachmann, *Musik des Orients* (1929); Ferdinand Hirt



*Rāga*s and *maqāmāt* embrace a wide variety of emotional content. They can be courageous, amorous, melancholy, cheerful, soothing, or ecstatic and are capable of conveying these qualities to the listener. Some melody types are supposed to influence illnesses, calm storms, tame wild animals, and increase the fertility of the soil. Others have pernicious power; for example, a certain *maqām* attracts evil spirits, while a certain *rāga* may cause fire. According to an ancient legend, a singer who performed such a *rāga* under the reign of Emperor Akbar (1556–1605) was reduced to ashes although he had submerged his body in the Ganges River. Cosmological connotations have caused melody types to be assigned to different parts of the day; for example, a mid-morning *rāga* may not be performed during the evening. The Westernized younger generation, however, tends to disregard these traditional restrictions.

The music of many tribal societies shows well-organized scalar and modal patterns, but they have not been explicitly formulated and systematized by the peoples involved. A comparative analysis reveals that, in spite of a great structural diversity, scalar and modal correlations exist not only among culturally unrelated tribal societies but also among tribal, Western folk, and Eastern art music. Modal systems and their unconscious counterparts thus may be said to have substantially contributed to the structure of a vast amount of the world's music.

BIBLIOGRAPHY. GUSTAVE REESE, *Music in the Middle Ages* (1940), provides a competent account and critique of a wide variety of modal concepts prevailing during the Western Middle Ages and discusses their origins and repercussions. WILLI APEL, "The Tonality," in *Gregorian Chant*, ch. 3 (1958), describes thoroughly the system of church modes. EGON WELLESZ, *Eastern Elements in Western Chant: Studies in the Early History of Ecclesiastical Music* (1947, reprinted 1967), discusses the modal structure of the melodies of Byzantine hymns and compares them with those of the Western Church. ABRAHAM Z. IDELSOHN, *Jewish Music in Its Historical Development* (1929, reprinted 1967), an authoritative work on Jewish music, contains chapters on "The Modes of the Bible" and "The Modes of the Prayers." EGON WELLESZ (ed.), *The New Oxford History of Music*, vol. 1 (1957), includes the following chapters discussing modes: ARNOLD BAKE, "The Music of India," CARL H. KRAELING and LUCETTA MOWRY, "Music in the Bible," ERIC WERNER, "The Music of Post-Biblical Judaism," ISOBEL HENDERSON, "Ancient Greek Music," and HENRY G. FARMER, "The Music of Islam." MIECZYSLAW KOLINSKI, "Classification of Tonal Structures, Illustrated by a Comparative Chart of American Indian, African Negro, Afro-American and English-American Structures," *Studies in Ethnomusicology*, 1:38–76 (1961), comprises a new system of modal classification accommodating a wide range of Western and non-Western modal patterns. ROBERT LACHMANN, *Musik des Orients* (1929, reprinted 1966), is the first scholarly survey of the music of Eastern cultures.

(M.Ko.)

Modigliani, Amedeo

For the importance and quality of his work, as well as for the drama of his brief life, the Italian painter and sculptor Amedeo Modigliani was one of the most significant figures in the international artistic milieu that existed in Paris in the first decades of the 20th century and that is known as the School of Paris. His work consists of a small number of stone sculptures, all executed between 1909 and 1913, about 350 paintings, and many drawings. The subjects are almost exclusively portraits and female nudes. The style has easily recognizable and relatively consistent characteristics: the use of broad expanses of colour for the plastic construction of the image, rhythmic modulation of line, reduction of the third dimension, asymmetry of composition, and elongation of the figure.

Modigliani was born in Leghorn (Livorno), Italy, on July 12, 1884, into a Jewish family of small merchants. After suffering from pleurisy and typhus in 1895 and 1898, he was forced to give up a conventional education, and it was then that he began to study painting. Because of his poor health, he spent the winter of 1901 at Naples and Capri; his letters to the painter Oscar Ghiglia, in which he writes of his artistic vocation and expounds his idea on art, began during this period. After a brief stay in Florence in 1902, he continued his artistic studies in Venice, remaining there until the winter of 1906, when he left for Paris.

In Paris, he was overwhelmed by the painting of Paul Cézanne, which exerted a decisive influence on the earliest phase of his work. His initial important contacts were with the poets André Salmon and Max Jacob, with Pablo Picasso, and—in 1907—with Dr. Paul Alexandre, a friend of the avant-garde artists and the first to become interested in Modigliani and to buy his works. In 1908 he exhibited six works at the Salon des Indépendants. For a time his interests turned toward sculpture. After the summer of 1909, which he spent with his family in Leghorn, he returned in the fall to Paris. There, through Dr. Alexandre, he met the Romanian sculptor Constantin Brancusi, by whose work he was impressed and on whose advice he made a serious study of Negro sculpture. To prepare himself for sculpture, he intensified his graphic experiments. He detested what he considered the false Impressionism of Auguste Rodin, with its pictorial modelling and susceptibility to the play of light. In his drawings he tried to give to his contours the function of limiting or enclosing the volume; in his sculpture, he carved the stone directly in an attempt to preserve in the form the plastic unity of the block. In 1912 he exhibited a "decorative ensemble" of eight stone heads at the Salon d'Automne.

He soon returned entirely to painting, but his experience as a sculptor had fundamental consequences for his style, both in the simplification and essentiality of the image and in the use of colour, which is employed as the con-

"Self-Portrait," oil on canvas by Modigliani, 1919. In the Museu de Arte Contemporânea da Universidade de São Paulo, Brazil.

By courtesy of the Museu de Arte Contemporanea da Universidade de Sao Paulo, gift of Mr. Francisco M. Sobrinho and Mrs. Yolanda Penteado; photograph, Gerson Zanini

structive material of the painting in the same way that stone is the constructive material of sculpture. By reducing and almost eliminating chiaroscuro—that is, the illusory third dimension—he achieved, by the strength of his contours and juxtaposed colours, a solidity in the flat image that is similar to that of sculpture.

Influence of Cézanne

The outbreak of war in 1914 increased the difficulties of Modigliani's life. Alexandre and other friends were at the front; his pictures did not sell; his already delicate health was deteriorating because of his poverty, feverish work, and the abuse of alcohol and drugs. He was in the midst of a troubled affair with the English poetess Beatrice Hastings, with whom he lived for two years, from 1914 to 1916. He was assisted, however, by the art dealer Paul Guillaume and especially by the devoted friendship of the Polish poet Leopold Zborowski, who bought or helped him to place a few paintings and drawings.

Modigliani was not a professional portraitist; for him the portrait was only an occasion to isolate a figure in an absolute unity of space and colour. He painted his friends, personalities of the Parisian artistic and literary world, but also unimportant people: models, servants, girls from the neighbourhood. In 1917 he began a series of large female nudes that are among his best works. In December, Berthe Weill organized a one-man show for him in her gallery, but the police judged the nudes indecent and had them removed.

His last, touching love affair began in the same year, 1917, with the young painter Jeanne Hébuterne, with whom he went to live on the Côte d'Azur. Their daughter Jeanne was born in November 1918. This was also a happy period for his painting, which became increasingly refined in line and delicate in colour. A more tranquil life and the climate of the Mediterranean, however, did not restore the artist's undermined health. After returning to Paris in May 1919, he became ill in January 1920 and ten days later, on the 24th, died of tubercular meningitis. Next day, Jeanne Hébuterne killed herself by jumping from a window.

Little known outside avant-garde Parisian circles, Modigliani seldom participated in official exhibitions, and his single one-man show was the one held in 1917 at Berthe Weill's. Fame came after his death with the 1922 exhibition at the Galerie Bernheim-Jeune and, later, with a monograph by the poet André Salmon.

Posthumous fame

His original sculptures, mostly in sandstone, are 25 in number; the number of drawings cannot be determined. Except for some 30 large female nudes (1916–19) and four landscapes (1919), his paintings are almost always portraits of relatives, artists, writers, musicians, actors, dealers, and collectors, along with many of unidentified persons. The names of the individuals portrayed may give an idea of the Montparnasse milieu frequented by the artist: Constantin Brancusi, Diego Rivera, Henri Laurens, Pablo Picasso, Chaim Soutine, Juan Gris, Max Jacob, Jean Cocteau, Jacques Lipchitz. Some, such as Paul Guillaume, Hanka and Leopold Zborowski, Beatrice Hastings, and Jeanne Hébuterne, were painted several times. Of Modigliani himself, there is only a single self-portrait, painted in 1919, shortly before his death.

MAJOR WORKS

"Head of a Young Woman" (1908; Jean Masurel Collection, Mouvaux, France); "La Juive" (1908; private collection, Paris); "Paul Alexandre" (1909; private collection, Paris); "Cariatide in piedi" (*c.* 1912; R.D. Judah Collection, London); "Seated Nude" (*c.* 1912–13; Courtauld Institute Galleries, London); "Diego Rivera" (1914; Museu de Arte, São Paulo, Brazil); "Bride and Groom" (1915; Museum of Modern Art, New York); "Paul Guillaume" (1916; Galleria d'Arte Moderna, Milan); "Beatrice Hastings con cappello" (1916; Barnes Foundation, Merion, Pennsylvania); "Jacques Lipchitz and His Wife" (1916–17; Art Institute of Chicago); "Léon Bakst" (1917; National Gallery of Art, Washington, D.C.); "Girl with Braids" (1917; E. Kahn Collection, Cambridge, Massachusetts); "Chaim Soutine" (1917; National Gallery of Art, Washington, D.C.); "Nude with Necklace" (1917; Ralph Colin Collection, New York); "Nudo sdraiato" (1917; Gianni Mattioli Collection, Milan); Hanka Zborowska seduta su un divano (1917; Museum of Modern Art, New York); "Madame Amédée (Woman with Cigarette)" (1918; National Gallery of Art, Washington, D.C.); "Anna (Hanka Zborowski)" (1918; Galleria Nazionale d'Arte Moderna, Rome); "Little Girl in Blue" (1918; private collection, Paris); "Leopold Zborowski" (*c.* 1918; Museu de Arte, São Paulo, Brazil); "Nudo in piedi (Elvira)" (1918; Walter Hadorn Collection, Bern); "Ritratto di Baranowski" (1918; Mr. and Mrs. R.J. Sainsbury Collection, London); "La giovane Lattaia" (1918; Mr. and Mrs. George Friedland Collection, Merion, Pennsylvania); "Jeanne Hebuterne in camicia" (1918; Nate B. and Frances Spingold Collection, New York); "Gypsy Woman with Baby" (1919; National Gallery of Art, Washington, D.C.); "Reclining Nude" ("Le Grand Nu," *c.* 1919; Museum of Modern Art, New York); "Yellow Sweater" *c.* 1919; Solomon R. Guggenheim Museum, New York); "The Servant Girl" (1919; Albright-Knox Art Gallery, Buffalo, N.Y.); "Self-Portrait" (1919; Museu de Arte Contemporânea da Universidade de São Paulo, Brazil).

BIBLIOGRAPHY. ANDRE SALMON, *Modigliani, sa vie et son oeuvre* (1926), the first biography of the artist, is important because it was written shortly after Modigliani's death by a poet and writer who was close to him. Salmon's *Vie passionnée de Modigliani* (1957; Eng. trans., *Modigliani: A Memoir*, 1961), is a biography in the form of a novel. ARTHUR PFANNSTIEL, *Modigliani* (1929), is a fundamental critical-biographical study with an annotated catalog of 289 works; his *Modigliani et son oeuvre, étude critique et catalogue raisonné* (1956), amplifies his earlier work, particularly in the catalog additions; Pfannstiel's *Modigliani: dessins* (1958), is a short text on the drawings, with reproductions and an annotated catalog. AMBROGIO CERONI, *Amedeo Modigliani, peintre suivi des "souvenirs" de Lunia Czechowska* (1958), is important for Modigliani's letters to his friend Czechowska and for her memoirs; *Amedeo Modigliani: dessins et sculptures avec suite du catalogue illustré des peintures* (1965), contains an important catalog and a bibliography focussing on the drawings and sculpture; Ceroni's *I dipinti di Modigliani* (1970), with an introduction by LEONE PICCIONI, is a very important critical work, with essential bibliography up to 1969, documentary photographs of the artist and of his exhibitions, chronology, an annotated catalog of his works, and biographical data on the people painted. JEANE MODIGLIANI, *Modigliani senza leggenda* (1958; Eng. trans., *Modigliani: Man and Myth*, 1958), written by Modigliani's daughter, is the most reliable and exhaustive biography and includes

letters and unedited documents. JAMES T. SOBY, *Modigliani* (1954), is one of the few noteworthy publications on the artist written in English. LAMBERTO VITALLI, *Disegni di Modigliani* (1929), is dedicated exclusively to the designs that are particularly numerous and important in the range of Modigliani's work. JEAN COCTEAU, *Modigliani* (1950), is another interesting volume written as a testimonial by a poet who was a friend and of whom Modigliani did one of his most famous portraits; JACQUES LIPCHITZ, *Amedeo Modigliani* (1953), is a small book, also written as a testimonial, by a friend who was a sculptor, whose "Man with His Wife" is one of Modigliani's most famous works. CLAUDE ROY, *Modigliani* (Eng. trans. 1958), includes biographical data, bibliography, and a list of exhibitions. ALFRED WERNER, *Modigliani the Sculptor* (1962), is the most important study of his sculpture.

(P.Bu.)

Mohammed I Askia of Songhai

Mohammed I Askia, one of the greatest rulers of the Songhai Empire in the western Sudan of West Africa, introduced important political and administrative reforms that brought security and prosperity to his people.

Both Mohammed's place and date of birth are unknown. For a long time, he was thought to be a Silla (a Tukulor clan of Senegal) or a Touré of Sarakole (Soninke) origin, but it now seems that his name, as spelled in Arabic by 18th-century Timbuktu chroniclers, was Mohammed Towri, or Mohammed of the Toro (Fouta Toro of Senegal). So he was probably of Tukulor origin, from a Senegalese family that had settled in Gao. The name of his clan was probably Kan, or Dyallo. Oral tradition, however, which is still very much alive, makes Mamar (Mohammed's popular name) out to be Sonni ʿAlī's nephew, his sister Kasey's son by a *jinnī*.

After the death of Sonni ʿAlī Ber, the ruler who had solidified the Songhai empire from 1464 to 1492, Mohammed tried, as early as February 1493, to wrest power from Sonni ʿAlī's son Sonni Baru, who had been elected by acclamation on January 21. In the Battle of Anfao on April 12, 1493, Mohammed's forces, though inferior in number, were victorious. Paganism tinged with the esoteric Songhai Islām of the Sonnis gave way to an Islāmic state whose civil code was the Qur'ān and whose official writing was Arabic. After conquering the enemy, Mohammed assumed the title of Askia in order to ridicule, it is said, the daughters of the fallen Sonnis who said of him *a si tya*, or "he will not be."

While Sonni ʿAlī had been a warrior, Askia Mohammed was above all a statesman. He set up an efficient administration of the regions conquered by his predecessor. He began by dividing Songhai into provinces and placed each under a governor. A standing army and a fleet of war canoes were organized under the command of a general and an admiral. Moreover, Askia Mohammed created the positions of director of finance, justice, interior, protocol, agriculture, waters and forests, and of "tribes of the white race" (Moors and Tuaregs who at that time were vassals of the Songhai and furnished them with squadrons of dromedary-mounted troops). All these officials were for the most part chosen from among the nobles and were brothers, sons, or cousins of Askia.

Organization of the Songhai Empire

This exemplary organization of an African state was completed by a religious organization. Although a faithful believer, Mohammed was not very well informed in matters of religious orthodoxy and therefore took as an adviser the Moroccan reformer, el-Mehrili, persecutor of the Jews of Touat, to help him put his realm in order, in particular to recover the possessions belonging to the descendants of the defeated Sonnis and to "servile tribes" not converted to Islām. Establishing Islām as the official religion of the nobles was without doubt the only error of this statesman. From then on, it was no longer a popular religion but an imported one that later was to justify the conquest of the Songhai by Moroccan Muslims.

And yet it was to receive the necessary counsel directly from "God's House" that in 1495, two years before his accession to power, Askia Mohammed undertook a pilgrimage to Mecca. This pilgrimage, from 1495–97, has remained famous as much for the pomp with which it was carried out as for the marvelous tales to which it gave rise. The chronicler Mahmud Kati, who accompanied Askia, wrote in *Tarīkh al-fattāsh* that the *jinnī* of Mecca had had Askia Mohammed named caliph and had told him what his rights were over the former vassal tribes of the Sonnis. By the time he returned in 1497, he was a leader deeply converted to Islām. Next he would proceed to consolidate and enlarge Songhai.

Militarily he met with uneven success. Although between 1498 and 1502 he was victorious over the Mossi of Yatenga and the inhabitants of the Aïr (Niger), a few years later (1505–06) he undertook an unsuccessful campaign against Borgu (the present boundary region of Niger and Nigeria). And similarly, although between 1507 and 1514 he reduced the insurgent Fulani factions in Senegal and the Bornu factions near Agadez, one of his lieutenants, the Karta of Kabi, revolted against him in turn and, despite his efforts between 1516 and 1517, remained independent. As an organizer of an effective administrative system he was more successful.

During the course of his lengthy sojourns in the capital, Gao (1502–04 and 1506–07), he set up with rare talent the system of tithes and taxes, the regulation of agriculture and fishing, and the recruitment and training of his administrators and governors.

The extent of the Songhai Empire of this period remains conjectural. Sadi, the Timbuktu chronicler, has said that the territory that Askia Mohammed conquered "by fire and sword" extended west as far as the Atlantic Ocean, northwest to the salt mines of Teghaza (on the northern border of present-day Mali), southwest as far as Bendugu (Segu), southeast to Bussa, and northeast to Agadez. It is certain that the influence of Songhai during Askia Mohammed's time was considerable and extended even beyond these boundaries. All the surrounding states, whether allies or enemies, experienced its civilizing ferment.

This influence was reinforced by an indirect, though nonetheless profound, Islāmic propaganda. Muslim scholars went into areas they would not have been able to penetrate without the Gao ruler's support. And for several centuries to come, the small African states and the neighbouring leaders would take as their model the Islāmic empire of Songhai and its prestigious leader, Askia Mohammed. Even today, according to oral tradition, Askia Mohammed appears as a supernatural being, a *jinnī*, who either took after his father or after those with whom, by a special gift, he was able to consult during his pilgrimage to Mecca.

Fall from power and death

The end of his reign was, however, tragic. Little by little his dream of an Islāmized Sudan, whose emir he would be, evaporated. Even during his lifetime, his children were quarrelling over the spoils. After the death of his commander in chief, Kanfari Omar, one of his brothers, in 1519, Askia Mohammed was no longer safe even in Gao, and the Songhai people seemed to him "as crooked as the course of the Niger River . . ." Embittered, half blind, the old man had no one left but his friend and adviser, his servant Ali Folen. The almost religious fear that he inspired gave way to contempt. Musa, his eldest son, plotted against him and in 1528 killed his new general in chief, Yaya, another of Askia's brothers, who had remained faithful to him. Musa then dispossessed his father, taking the name Askia Musa. He kept this title for three years before being assassinated himself by one of his brothers. Now deposed, the old Askia Mohammed was banished to an island in the river, a place "infested with mosquitos and toads." There, from 1528 to 1537, he was a blind and despairing witness to the murderous quarrels of his children over the territory of Songhai.

In 1537 his third successor, his son Askia Ismaïl, recalled his father to Gao. To reward him, Mohammed bequeathed to him his green turban and his caliph's sabre. On March 2, 1538, during a period of temporary calm, this founder of a dynasty died. He was buried in Gao, under a pyramid of earth surmounted by wooden spikes. His tomb is still standing and has become one of the most venerated mosques in all of West Africa.

BIBLIOGRAPHY. A detailed bibliography of the history of Songhai may be found in JEAN ROUCH, *Contribution à l'histoire des Songhay* (1953). ABD AL-RAHMAN IBN ʿABD ALLAH,

AL-SA'DI, *Tarikh-es-Soudan* (written *c.* 1655; published in 2 vol., 1898–1900), is the classic source on the history of Songhai to 1655. See also HEINRICH BARTH, *Travels and Discoveries in North and Central Africa*, 5 vol. (1857–58; reprinted in 3 vol., 1965), based on the Sa'di chronicle. E.W. BOVILL, "The Niger and the Songhai Empire," *Journal of the African Society*, 25:138–146 (1925), is a good account by one of the few historians to emphasize the essential role of the Niger River in Songhai history.

(J.P.R.)

Moldavian Soviet Socialist Republic

The hilly land of Moldavia (formally the Moldavian Soviet Socialist Republic) lies in the extreme southwest of the European portion of the Soviet Union, surrounded by the Ukraine on the north, east, and south and bordered on the west by the Prut River, forming the Soviet frontier with Romania. Occupying an area of 13,000 square miles (33,700 square kilometres), it is the second smallest of the 15 constituent union republics, with a population of more than 3,800,000. The capital is at Kishinyov.

The greater part of the republic lies between the great, meandering rivers Prut and Dnestr (Dniester), as they flow to their Black Sea estuaries. This region was formerly known as Bessarabia. It was an integral part of the Romanian principality of Moldavia until 1812, when it was ceded to Russia by the suzerain Ottoman Empire. It remained a province of the Russian Empire until 1918, except for the southern districts along the Danube River, which were joined to Romania after the Crimean War in 1856, only to be retroceded to Russia in 1878. In 1918 Bessarabia became part of Greater Romania, and in the interwar period it was gradually integrated into the political and economic structure of that country. The Soviet Union did not recognize this cession of territory, and in 1924 it established the Moldavian Autonomous Soviet Socialist Republic, a small strip of territory on the left bank of the Dnestr. In June 1940 the Soviet Union obliged Romania to cede Bessarabia, and in August it established the Moldavian S.S.R. Between 1941 and 1944, while Romania and the Soviet Union were at war, Bessarabia again became part of Romania. In 1944 Soviet troops retook the territory, and the Moldavian S.S.R. was re-established. The territory has been fully incorporated into the Soviet political and economic system, but since the 1960s Bessarabia has again become a subject of controversy between Romania and the Soviet Union.

For related information, see DNESTR RIVER; BALKANS, HISTORY OF THE; RUSSIA AND THE SOVIET UNION, HISTORY OF.

THE LANDSCAPE

Topography. Moldavia lies beyond the great arc of the Carpathian Mountains, mostly on deep sedimentary rocks covering the southwest portion of the ancient structural block known as the Russian Platform. Harder crystalline rocks outcrop only in the north. Its surface is a hilly plain, with an average elevation of 482 feet (147 metres), cut by a deep network of river valleys, ravines, and gullies.

Kodry Hills

The uplands of the centre of the republic, the Kodry Hills, lie at an average elevation of about 1,150–1,300 feet (350–400 metres), and the highest point, Mt. Balaneshty in the west, reaches 1,409 feet (430 metres). These uplands are interlaced by deep, flat valleys, ravines, and landslide-scoured depressions, separated by sharp ridges, and steep, forest-clad slopes account for much of the terrain. The Dnestr uplands, their eastern slopes forming the high right bank of the Dnestr River, border the central uplands on the east and northeast.

The northern Moldavian landscape is characterized by the level plain of the Beltsy steppe (500 to 650 feet, or 150 to 200 metres, in elevation) and also by uplands averaging twice this height, culminating in the 1,053-foot (321-metre) Vysokaya Hill. The northern uplands include the strikingly eroded Medobory-Toltry limestone ridges, which border the Prut River.

In the south, the extensive Budzhak Plain is broken by numerous ravines and gullies, while in the east, left-bank Moldavia includes spurs of the Volyn-Podolsk Upland, cut into by tributaries of the Dnestr.

Climate. Moldavia's climate, warm and moderately continental, is characterized by a lengthy frost-free period, a comparatively mild winter, considerable temperature fluctuations, and, in the south, long dry periods.

Average annual temperature reaches 45.9° F (7.7° C) in the north and 49.8° F (9.9° C) in the south, but the July averages rise to 67° and 73° F (19.5° and 23° C), respectively, and the mercury drops to as low as 23° and 27° F (−5° and −3° C) in January. Absolute temperatures recorded range from a record low of −31° F (−35° C) and a high of 106° F (41° C). Moldavia experiences a very variable rainfall, with the average figures of 18 to 22 inches (450 to 550 millimetres; the figure is a little lower in the south) concealing variations that may double the figure in some years and result in prolonged dry spells in others. Most rainfall occurs in the warmer months, and heavy summer showers, coupled with the irregular terrain, cause erosion problems and river silting. Winter snow cover is thin. Winds tend to come from either the northwest or the southeast.

Drainage. Moldavia has a well-developed network of about 3,000 rivers, all draining south to the Black Sea, but only 246 of these exceed six miles (10 kilometres) in length, and only eight exceed 60 miles. The Dnestr, the rapidly flowing main artery, is navigable almost throughout the republic and has an average discharge of about 10,000 cubic feet (300 cubic metres) per second, swollen by spring snowmelt from the Carpathians and by heavy summer rains. In the warmer winters, it is not frozen. The other, smaller, main artery, the Prut, is a tributary of the Danube, which it joins at the extreme southern tip of the country. The Yalpukh, Kogilnik, and other small southern rivers drain largely into the Danubian estuary in the nearby Ukraine.

Underground water, extensively used for the republic's water supply, includes about 2,200 natural springs. The terrain favours construction of reservoirs.

Soils. Moldavian soils are varied and very fertile, with chernozem—rich, black soils—covering three-quarters of the republic. The best developed chernozem, fostering the growth of grain, tobacco, and sugar beets, is found in the north, in the low-lying parts of the central Moldavian and Dnestr uplands, as well as in the left-bank regions. Chernozem quality diminishes southward, but grapes and sunflowers can still be grown. Woodland soils characterize the uplands: 40 percent are covered by forests, the rest by orchards, vineyards, and fields of grain. Floodplain soils characterize the riverside areas, while the lower reaches of the Prut and southern river valleys have saline and marshland soils.

Chernozem soils

Plant and animal life. Northern and central Moldavia is a forest zone, while a steppe belt crosses the south. There are more than 1,500 species of plants in the republic, with scenic expanses of forests, covering about 1,150 square miles (3,000 square kilometres), of particular importance, especially in the central Kodry Hills region. The commonest trees are hornbeam and oak, followed by a rich variety including linden, maple, wild pear, and wild cherry. There are beech forests at the sources of the Ikel and Byk rivers. Large-scale reforestation projects have been carried out in the republic.

Moldavia's steppes were originally grass covered, but most of them are now cultivated. Lush meadows and reed growths occur in the floodlands of the Dnestr and portions of the Prut, while salt-marsh grasslands flourish in the saline valleys of the Kogilnik, Yalpukh, Botna, and lower Prut.

Moldavian animal life is rich, despite the republic's small size. Mammals include boar, roe deer, hare, wolf, fox, badger, wildcat, ermine, and polecat. Siberian stag, fallow deer, spotted deer, and muskrat have been successively introduced and are firmly established. Roe deer, hares, foxes, and muskrats are of commercial importance.

There are many species of birds, both resident and migratory. The marshy lower reaches of Moldavia's rivers provide sanctuary for wild geese, migratory ducks, and herons, while white-tailed sea eagles are to be found in

the floodland forests. The wood lark, jay, song thrush, blackbird, hawk, and long-eared owl frequent the republic's forests. Plentiful fish supplies include carp (raised in artificial reservoirs), perch, bream, ruff, and pike.

THE PEOPLE

By the mid-1970s, Moldavia was the home of more than 3,800,000 people, an increase of 32 percent since 1959 and of some 54 percent since 1940. The annual population increase is about 1.2 percent, and Moldavia has a higher average population density than any other Soviet Union republic.

Ethnic composition

Of the total population, according to the 1970 census, 2,304,000 (64.6 percent) were Moldavian, 507,000 (14.2 percent) Ukrainian, 414,000 (11.6 percent) Russian, 125,000 (3.5 percent) Gagauzi (a Turkic people), 98,000 (2.7 percent) Jewish, and 74,000 (2.1 percent) Bulgarian. Between 1959 and 1970 the Moldavians and Ukrainians in the total population declined by 0.8 and 0.4 percent, respectively, while the Russians increased by 1.4 percent.

The urban population has increased as a result of industrialization: by 1974 about 35 percent of the people lived in towns, compared with 13 percent in 1940. Most of the towns are small; in 1974 Kishinyov had an estimated 432,000 inhabitants, Tiraspol and Beltsy each had a little over 100,000, and Bendery had 89,000, with other towns below 50,000.

Moldavian villages are large, most of them having a thousand and more inhabitants; all are supplied with electricity, and the gas network is being extended.

THE ECONOMY

A diversified industry has been established in Moldavia, agriculture has been modernized, and transport and the building industry have been overhauled. Investment programs had borne fruit by the mid-1970s, when gross industrial output was more than 35 times the level of 1940. The same period witnessed a more rational distribution of productive forces over Moldavian territory, equalizing the economic development of various regions. By 1973 the specialized labour force approached 1,150,000, or 11 times the total in 1940. Moldavian industrial and agricultural products are supplied to all the union republics and are exported to some 50 foreign countries.

Industry. The leading branch of the Moldavian economy, industry concentrates in the main on food processing, with the machine-building, power-engineering, consumer-goods, and building-materials industries undergoing development as well.

The food industry

The food industry has numerous branches: wine making, canning, sugar refining, and oil pressing, as well as production of essential oils, are of all-union significance. In particular, Moldavia is an important supplier of wine, including champagne and brandy, and is responsible for one-fifth of the Soviet output.

The canning industry processes mainly vegetables and fruits, and Moldavia is responsible for 9 percent of Soviet production of these and for a full third of Soviet fruit juice output. Southeastern Moldavia alone produces about 1,000,000,000 cans annually.

Moldavia holds the third place in Soviet sugar production, the sugar-refining industry being located in the north.

The republic produces more than 5 percent of Soviet vegetable oil and one-third of Soviet tobacco. For local needs the republic has flour and other mills and well-developed meat, dairy, and confectionary industries.

Machine building, established in the mid-1950s and centred on Kishinyov, Beltsy, Tiraspol, and Bendery, is acquiring increasing importance. Tractors made in Moldavia, specially equipped for use in orchards and vineyards, are supplied to other Soviet republics.

Light industry includes the production of furs at Beltsy, garments and knitwear at Kishinyov and Tiraspol, footwear at Kishinyov, and silk fabrics at Bendery.

Building materials produced in Moldavia include bricks, limestone, tiles, cement, slate, and concrete blocks. Rybnitsa is the leading centre of this industry.

The power industry of the republic is expanding; thermoelectric power plants are located in Kishinyov, Beltsy, and Tiraspol, and there is a hydroelectric station in Dubossary on the Dnestr. A Moldavian hydroelectric station in the town of Dnestrovsk was completed in 1971 (and enlarged during 1973–74), with a final installed capacity of 2,000,000 kilowatts. Some 12,200,000,000 kilowatt-hours of electricity are generated annually, compared with a total of 17,200,000 in 1940. The republic provides electricity to the southern regions of the Ukraine and since 1972 has supplied power to Bulgaria through a transmission line. The republic's power system is connected to the unified southern power system of the Soviet Union.

Agriculture. There are about 480 collective and 240 state farms, and the total sown area is 4,650,000 acres (1,900,000 hectares). There is no large-scale private farming, but collective farmers do have small plots for their own use. Collective farms cultivate four-fifths of the area under grain crops and are generally engaged in mixed farming. On an average a collective farm has 7,000 acres (2,800 hectares) under crops and more than 1,200 acres (500 hectares) of vineyards and orchards. The state farms usually specialize in cultivating and processing a particular crop, some of them producing grapes and wines, others growing aromatic plants and producing essential oils.

Agriculture is highly mechanized, 98 percent of agricultural jobs being carried out by machines. All collective and state farms are provided with electricity, and use is made of chemical means of pest control and mineral fertilizers. The irrigation system is being expanded.

Arable farming

Arable farming is responsible for slightly more than 70 percent of the gross agricultural output, while the share of animal husbandry is not quite 30 percent. Arable farming accounts for up to 72–73 percent of the cash income of the collective farms.

The structure of arable farming has undergone radical changes in the Soviet period: the share of grain crops is about 45 percent, compared with 80 percent in 1940, while the share of industrial crops has increased, accounting for more than 20 percent by the 1970s. In order to consolidate animal farming, the share of fodder has also been raised and amounts to 30 percent. Since 1940 the area under vegetables, orchards, berries, and vineyards has undergone significant expansion. Viticulture, fruit and vegetable growing, and other specialized farming are particularly important, comprising 90 percent of the commodity output of arable farming. The largest vineyards are to be found in the southern and central regions, and most of the orchards are situated in northern and southeastern Moldavia. Walnuts are also significant.

Sunflower seed is the most important industrial crop, and, though it is grown all over Moldavia, the southeastern regions have the largest plantations. Sugar beets, a relatively new crop in Moldavia, are cultivated in the north. Moldavia is also the major Soviet tobacco grower. Vegetables are grown mainly in the southeast.

The chief grain crops are winter wheat and maize (corn). Wheat is used for the republic's own needs, and maize is supplied as seed to other republics. While the areas under wheat have decreased considerably, its gross output, aided by measures to increase productivity, has gone up. Moldavia holds first place in cereal crop capacity in the Soviet Union. Most of the grain is grown in the north.

Dairying and cattle breeding account for almost 30 percent of agricultural output, followed by pig farming. The sturdy Karakul and *tsigayk* breeds constitute a large portion of the sheep, raising of which is particularly important in southern Moldavia.

Economic subregions. The Moldavian S.S.R. is divided into four subregions: Northern, Central, Southern, and Southeastern. Each of the subregions has distinct natural and economic features, possessing unique problems of development.

The Northern subregion. This stretches from the northern border to the central Moldavian upland. It covers 13 districts, and its area constitutes about a third of the total area of the republic and the same proportion

of the total population. Urban and rural populations constitute, respectively, 21 and 79 percent. Agriculture dominates the economy: the subregion produces about 40 percent of the republic's grain, 95 percent of sugar beets, half the tobacco, 30 percent of fruit, 40 percent of sunflower, and 45 percent of animal-husbandry products. A diversified industry has been established. The subregion is responsible for 27 percent of the total industrial output, including all the cement and sugar production, and more than 80 percent of processed tobacco. Instrument making and electrical engineering are important industries. Beltsy is a prominent industrial centre, and Soroki and Rybnitsa are also developing industrial towns.

The Central subregion. This lies in the middle of Moldavia, between the Dnestr and the Prut. It occupies about a quarter of the republic's territory and is the home of a third of the population. Engineering and light industry are of great importance, alongside the food industry. The subregion also has a developed agriculture, concentrating on viticulture, fruit, and dairy and poultry farming.

Kishinyov

Kishinyov, the Moldavian capital, is the largest industrial centre of the subregion and of the republic. The urban complex produces more than one-third of the republic's engineering industry output, more than two-fifths of the light industry commodities, more than 25 percent of the building materials, and about 20 percent of the food industry products.

The Southern subregion. Occupying 20 percent of the republic's territory, and the home of only 14 percent of its people, the Southern region is predominantly agricultural. There are no large cities, and the most important town is Kagul, on the Prut River, which has wine-making and building-materials industries.

The Southeastern subregion. This occupies about 18 percent of the republic's territory and has about the same proportion of the population. It is noteworthy for its highly developed intensive agriculture and industry, specializing in the production of vegetables, fruits, grapes, and sunflower seeds and in animal husbandry. It also has silk-manufacturing, canning, oil-pressing, meat-packing, and heavy industrial plants, with Tiraspol and Bendery the main industrial centres.

Transport. Railway and motor transport are the basis of the Moldavian transport system. The railway network contains two lines—one linking Tiraspol, Kishinyov, and Ungeny and the other linking Tiraspol and Reni—that are of importance in the Soviet Union as a whole.

By 1972 the railways were carrying nearly 13 times as much outgoing freight as they had in 1940, while incoming freight, a good gauge of economic development, had increased more than 30-fold. Incoming goods include coal, petroleum products, iron and nonferrous metals, timber, mineral fertilizers, and machines and equipment. Agricultural goods and products of the food industry prevail among the outgoing commodities, although a growing quantity of machines and light industrial goods is sent to other parts of the Soviet Union. In addition, Moldavian railways carry 10,000,000 passengers annually.

Motor transport generally carries freight inside the republic, over a road network of which somewhat less than half is asphalt covered. By 1973 road transport was responsible for hauling 224,100,000 metric tons of freight, 150 times the 1940 figure, while buses carried more than 240,000,000 passengers.

River transport is of local importance, and air transport links Moldavia with other parts of the Soviet Union, Kishinyov being the site of the main airport.

POLITICAL AND SOCIAL STRUCTURE

The constitutional framework. According to the constitution of the Soviet Union and the constitution of the Moldavian Soviet Socialist Republic (adopted in February 1941), Moldavia is a sovereign socialist state of workers and peasants. Its political foundation is the Soviets of Working People's Deputies, and its economic foundation is the socialist ownership of the means of production. It has a coat of arms, flag, and national anthem.

A one-chamber Supreme Soviet, elected for a four-year period, is the highest organ of state power, supplemented, between sessions, by the Presidium of the Supreme Soviet, which it elects. The Supreme Soviet appoints the government—namely, the Council of Ministers, which is the highest executive and administrative organ.

The organs of state power in districts, cities, and rural localities are the Soviets of Working People's Deputies, which are elected for two-year periods. The Supreme Soviet also appoints, for a five-year period, the Supreme Court, which is the highest judicial establishment of the republic. The general procurator of the Soviet Union appoints for a five-year period the chief procurator of Moldavia, who is responsible to him.

The Communist Party

The Communist Party of Moldavia, part of the Communist Party of the Soviet Union (CPSU), is the only political party allowed to function. There is no organized political opposition. So-called nonparty candidates may run for election to the Supreme Soviet and other bodies, but only on a single list with Communist candidates approved by the Communist Party. Nonparty deputies do not pursue policies distinct from those of the Communist Party. As in the other union republics, the party employs the various state organs mentioned above to carry out its policies in all spheres of activity. The primary purpose of elections and the various public bodies is, consequently, to rally mass support for party policies. The party had about 125,000 members and candidates to membership by the mid-1970s. The Young Communist League (Komsomol) of Moldavia, with a membership of about 450,000 young people, is in close collaboration with, and under the leadership of, the Communist Party of Moldavia.

The trade unions of the republic have a membership of 1,220,000. As in other socialist countries, they concern themselves with the recreational and working conditions of their members, the stimulation of fulfillment of various economic goals, and the organization of the emulation process known as socialist competition. The trade unions have little independence or power of initiative. No strikes in Moldavian industry have been recorded since the end of World War II.

Education and social services. Significant changes have occurred in Moldavian society during the years of Soviet power. Illiteracy was eradicated, and, as in other union republics, a general eight-year education scheme—changing over to a general 10-year program by the early 1970s—was introduced. Emphasis has been placed upon technical education in order to satisfy the steadily growing needs of agriculture and industry for specialists and a highly skilled work force. In 1975 there were nearly 700,000 school children; 92 percent of those who finished the eight-year program continued their education. Before 1940 Moldavia had only one institution of higher education, a teacher-training college. In 1973–74 there were 42,400 students in higher educational institutions and 52,500 in specialized middle schools.

Academy of Sciences

At the highest level, the Moldavian S.S.R. Academy of Sciences, established in Kishinyov in 1961, coordinates the activities of some 20 scientific institutions. An additional 50 or so research centres in the fields of viticulture, horticulture, beet growing, grain cultivation, and wine making have been set up, and Moldavian scientists have won international acclaim in these fields.

The rising standards of living in the republic are closely connected with the development of an economy utilizing scientific methods to improve techniques and overall productivity. The monthly income of the average Moldavian is enhanced by a third or so by the provision of social services from the republic's national income. In addition to free education in elementary, secondary, and high schools, this "social consumption fund" is used for medical help, for pensions and stipends to students, for free stays in sanatoriums and the like, for kindergartens and nurseries, and for Pioneer camps for younger children. Social insurance funds created for agricultural workers are used to provide medical assistance during sickness and temporary disability. All payments for education and social services come out of the state budget.

By the early 1970s the annual rate of completion for new apartments and houses in both town and country was around 36,000 annually. This figure nevertheless re-

mained inadequate to meet all needs, particularly in the light of the movement to the cities, and the need for further improvement in housing continued to occupy the attention of Moldavian planners.

Medical facilities

Moldavia has about 360 hospitals, with about 40,000 beds, in addition to some 230 clinics and consulting medical stations for women and children. There are 24 doctors and 82 medical professionals for every 10,000 people; all kinds of medical aid are free.

CULTURAL LIFE AND INSTITUTIONS

Because of the historical ties between Bessarabia and Romania and the ethnic identity of Moldavians and Romanians, the culture of the Moldavian S.S.R. to a considerable extent still reflects these primordial sources. Although the Moldavian language, as it is officially called, employs the Cyrillic alphabet, the differences between it and literary Romanian are of little significance and are confined to phonetics and vocabulary. Soviet scholars, mainly for political reasons, have insisted since the 1920s that Moldavian is an independent Romance language distinct from Romanian, a view rejected in the West.

Since World War II the development of Moldavian culture and literature has followed the prevailing pattern of the Soviet Union as a whole. Moldavian literature experienced the vicissitudes of Soviet literature generally during the Zhdanov era of the late 1940s and early '50s. Nonetheless, some notable work was done by the poets Emilian Bucov and Andrei Lupan, among the older generation, and by the dramatist and prose writer Ion Druţa, whose novel *Frunze de dor* (1957; "Leaves of Yearning") marks the coming of age of literature in Romanian in the Moldavian republic.

Particular attention has been given to the expansion of cultural opportunities for the general population. There are numerous amateur theatres and musical and art groups. The Doina choir and the Zhok popular ballet company have won international reputations.

Publishing has grown with the expansion of education. In 1973, 1,627 books (548 in Moldavian) in 12,700,000 copies (7,300,000 in Moldavian), 129 newspapers (58 in Moldavian), and 63 magazines and journals (18 in Moldavian) were published.

PROSPECTS

Plans for the further development of Moldavian society and economy will continue to emphasize the expansion of industry, especially in the areas of power, machinery and instruments, food processing, and light industry. Some attention will also be given to an increase in consumer goods, particularly wines. In agriculture, efforts are to be made to increase the amount of land under cultivation and to improve crop yields, especially through the expansion of irrigation projects.

(F.N.S./K.A.H.)

Molecular and Atomic Beams

Molecular beams are streams or rays of molecules moving in the same general direction, usually in a vacuum; *i.e.*, inside an evacuated chamber. In this context the word molecule includes atoms as a special case. Most commonly, the molecules comprising the beam are at a low density; that is, they are far enough apart to move independently of each other. Because of the one-directional motion of the atoms or molecules, their properties can be studied in experiments that involve deflecting the beam in electric and magnetic fields or directing the beam onto a target. The target may be a solid, a gas, or a second beam of atoms or molecules.

Deflections of beams in electric and magnetic fields can give information about the structure and properties (such as rotation and spin) of the molecules, or atoms, in the beam. In more sophisticated experiments, two beams are allowed to intersect, producing scattering interactions or collisions between molecules in pairs, one from each beam. Scattering can demonstrate such properties of these pairs as the potential energy of their interaction as it varies with the distance of separation, their chemical reactivity, and the probability that they will exchange internal energy on collision. Reactivity and internal energy will be treated in detail below.

First experiment

The first experiment with molecular beams, in 1911, confirmed a postulate of kinetic theory that molecules of a gas at a very low pressure travel in straight lines until they hit the walls of their container. At higher pressures, molecules have a shorter free path because they collide with each other before arriving at the wall. The first extensive experiments with molecular beams were made in Germany between 1920 and 1933. The use of beams to study chemical reactions and the transfer of energy between colliding molecules increased rapidly after 1955.

PRODUCTION, CONTROL, AND DETECTION

A molecular beam is produced by allowing gas to enter a vacuum chamber through a small hole or slit in a box containing vapour of the molecules that are to make up the beam (see Figure 1). The vapour often comes from

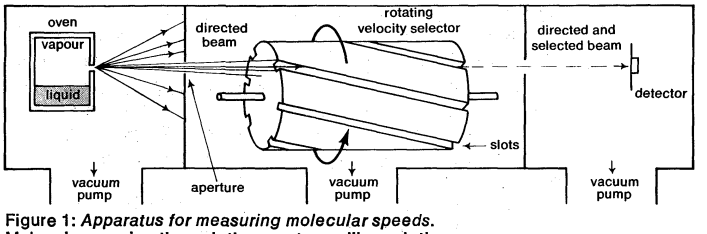


Figure 1: *Apparatus for measuring molecular speeds.* Molecules passing through the aperture will reach the detector only if the cylinder is rotating with such speed that the molecules have clear passage through the cylinder slots (see text).

evaporation of a liquid sample in the box, called an oven, that can be heated to a suitable temperature. At low pressures of vapour in the box, when the free path of the molecules is greater than the width of the exit hole, the molecules will effuse (as will be discussed later) through the hole; at higher pressures they will flow through the hole as fluids do under pressure, forming a jet. The molecules in the jet are at first close enough together to interact with each other, but the jet rapidly expands in the vacuum until the molecules move independently. Only those molecules from the oven that happen to be moving in just the proper direction to pass through a second hole become part of the beam that is used for the experiment. The others are pumped away.

Controlling speed with a velocity filter

Molecules in the beam move at various speeds. If molecules of nearly uniform speed are needed for a particular experiment, the beam can be put through a filter (called a velocity selector) that permits only molecules within a small range of speeds to pass through. These selectors are often made of slotted disks or cylinders spinning rapidly on an axis parallel to the beam. The molecules that emerge from the selector are those with the right speed to stay in a given slot as they move along the cylinder. Molecules of other speeds are removed as they stick to or reflect from the sides of the slots. Changing the speed of rotation of the cylinder changes the speed at which the molecules are transmitted.

To be useful in experiments the deflection or scattering of a beam must be detected. This detection may be difficult because there are relatively few molecules in a typical beam and their velocities, and hence kinetic energies, are low. A detector should have high sensitivity, and there must be little interference from molecules coming from other sources, such as the residual gas in the vacuum chamber. When an experiment is performed with beams consisting of atoms of metals such as the alkalis, which are easily ionized (*i.e.*, given a net positive charge) by the loss of one electron, an efficient detector can be made with a tungsten wire heated to redness. Because of the relatively high energy available for the capture of an electron by a hot tungsten surface, almost all alkali atoms hitting the wire give up one of their electrons to the wire, passing on as ions to be recorded as an electric current at a collecting electrode. A different kind of detector is needed for other kinds of beam molecules. Atoms and molecules can be ionized by bombarding them with a

stream of electrons, and the resulting ions can then be sorted and identified as to mass and charge by directing them into an instrument called a mass spectrometer. Although versatile, this type of detector is much less sensitive for alkali atoms than is the tungsten wire, because the mass spectrometer is generally able to register no more than one-tenth of one percent of all the beam molecules that enter it.

PROPERTIES OF MOLECULAR BEAMS AND THEIR COMPONENTS

The velocity distribution of molecules. According to the kinetic theory, molecules in a gas at thermal equilibrium (*i.e.*, all parts in the system being at the same temperature) in a container at a constant temperature are in rapid, random motion (the average speed of a hydrogen molecule at room temperature is over 4,000 miles [6,400 kilometres] per hour), making a large number of collisions with each other and with the walls of the container. Motion is equally probable in any direction, and there is a range of speeds in any volume of gas running all the way down to zero. The use of molecular beams permits a quantitative test of an expression first derived by the British physicist James Clerk Maxwell, who showed that for a large number of molecules in thermal equilibrium at a given temperature, there is a well-defined distribution of speeds in which some molecules are faster and some slower than those that have the most probable speed. If the hole in a box of gas that is the source for the molecular beam, pictured in Figure 1, has a diameter much smaller than the average distance travelled by molecules between collisions, a small fraction of the molecules will leave the box through the hole. This fraction can be made so small that the loss of molecules does not appreciably disturb the distribution of the speeds (numbers per unit of speed) of the molecules remaining in the box. The molecules leaving are said to effuse from the hole. They are then a representative sample of those in the box. If the beam formed from these molecules passes through the velocity selector and produces an electric current at the detector, the magnitude of this signal is directly related to the number of molecules in the gas of the source having the particular speed transmitted by the selector. Repeating the experiment at many different speeds gives an experimental set of values for the numbers of molecules versus speeds, which when plotted as points fit closely to the theoretical distribution curve derived by Maxwell. Such fits provide direct confirmation of the kinetic theory of gases.

Maxwellian distribution

Deflection processes and experiments. Beams of ionized molecules or atoms can be accelerated or deflected with comparative ease because of the strong forces exerted on charged particles by electric and magnetic fields. Beams of electrically neutral molecules, on the other hand, are not influenced when they move through uniform fields. If the charge distribution in a molecule or atom is asymmetric, if, that is, one part of the molecule is positively charged compared to another part, it has what is called an electric dipole moment and will be subjected to a force when it is in a nonuniform, or inhomogeneous, field, one that varies from point to point. Although this force is small, it can produce measurable deflections of a molecular beam.

Stern–Gerlach experiment

A famous experiment involving the deflection of an atomic beam was performed in 1922 by Otto Stern and his colleague, Walther Gerlach. They prepared a beam of silver atoms (in an apparatus like that of Figure 1, but without the velocity selector) and then let the neutral beam pass between the poles of a magnet shaped so that the field was much stronger near one pole than the other. After leaving the magnetic field, the beam hit a cold surface, where the atoms condensed. The shape of the deposit showed that the beam was deflected away from its initial direction to form two spots, one corresponding to deflection toward a higher value and one toward a lower value of the magnetic field, and equally spaced from the position of the undeflected beam. This result, as will be seen below, is exactly the one required by the quantum theory. Certain species of atoms have magnetic moments (or magnetic dipole moments); that is, they behave like tiny magnets because of the orbital motion or spin of their electrons. For silver atoms, the magnetic moment is caused by the spin of a single electron outside a closed shell of electrons, and, according to quantum mechanics, there are only two orientations of the magnetic moment permitted in an external magnetic field: parallel to or opposed to the direction of the field. The corresponding classical theory permits all orientations. If the classical theory had been applicable, the beam would have produced only a single, widened spot centred at the position of the undeflected beam. In showing dramatically that only definite orientations of the magnetic moment of an atom are permitted in a magnetic field, the experiment provided strong confirmation for the predictions of the quantum theory.

For reasons inherent in the phenomena of the spectra and the structure of atoms, the observation of very closely spaced lines, called hyperfine structure, in optical spectra suggested that atomic nuclei might also act like magnets—*i.e.*, have magnetic moments—although they were expected to be much smaller than those of the unpaired electrons of atoms like silver. (Spins of paired electrons are of opposite direction and thus can cancel one another's magnetic moment.) The magnetic moment of the proton, the positively charged nucleus of the hydrogen atom, which is also caused by its spin, was of particular interest because the proton is the simplest of all nuclei. At low temperatures, hydrogen molecules exist almost entirely in two energy states: in the first of these, the molecule is not rotating and the magnetic moments or spins of the two nuclei (protons) are oriented opposite to each other; in the second, the molecule itself is rotating and the spins of the two nuclei are parallel to each other. In a refinement of the Stern–Gerlach experiment, it was found that molecules in the first energy state were not deflected in the inhomogeneous field (because the nuclear moments cancel each other) and molecules in the second energy state were split into a pattern consistent with quantum theory. The measured value of the magnetic moment of the proton, however, was bigger than had been calculated, suggesting that the proton must have a more complex structure than does an electron.

Magnetic moment of protons

Most molecules in which the atoms are arranged unsymmetrically have electric dipole moments as a result of uneven distribution of their electrons and may be deflected when they move through an inhomogeneous electric field, just as atoms having magnetic dipole moments can be deflected by inhomogeneous magnetic fields. In one experiment, a molecular beam passes through a field that can be switched on or off. A small stop is arranged to block the direct beam to the detector when the field is off. When the field is then switched on, molecules having electric dipole moments greater than some small value are deflected around the stop and reach the detector. In this way, molecules such as those of barium chloride were shown to have electric dipole moments and thus to have an unsymmetrical structure of the form $\left(\mathrm{Ba}\begin{smallmatrix}\diagup\mathrm{Cl}\\ \diagdown\mathrm{Cl}\end{smallmatrix}\right)$. This structure indicates that the barium atom (symbolized by Ba) is bonded to two chlorine atoms (symbolized by Cl) that are not balanced against one another on either side of the barium. The symmetric, linear form of the molecule can be indicated as (Cl—Ba—Cl), and in this arrangement of atoms, the dipole moment would necessarily be zero. Calcium chloride, unlike barium chloride, is apparently linear and of the form (Cl—Ca—Cl), because it is not deflected. Information about the shapes of these molecules was not known before the deflection experiments.

Spectroscopy with beams. Spectroscopy is the study of the interaction of electromagnetic radiation, including visible light and radio waves, with matter.

The internal energy of a molecule is due, in part, to the fact that it can rotate about an axis and its atoms can vibrate with respect to one another. These motions are quantized—that is, only certain frequencies of rotation

and vibration are allowed, depending on quantum numbers arising from quantum theory. Each set of quantum numbers specifies a distinct energy level or state. Rotational levels are separated by energies corresponding to ultra-high-frequency radio waves and far infrared radiation. Different vibrational levels are separated by energies corresponding to the near infrared. Thus, the absorption of a quantum of energy in the infrared may be sufficient to raise the energy of a molecule by increasing the vibration rate from that associated with a particular level to that of another, higher level. Or, the emission of a quantum of energy may be caused by the molecule, changing from one rotational state to another one, of lower energy.

The frequency of the radiation absorbed when a molecule is excited, or raised, from an initial state of internal energy to a final state of energy is equal to the difference of the two energies divided by Planck's constant and is given by the equation $\nu = (E_f - E_i)/h$, in which ν is the frequency, E_f is the final energy, E_i is the initial energy, and h is Planck's constant. This frequency varies somewhat with the speed of the molecule with respect to the source of the energy, just as the pitch of sound waves varies when the source moves away or toward the listener. For a gas, in which molecules are moving at different speeds and directions, there is a range of frequencies in which the spectroscopic line is no longer sharp but is widened, in what is known as Doppler broadening. This variation of the frequency absorbed with the molecular speed often limits the accuracy with which the energy difference can be measured in ordinary optical spectroscopy and thus conceals fine differences in the energy states. Molecular beam measurements are advantageous in that the method reduces the Doppler broadening of spectral lines because the beams can be observed at an angle of 90 degrees from the direction in which the molecules are moving so that their speed away from or toward the observer becomes zero.

Doppler broadening

The most notable contributions of molecular beams to spectroscopy come from the possibility they offer of observing in a sensitive way the absorption of small amounts of energy by molecules that are electrically asymmetric; *e.g.*, those having spin or dipole moments. The trajectories that such molecules in a beam follow through a nonuniform magnetic field depends on the energy state of the molecules; if the state can be changed by the absorption of radiation, the molecules follow changed trajectories that can be designed to make them miss a detector they would otherwise hit. In this arrangement, a drop in the beam current at the detector indicates that the molecules in the beam are absorbing the radiation. The frequencies absorbed can be measured accurately, and the measurements in turn yield information about the molecules.

Resonance phenomena. The internal energies of atoms and molecules, the energy, for example, with which a molecule rotates, are not distributed continuously over a range. Instead, each molecule is able to exist only in certain discrete energy states. When a molecule is in an electric or magnetic field, however, the energy of a particular state varies as the strength of the field varies. Transitions between these states, which may occur on the absorption or emission of a quantum of energy, occur at frequencies that also vary with the strength of the field. Resonance occurs for a molecule when the irradiating frequency just corresponds to the energy difference existing between two states for that molecule.

Although in the past measurements of the deflections of molecular beams in inhomogeneous fields provided new information of fundamental importance for physics, they were difficult to render quantitative both because of the need of measuring accurately the values of the inhomogeneous fields and because of the large range of speeds of the molecules in the beam. A remarkable improvement in accuracy was provided by a resonance method introduced in 1938.

Resonance experiment

A schematic view of the apparatus used in this method is shown in Figure 2. At the left is a source from which the beam moves into a steady, nonuniform field (labelled A)

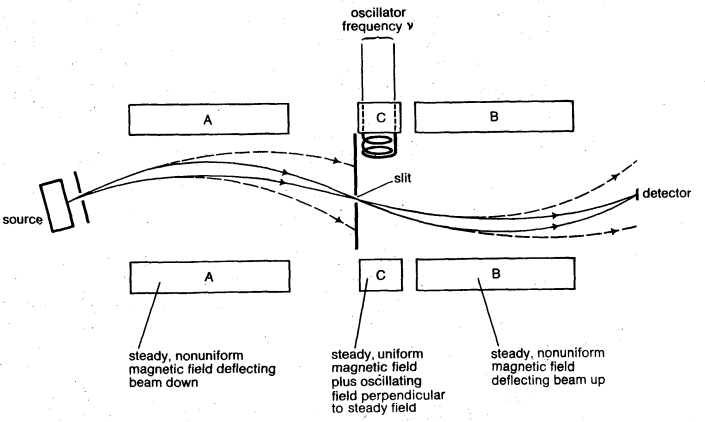


Figure 2: Molecular beam apparatus for magnetic resonance experiments (see text).

like that used for the Stern–Gerlach experiment. The solid lines show two trajectories for molecules, of the particular energy state for which they are deflected just enough to pass through the slit. Molecules in other states follow trajectories, such as those shown as dashed lines, that miss the slit. Beyond the slit the molecules enter a region (labelled C) in which the main magnetic field is uniform but where there is also a second field, perpendicular to the first and oscillating at a given frequency. There some molecules may absorb radiation from the oscillating electromagnetic field and be raised to a higher energy state. From C, the beam enters the field marked B, which is like the field at A except that it causes deflection in the opposite direction. Molecules that have not absorbed radiation from the oscillating field reach the detector (solid lines), but those that have undergone transitions follow new trajectories (dashed lines), which miss the detector. An essential feature of the apparatus is that the inhomogeneous fields serve solely to take molecules in particular energy states to the region C and then on to the detector. The only quantities that must be determined accurately are the strength of the uniform field at C and the oscillator frequency. The technique has been so effective that since its development most studies of molecular and atomic structure employing molecular beams have used it.

An important advance in the technique of making accurate time measurements is the use of an atomic beam of cesium in a resonance experiment to provide a standard frequency. Although there are many transitions between energy states of isolated cesium atoms, which can be caused by the absorption of radiation, for example, in the visible region, there is one transition that is almost independent of the strength of the homogeneous magnetic field in the C region (Figure 2). Thus, the frequency absorbed in this transition is essentially a property of the cesium atom alone and independent of the apparatus. When the signal at the detector is used to control and stabilize the radio frequency used in the C region, this frequency becomes one of the most accurately measured physical quantities known. It is stable to one part in 10^{11}. In 1967 its value, 9,192.631770 megahertz (or units of 10^6 cycles per second), was adopted as a new standard for the definition of the second as a unit of time. The cesium beam can be used as part of an extremely stable clock if a counter is set up to keep track of the number of oscillations.

Cesium clock time standard

In the development of physics, measurements of the properties of the simplest particles of matter have been of especial importance. The electron is one of these particles, and a significant application of molecular beams was to obtain a value for the magnetic moment of a free electron (that is, an electron not bound to a nucleus in an atom) relative to its intrinsic angular momentum or spin. This ratio, called the gyromagnetic ratio (g), originally was given the value of exactly two according to the electron theory of an English physicist, P.A.M. Dirac. The uncertainty principle (*i.e.*, it is impossible to specify or determine simultaneously both the position and mo-

mentum of a particle with full accuracy), prevents a direct measurement of the gyromagnetic ratio for a free electron in a typical resonance experiment. This restriction, however, does not prevent measurements of electrons bound in atoms; so the value of the gyromagnetic ratio can be deduced from experiments conducted with beams of heavy, neutral atoms, such as the alkali metals, each of which has a single electron in its outer shell, in resonance apparatus like that shown schematically in Figure 2. For these alkali-like atoms there is no net magnetic moment due to the orbital motion of the electrons about the nucleus, and so the total magnetic moment may be expected to be the same as that of a free electron. The value of the gyromagnetic ratio found, and confirmed later by other, more precise experiments not involving molecular beams, was larger than two by 0.119 percent. This small difference helped to lead to significant corrections to the theory of the electron.

The energy of a state can be expressed as a wave number, reciprocal (1 divided by the value) of the wavelength, that is, $1/\lambda$ (in which λ is the radiation wavelength), given in units of reciprocal centimetres, cm^{-1}. The two lowest excited states of the hydrogen atom are the ones labelled in spectroscopic notation $2^2S_{1/2}$ and $2^2P_{1/2}$, which lie about 82260 cm^{-1} above the ground state (state of lowest energy). According to Dirac's theory of the electron, these states should have the same energy and be about 0.365 cm^{-1} below the $2^2P_{3/2}$ state. In an experiment in 1947, two U.S. physicists, Willis Eugene Lamb and R.C. Retherford, showed that the $^2S_{1/2}$ state is higher in energy than the $^2P_{1/2}$ state by about 1,060 megahertz (0.035 cm^{-1}). This difference, which is called the Lamb shift, established that the Dirac theory required modification.

Lamb shift

In the experiment, a beam of hydrogen atoms was first formed by heating hydrogen molecules to a temperature of 2,500° K. Bombardment of the hydrogen atoms by electrons produced some atoms in the $^2S_{1/2}$ state, which survive for a lifetime of about 0.1 second before returning spontaneously to the ground state with the emission of a photon. This time was enough for most of them to reach the detector, a tungsten surface, where they produced a current by ejecting electrons from the surface. Between the source and the detector, however, the atoms were in a uniform magnetic field where they were irradiated (as in the C-labelled field of Figure 2) at the right microwave frequency to excite the $^2S_{1/2}$ state to the $^2P_{3/2}$ state. The latter state has a lifetime of only about 10^{-9} second, decaying to the ground state before reaching the detector and thus being unable to eject electrons at the detector. Thus a dip in the detector current showed when a resonance frequency had been reached. The resonance frequency, of course, gave the energy separation between the $^2S_{1/2}$ and the $^2P_{3/2}$ levels, which turned out to be higher than the Dirac theory predicted, higher by the amount called the Lamb shift.

Electric resonance

Apparatus of the general type shown in Figure 2 can be used for resonance experiments when the fields labelled A, B, and C are electric instead of magnetic. The molecules in one particular energy state are deflected in the A field, irradiated in the C field, where they may make transitions to other states, and then directed by the B field toward the detector. Because the resonance frequencies that cause diatomic molecules such as lithium fluoride to make transitions between different rotational states vary slightly from one vibrational state to another, the dipole moments for several of the lower vibrational states can be measured. The accuracy with which these measured differences are predicted serves as a sensitive test for theories of the electronic structures of molecules.

COLLISION PROCESSES

Crossed beam studies. Molecules that collide with each other are said to be scattered, and this scattering usually produces a change in the direction of relative motion of the molecules. If there is no change in speed as a result of the collision, the scattering is elastic; whereas if the speed changes, kinetic energy has been lost and the scattering is inelastic. If the atoms in the colliding molecules are rearranged as a result of inelastic collisions, the scattering is called reactive.

Scattering experiments with molecular beams provide a particularly direct way of studying the collisions between molecules because changes in direction and speed of the relative motion can both be observed. The apparatus required is similar to that shown in Figure 1 with the addition of a second beam intersecting the first in a small region of space called the scattering volume (Figure 3).

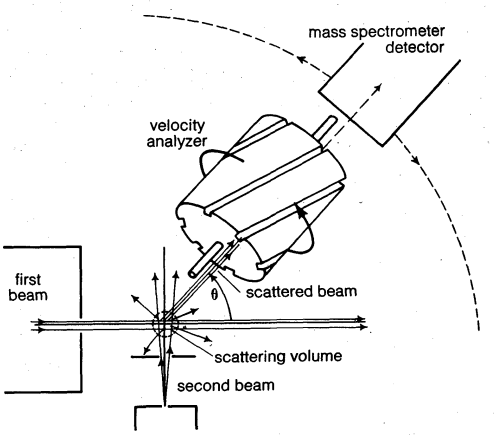


Figure 3: *Arrangement for a scattering experiment with molecular beams.*
A mass spectrometer used in tandem with a velocity analyzer measures the angular distribution, θ, of molecules scattered from intersecting beams (see text).

As the detector is moved around this centre, it can record the relative numbers of molecules scattered for different angles. Furthermore, if another rotor, like that used to select the speed for the beam in Figure 1, is inserted between the scattering centre and the detector, the distribution of speeds of the scattered molecules can be analyzed.

When molecules of one kind are scattered by molecules of another kind, the number of the first kind arriving in unit time at a detector placed at an angle from the original direction of the first beam is equal to a proportionality constant times the densities of the two kinds of molecules, the relative speed of the two kinds of molecules with respect to one another, the volume of the scattering region, and the solid angle subtended by (included by) the detector at the scattering volume. The constant of proportionality is called the differential cross section for that angle. The total cross section, which has units of area, is the integral or summation of the differential cross section over all angles; giving the cross section is a convenient way of describing a particular type of scattering.

The elastic scattering of one atom or molecule by another has been studied thoroughly. Most of the scattering produces deflections of small angles because the molecules pass each other at such relatively large distances that the forces between them are small. For some collisions, however, the molecules approach more nearly head-on so that they interact relatively strongly and scatter to large angles. The differential cross sections show a complex structure that often has a series of many maxima as the scattering angle is changed. The closely spaced maxima cannot be explained on the basis of classical mechanics; they show directly that even particles as large as heavy atoms behave like waves. Interpretation of scattering results like these indicates how the potential energy of interaction between molecules varies with their distance of separation.

Similar experiments, in which a second, analyzing rotor is used to measure the change in speed occurring as a result of inelastic collisions, show the likelihood for the transfer of varying amounts of kinetic energy of relative motion of the molecular pair to rotational or vibrational energy of one or both of the individual molecules. Inelastic collisions play an important part in the sequence of individual collisions that together make up the overall change called a chemical reaction.

Reactive collisions

Molecular beams have been of special use in the study of reactive collisions, but, of course, in this application the detector normally must be able to distinguish the reactants from the products. Mass spectrometers are frequently used as detectors for this purpose. In some experiments the amount of product formed is measured for a range of speeds of relative motion to see if there is a minimum speed below which reaction cannot take place. If there is one, this is the threshold below which the total cross section for reaction is zero. The amount of product, its angular distribution around the scattering volume, and the extent to which the molecules produced are rotating or vibrating can also be determined. Some reactions, like that of potassium atoms with bromine molecules, occur on almost every collision for which the molecules come within ten angstroms (10^{-7} centimetre) of each other, an unusually large distance compared to the maximum for most other chemical reactions. The threshold energy seems to be zero for this reaction, and most of the chemical energy released is converted to rotational and vibrational energy of the product, potassium bromide. For another reaction, that of methyl iodide with potassium atoms, measurements of the differential cross section for nonreactive scattering can be interpreted to show how the probability of reaction varies with the potential energy of interaction or the distance of separation. For this same reaction the methyl iodide can be oriented in space by the use of electric fields similar to those needed for the A field in an electric resonance experiment. The products, potassium iodide and methyl, are formed with a substantially greater probability when the beam of potassium atoms approaches the methyl iodide from the side of the iodine atom rather than from the side of the methyl group.

Beam experiments can also show if molecules react abruptly, so that the products are formed and separate directly after the first approach, or if the two reactants stay together as one aggregate at least long enough to make several revolutions about each other. In the latter case, the products are scattered symmetrically forward and backward along the original direction of approach.

Merging beam studies. One of the difficulties of experiments with beams is that of detecting slowly moving molecules. In most experiments with faster particles, which are easier to detect, the relative velocities, and hence energies, of beams intersecting at right angles are so large that they greatly exceed the relatively smaller energies of chemical reactions. A way of overcoming this problem is by the use of merging beams. In this experiment two beams of different molecules or atoms are separately ionized and accelerated in an electric field along the same path and in the same direction to similar but slightly different high speeds. Either or both beams may be neutralized by being passed through a dilute gas in which the fast particles of the beam transfer their charge to the slow molecules of the gas. In the merged beam the two species move at low speeds relative to each other, and so reactions between the molecules can be studied at low energies. The kinetic energy relative to the stationary detector remains large, however, and thus the result of reaction in the merged beam can be more easily detected than it can in the usual arrangement. This technique gives total but not differential cross sections.

INTERACTIONS OF BEAMS WITH CRYSTALS

When beams are used to study the interaction of individual atoms or molecules with the surfaces of solids, the results vary greatly as the energy of the beam varies. When an atom in the beam is moving rapidly, the relatively weak attractive forces existing between atoms have little effect on its trajectory. Only the much stronger repulsive forces that it encounters on approaching close to the nucleus of another atom are significant. Because of the extremely short range over which these forces are appreciable, the incoming atom interacts essentially with only one atom of the solid at a time. At each impact this atom from the beam loses some of its energy, but its large initial energy permits it to penetrate deeply into the structure of the solid. At lower energies the atom from the beam interacts at any given time with a larger number of atoms because of the longer range of the relatively weak attractive forces between atoms. At the lowest energies the incident atom may be trapped on the surface or be reflected from it, but the atom does not penetrate appreciably into the bulk of the solid.

Channelling. When a fast incoming atom hits a target that is an amorphous (noncrystalline) solid, the atom collides with successive atoms of the target much as it would in passing through a dense gas. The atom's trajectory is a random, zigzag one. When the target is a crystalline solid, however, the regular arrangement of its atoms makes penetration by the incoming atom relatively easy or difficult depending on the angle of its approach. When it enters a crystal in a direction that takes it through the more densely packed planes of atoms, the incoming atom moves less freely than it does when its direction is parallel to these planes of atoms. The spaces between these planes form channels. Moving in a channel the incoming atom makes glancing collisions with the atoms forming the walls of the channel, using up relatively little energy per impact. Thus it penetrates farther into the crystal than it would have if it had entered the crystal at a different angle. The idea of channelling offers possibilities for new interpretations of damage to crystals that have been exposed to radiation.

Crystal channels

Sputtering. When a fast atom from a beam hits a solid target such as a crystalline surface, the atom transfers some of its kinetic energy and momentum to the atoms of the solid. As a result, these target atoms may even receive enough momentum to leave the surface of the solid. This process is called sputtering of the target by the beam. It has been theorized that the atoms of the target recoil after being hit by the incoming atom and in turn hit other atoms of the crystal. These chains of successive impacts may be focussed along crystal directions in which the atoms are densely packed and ultimately cause the ejection of atoms when chains reach the surface of the crystal.

Diffraction. It was suggested in 1924 that beams of heavy particles such as atoms would behave like waves in which the wavelength could be given by Planck's constant divided by the momentum. This suggestion led to an experiment involving scattering a beam of helium atoms from a clean surface of a crystal of lithium fluoride. By the use of a sensitive pressure gauge, specially developed to detect the scattered helium, it was found that when the intensity of the scattered helium atoms was plotted, a strong scattered peak appeared at the angle corresponding to mirror-like reflection from the surface of the crystal. On both sides of this peak, however, there were smaller peaks exactly where they should have been if the atoms were behaving like waves. The helium had been diffracted by the crystal. This was an important confirmation of the wave nature of matter.

Gas-surface interactions. In spite of the successful observation of the diffraction of a beam of helium atoms from a crystal of lithium fluoride, further studies of the scattering of atoms of low energies from surfaces often led to confusing results. One of the main difficulties is that even in a good vacuum enough gas molecules are present so that any freshly prepared clean surface is quickly covered by unwanted impurities, such as water or carbon monoxide molecules, which often stick firmly and change the nature of the surface in an unknown way. The effects of these impurities are reduced substantially if a fresh surface is continuously created in the vacuum. One way to do this is to let a stream of atoms, such as those of silver vapour, condense steadily on the surface. The impurities are then buried beneath the much larger number of incoming silver atoms even though the latter are still few enough not to interfere with another beam, which is being reflected from the slowly growing surface. If the silver crystal grows on an underlying crystal, such as a piece of mica, chosen to have an appropriate spacing of its atoms, the mica directs the deposition of the silver, which is said to grow epitaxially; *i.e.*, in a pattern established by the underlying crystal. If a beam of helium atoms hits a silver surface of this kind, the helium atoms leave the surface as a directed beam, gen-

erally at the angle expected if the silver surface were to be considered a mirror. Diffracted peaks—that is, angular variations in intensity caused by different energies and crystal spacing—are not seen. A beam of ammonia molecules, on the other hand, which interacts much more strongly with the silver crystal, is scattered diffusely with no direction being preferred over another. Although the scattering of molecular beams from single crystals is one of the least ambiguous experiments that can be done with surfaces, the results are not well understood theoretically.

BIBLIOGRAPHY. K.F. SMITH, *Molecular Beams* (1955), a short, clearly written introduction; N.F. RAMSAY, *Molecular Beams* (1956), a comprehensive work dealing particularly with spectroscopy; C. SCHLIER (ed.), *Molecular Beams and Reaction Kinetics, Course 44: Proceedings of the International School of Physics "Enrico Fermi"* (1970), emphasis on molecular interactions and chemistry; J. ROSS (ed.), *Molecular Beams*, in *Advances in Chemical Physics*, vol. 10 (1966), on molecule scattering; AMERICAN ASSOCIATION OF PHYSICS TEACHERS, *Molecular Beams*, vol. 1, *Experiments with Molecular Beams*, vol. 2, *Atomic and Molecular Beam Spectroscopy* (1965), reprints of famous papers, particularly on spectroscopy, and a bibliography; O.R. FRISCH, "Molecular Beams," *Scient. Am.*, 212:58–74 (1965), a popular article.

(E.F.G.)

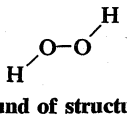
Molecular Structure

Atoms and molecules are the chief intellectual abstractions of chemistry. Atoms are the fundamental building blocks from which molecules are constructed, and molecules the smallest particles of a pure compound that retain its characteristic properties. Although there are only about one hundred different kinds of atoms (see ATOMIC STRUCTURE), they combine in tens of millions of different molecules (see CHEMICAL COMPOUNDS, INORGANIC; CHEMICAL COMPOUNDS, ORGANIC), and chemists add new kinds of molecules at a rate of more than 100,000 per year. Atoms and molecules undergo physical change without change in number or kind; atoms participate in chemical reaction without change in number or kind, but they can be transformed in nuclear reactions; molecules can be transformed in chemical reactions. It is a fundamental problem of chemistry to determine the arrangement of atoms in molecules and to elucidate the manner in which atoms are reorganized in the course of chemical reactions.

GENERAL CONSIDERATIONS

The simplest molecule is that of hydrogen, H_2, in which two hydrogen atoms are joined together. The minute size of these particles is suggested by the fact that one gram of hydrogen contains more than 10^{23} hydrogen molecules. Molecules range in size and complexity from hydrogen to molecules within living organisms that may contain many millions of atoms. In the case of two-atom, diatomic, molecules, such as those of hydrogen, nitrogen (N_2), oxygen (O_2), or hydrogen chloride (HCl), the only question about the molecular structure would be the distance between the two joined atoms. Adding an oxygen atom to the hydrogen molecule to make water (H_2O) raises several other questions. Water molecules could consist of chains with a terminal or a central oxygen atom, H—H—O or H—O—H; the chain could be straight or bent. If it is not straight, the angle of bend as well as the distances between the centres of the atoms must be known in order to specify the molecular structure. For molecules with larger numbers of atoms, three-dimensional shapes must be considered. While water, for example, has two hydrogen atoms attached at an angle of 104.5° to a central oxygen atom, hydrogen peroxide (H_2O_2) has two hydrogen and two oxygen atoms joined in a structure known to have the shape:

Structural possibilities with water



Historical background of structural concepts. Early in the 19th century an English chemist, John Dalton, proposed his atomic theory, and the precise distinction between atoms and molecules was first drawn in 1811 by an Italian physicist, Amedeo Avogadro. It was not until 1858, however, that his countryman Stanislao Cannizzaro, a chemist, showed that all physical and chemical facts confirmed the validity of Avogadro's conclusions and that the distinction reconciled many contradictory experimental results accumulated since the beginning of chemistry as a science. It was then recognized that molecules can possess three-dimensional shapes formed by connected (bonded) atoms. The contributions of another generation of chemists and a great deal of controversy were required, however, to ready the problem of molecular structure for solution, but the final steps in discarding the earlier, flat-molecule theory were taken by chemists Jacobus Henricus van't Hoff of The Netherlands and Joseph-Achille Le Bel of France. Van't Hoff's ideas, summarized in his book *Chemistry in Space*, were ferociously attacked by one of the leading chemists of the day. Thanks to this controversy the idea of geometrical relations within molecules became widely known and accepted.

During the same period, the concept that matter is in continual motion (kinetic-molecular theory) prevailed over the concept that matter is in a state of static equilibrium. Thus, the geometrical relationships of atoms within molecules cannot be considered as fixed. In addition to the movements of the molecules themselves through the gases and liquids that they constitute (translation), the various parts of the molecular structure itself are in continuous motion. This internal motion may take the form of vibrations in which atoms and groups of atoms move back and forth, increasing and decreasing the distances between them (stretching) and opening and closing the angles made by their bonds (bending). In complex molecules, groups may rotate about portions of the structure. These motions are fast and frequent, typical vibrations taking place on a timescale of 10,000,000,000,000 (10^{13}) per second; rotations are somewhat slower (typically 100,000,000,000 [10^{11}] per second). Because of this flexibility, molecular dimensions are taken as time-averaged values of the positions of the atoms. At higher temperatures, the atoms take longer and more violent excursions from their equilibrium positions, and new patterns of motion sometimes become possible. Molecular structure thus depends upon temperature.

Internal motions of atoms

In addition, the exact dimensions of molecules vary according to the situation in which they are studied. Thus, somewhat different measurements are obtained for molecules in the three phases, solid, liquid, and gas, or dissolved in solution or adsorbed on a surface. Drastic changes from the normal form have been found in determining structures of molecules carrying excess charges (called ions or radicals) and those possessing excess energy (excited-state species).

Structural theories. *Quantum mechanical approach.* No truly adequate theory of molecular structure is yet available. The detailed structure of complex molecules cannot be predicted but must be obtained from experiment. Recent studies, however, have been specifically directed toward the development of theories that could yield structural information.

One approach is through quantum mechanics (see MECHANICS, QUANTUM), in which an orbiting electron is considered to have the characteristics of a wave; treating the electron standing waves (atomic orbitals) mathematically allows determination of how pairs and groups of atoms come together to form joint orbitals (bonds). Such an approach relies on data from experiment in conjunction with approximate mathematical methods utilizing high-speed computers, but only the simplest systems can be calculated exactly. To achieve a strong bond, the electron orbitals of the atoms involved must overlap in space as much as possible (see CHEMICAL BONDING); and since atomic orbitals are themselves directed in space, the resulting bonds must conform to their arrangements. The quantum-mechanical approach leads to correct predictions concerning the shape of only simple molecules; for example, the bent shape of the water molecule.

Electrostatic approach. An alternative approach is to consider the molecule as a system in electrostatic balance

—the attraction between electrons and their nuceli (because they have opposite charges) tending to be maximum and the repulsion between electrons (because they carry the same charge) tending to be minimum. The attraction between each nucleus and the electrons of an adjacent nucleus is also of primary importance in bonding atoms. In this view, the most likely shape of the molecule will be the one that best meets these electrostatic requirements. Those pairs of electrons farthest from the nucleus of a central atom of a group are considered to be important in determining the molecular structure of the group. To minimize the repulsion between centres of negative charges, the system adopts the most balanced arrangement possible; when there is more than one pair, each takes up a position as far as possible from the others, while remaining as close as possible to the nucleus. Two electron pairs thus place themselves on opposite sides of the nucleus to give a linear (180°) relationship; three pairs lead to a trigonal planar arrangement, with bond angles of 120°; four to a tetrahedral arrangement, with angles of 109°28′; six to an octahedron, with 90° angles; etc. The electron pairs can be either shared with a second atom (bonding) or lone pairs (nonbonded). Shared pairs are drawn out between two positive centres and lie farther from other pairs of electrons on the same atom, and their repulsive effect is thus somewhat reduced. Unshared pairs, on the other hand, are under the influence of only one positive centre and, because of their proximity, create the most severe repulsions. The relative repulsive power is thus, in descending order, lone pair–lone pair, lone pair–bonding pair, bonding pair–bonding pair; this is also the order in which distortion (opening) of bonding angles has been found experimentally for molecules taking the regular geometric shapes listed above.

Shared and unshared pairs

No perfectly regular solid figures exist with five or seven points (or vertices); the electron pairs are, however, well separated in the semiregular trigonal and pentagonal bipyramidal shapes, and there are many molecules of those kinds.

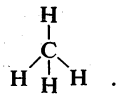
The structure adopted by a molecule must also meet the spatial requirements of its constituent atoms and groups of atoms. The exact dimensions of atoms are not known, but the volumes occupied by given numbers of atoms are largely incompressible. No amount of force generated in one part of a structure can bring about interpenetration of atomic volumes in another. These spatial requirements are a controlling factor in molecular geometry.

MOLECULAR BONDS AND SHAPES

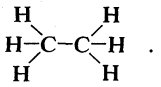
Spatial arrangement of atoms. The basic laws of chemistry give the number of attachments, usually variable, that each atom is capable of making with other atoms. The restriction against atoms occupying the same volume guides their packing together to fill space. Distances and angles between atoms are adjusted to maximize attractive (bonding) interactions and to minimize electrostatic repulsions. On these principles a bewildering variety of molecular shapes and sizes are constructed. The range of structural possibilities is difficult to comprehend and impossible to represent adequately on the printed page, first because most molecules have three-dimensional shapes and, second, because all are in rapid, dynamic motion. Static, two-dimensional figures in which atomic symbols are placed at equilibrium positions in the structure are nevertheless the most useful way of representing molecular structures.

Chains. The chemistry of carbon, the archetypal participant in shared-electron-pair (covalent) bonds, is the basis of organic chemistry. Carbon and hydrogen together form a fantastic number of compounds; those containing only the two elements are called hydrocarbons. These have derivatives that contain other elements as well. All are called organic compounds, because it was thought at one time they could be made only by living organisms. The simplest hydrocarbon is methane, with the formula CH_4, in which carbon directs four equivalent bonds toward the vertices of a tetrahedron, so that methane can be represented structurally as:

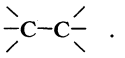
Organic compounds



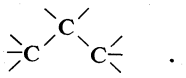
Ethane, with an empirical formula of C_2H_6, can be considered to be constructed by replacing one of the hydrogen atoms of methane by a CH_3 group, thus



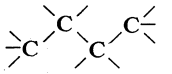
This arrangement can also be written without showing the hydrogens:



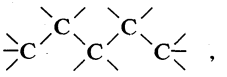
Repeating the same operation and replacing a hydrogen atom of ethane by another CH_3 group leads to propane, C_3H_8:



A hydrogen atom on a terminal, or end, carbon can again be replaced with a CH_3 group, to produce butane, C_4H_{10}:

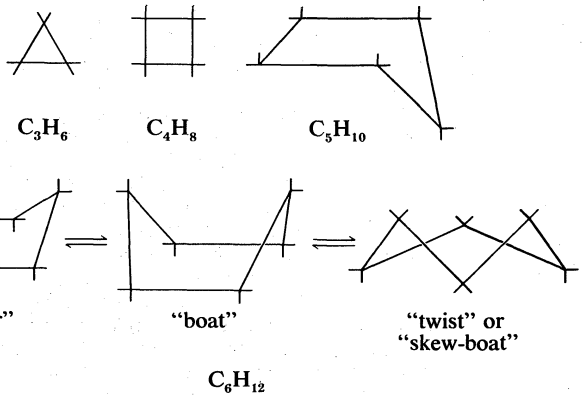


from which the same process produces pentane, C_5H_{12}:



etc. This series of hydrocarbon compounds in which the carbon atoms always form a straight chain can be extended indefinitely. In actuality, the light, shorter chain hydrocarbons are derived from natural gas, which is a mixture of straight-chain hydrocarbons having one to four carbon atoms (in abbreviated form, C_1 to C_4); bottled fuel gas is usually C_3 or C_4; gasoline C_6–C_{10}, kerosene C_{10}–C_{18}, paraffin waxes $C_{>20}$; and polyethylene consists of chains of 5,000 or more of the same CH_2 units that first occur in propane (see HYDROCARBONS).

Rings. Compounds that consist of ring structures of various sizes can be formed by connecting the ends of chains, as in C_3 to C_6 hydrocarbons represented diagrammatically below:



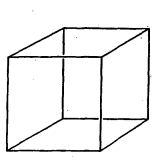
The names of ring compounds are made up of their straight-chain names combined with the prefix cyclo-. Cyclopropane, C_3H_6, is highly reactive, because the interior 60° angles of the three corners bring electrons that are mutually repulsive into close proximity. The cyclobutane (C_4H_8) molecule, with interior angles of 90°, is also easily decomposed. A pentagon has interior angles of 108°, close to the tetrahedral angles of 109°28′, which are the normal bonding angles of carbon, and, therefore, cyclopentane (C_5H_{10}) is quite stable. The puckered form that it adopts minimizes hydrogen contact. It might be thought that, as larger numbers of carbon atoms join rings, the molecules would become less stable again, since

Puckered rings

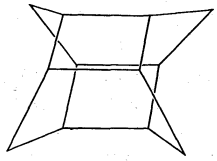
the interior angles of the planar figures become larger than 109°28′; for example, 120° in the hexagon. The problem does not arise, however, because of the possibility of reducing the angles by puckering the structure and eliminating a flat configuration. The C_6 cyclohexane molecule (C_6H_{12}) can exist in a "chair," a "boat," or a "twist," or "skew-boat," form, all involving tetrahedral angles at the carbon atoms while minimizing hydrogen-atom repulsions. The molecules flip rapidly from one form to another at room temperature. Stable hydrocarbon rings with more than 30 carbon atoms are known.

More complex ring types are represented by bicyclobutane and bicyclooctane and the fused-ring systems such as those found in steroids.

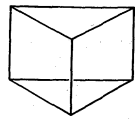
In recent years, chemists have synthesized a spectacular number of hydrocarbons with unusual polyhedral structures. Some examples, with their highly descriptive, nontechnical names, are given below:



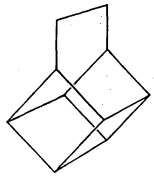
cubane



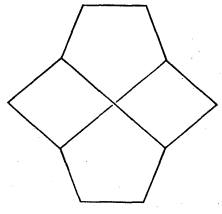
asterane



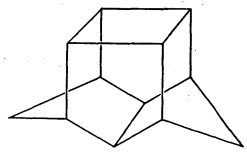
prismane



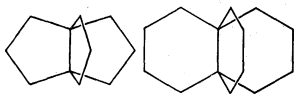
basketane



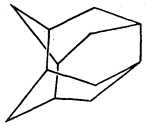
twistane



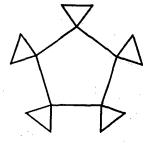
birdcage



propellanes

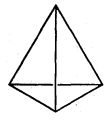


iceane

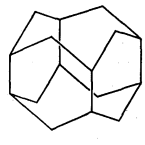


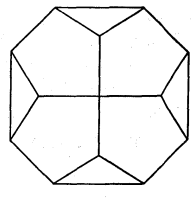
paddlewheelane

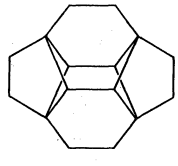
Efforts are under way to add the following:

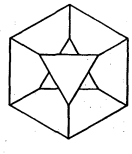


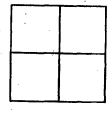
tetrahedrane





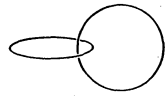






windowpane

The idea of what constitutes a molecule has itself been modified by the synthesis of linked rings:



Silicon and oxygen in minerals

As carbon and hydrogen form the basis of the organic world, so are silicon (Si) and oxygen (O) the basic elements of the mineral world, and the linkages represented by $-\overset{|}{\underset{|}{Si}}-O-$ form the backbone of molecular structures. Small numbers of these individual oxygen–silicon units are found in the soft rocks, such as talc; long chains of them give rise to fibrous minerals such as asbestos; planar sheets form the two-dimensional structure of mica, while quartz contains the same $-\overset{|}{\underset{|}{Si}}-O-$ units cross-linked into three-dimensional network arrays. Lately, materials have been synthesized that can be viewed as the organic derivatives of these rock forms. In one example, CH_3 groups are bonded to silicon atoms, which are linked through oxygen into long chains. These materials, called silicones, combine the advantages of the strong oxygen–silicon backbone with the tractable nature of organic compounds.

Chelates. Hydrocarbons can be modified by the substitution of other atoms for carbon. When a carbon atom in a ring is replaced by another kind of atom, the compound produced is called heterocyclic, and it adopts conformations that accommodate the particular spatial, or steric, requirements of the heteroatom, the noncarbon atom in the ring. Incorporation of heteroatoms and groups into hydrocarbon molecules often confers special properties, including the ability to form additional bonds, typically to a charged metal atom (ion), in what are called coordination compounds (*q.v.*). Important examples are provided by molecules with two or more heteroatom sites juxtaposed spatially in such a way as to enable them to attach simultaneously to a single metal ion. Such compounds, called chelates, after the Greek for "crab's claw," can be synthesized in bidentate (two active sites), tridentate, etc., varieties. If the spacing between the heteroatoms is just appropriate to hold a metal ion of a certain size, the chelating action will be selective. The hexadentate compound ethylenediaminetetraacetic acid (EDTA), for example, has its six points of attachment so arranged that it can wrap itself around a lead atom and remove that element from the body in cases of lead poisoning. Other complexing agents of biological importance include heterocycles fused into large, flat rings, leaving a central hole into which the heteroatoms protrude. The porphyrins are such a class, with four nitrogen atoms in a central opening just large enough to fit a metal ion. The biologically essential molecules hemoglobin and chlorophyll consist of porphyrin rings in which iron and magnesium are held, respectively. Recently, chelating agents have been synthesized that mimic the natural materials but differ from them in potentially important ways.

Heteroatoms

Novel shapes. Other novel molecular shapes are found in compounds in which metal atoms are sandwiched between the planar faces of hydrocarbon rings; for example, ferrocene, $(C_5H_5)_2Fe$, and dibenzenechromium, $(C_6H_6)_2Cr$, in which the ring planes are parallel, and stannocene, $(C_5H_5)_2Sn$, with an angular sandwich structure. These and other synthetic organometallic compounds are the fruits of attempts to break elements out of intractable, inorganic compounds such as salts and oxides or tough metallic forms and to incorporate them into molecules with useful and interesting properties. The operation involves the replacement of the inorganic or metallic lattice bonds with bonds to hydrogen or organic groups. The element boron, for example, forms an icosahedral structure (having 20 faces), and a hydrogen derivative of this structure has been prepared, with the formula $B_{12}H_{12}^{2-}$, along with several smaller borane molecules, which can be viewed as derivatives of icosahedral boron fragments. Carboranes, with formulas such as $B_{10}H_{10}C_2H_2$, have two icosahedral vertices substituted by carbon. In a similar way, there have been syntheses of organic derivatives of a large number of the familiar metals that preserve the clusters of atoms found in the metals themselves.

Polymers. When a simple molecule is bonded to itself or to a few other simple molecules to form long chains, the result is a polymer, or macromolecule. Polymers constitute a special class of molecular structure because of their giant size and because any given sample is always a mixture of molecules of various sizes. It is possible to distinguish one-dimensional, chain, or linear, struc-

Spatial structures

tures, such as the natural polymers in unvulcanized rubber, silk, and asbestos and the artificial polyethylene and silicone fluids; two-dimensional macrostructures such as those formed by mica and graphite; and three-dimensional structures such as quartz or diamond. In fact, a perfect crystal of diamond (one of the forms of carbon) could be considered to be a single, giant molecule, owing its extraordinary hardness to the four strong electron-pair bonds radiating from each carbon atom to four other carbon atoms. Some of the polyhedral hydrocarbon compounds mentioned above can be considered to be fragments of the diamond structure. The constituent units of a linear polymer may be arranged at random, as was the case of all artificially produced polymers made before 1954, or in some highly ordered manner, as is the case of materials such as natural rubber and the artificial materials produced after the discovery (1954) of a special type of polymerization, called stereospecific.

Protein structures

Special problems in molecular structure arise with proteins. The basic constituents of proteins are molecules with two different group properties (bifunctional molecules). These properties are provided by two functional groups attached to the molecule, in this case an acid group and a base group, in compounds called α- (alpha-) amino acids. The acid (carboxyl) end of one such molecule can act on the base (amino) end of another to form a molecule of water, which is released, and a bond called the peptide link between the remaining parts of each molecule. Using R— to represent the body of an α-amino acid, the formula is $H_2N{-}CHR{-}COOH$, H_2N being the amino group and COOH the acid group. The reaction between any two amino acids (R_1 and R_2) is:

$$H_2N{-}CHR_1{-}COOH + H_2N{-}CHR_2{-}COOH$$

$$\longrightarrow H_2N{-}CHR_1{-}CONH{-}CHR_2{-}COOH + H_2O.$$

Repeating this process over and over, each time eliminating a molecule of water, produces a polypeptide chain. The various amino acids are distinguished by the nature of the R groups; because 23 amino acids have been recognized as constituents of proteins, the selective order in which they polymerize is, practically speaking, infinite in number.

The order of the amino acids in the polypeptide chains is called the primary structure, and the first such sequence was determined in 1952 by an English chemist, Frederick Sanger, for the protein insulin. The regular way in which the polypeptide chains are arranged in space to form a protein molecule is called the secondary structure. The principal type of secondary structure is called the alpha helix. The polypeptide chain can be folded into the spiral form with about four amino acid units per turn. Many fibrous proteins, including hair and muscle, consist of these polypeptide helices twisted about one another to form cables. In 1960 investigation of the protein myoglobin revealed that its single polypeptide chain coils into eight short, helical segments separated by nonhelical sections. This arrangement of the three-dimensional structures of the polypeptide chain in space is the tertiary structure, while the combination of two or more polypeptide chains constitutes the quaternary structure.

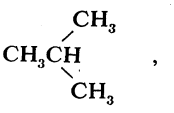
Isomerism. The possibilities for variety in molecular structure are enormously expanded by the freedom to arrange the same group of atoms in more than one way to form different structures, a phenomenon called isomerism. The properties of a compound are determined not only by its chemical composition but also by the arrangement of its constituent atoms in space. Two substances are said to be isomers if they have identical chemical composition and molecular weight but differ in molecular structure (see ISOMERISM).

Early studies of the structure of cyanates

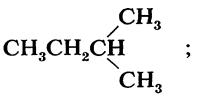
The first such sets of compounds identified were the explosive fulminates (salts of fulminic acid) prepared in 1824 by the German chemist Justus von Liebig and the inactive cyanates (salts of cyanic acid) prepared in the following year by another German chemist, Friedrich Wöhler. At first, each man thought the other wrong, since the formulas for the two salts proved to be the same, but they soon met and in a common series of experiments showed that both parent acid compounds possessed the same atoms, of carbon, hydrogen, oxygen, and nitrogen. The French scientist Joseph-Louis Gay-Lussac correctly concluded that the spatial arrangement of atoms in the two compounds must be different. Cyanic acid has the formula HOCN, while fulminic is HONC. It remained for Jöns Jacob Berzelius in Sweden to formulate a precise definition of the concept of isomerism.

The subject is best illuminated by listing briefly some of the many types of isomerism.

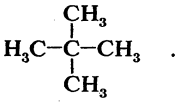
Branching. In building up the straight-chain hydrocarbons from methane as shown above, a CH_3 group was substituted for a terminal hydrogen atom in each step. Substitution at an interior atom would produce a hydrocarbon of the same formula but with a branched structure, as in isobutane, C_4H_{10}:



compared with *n*-(normal) butane: $CH_3CH_2CH_2CH_3$, also C_4H_{10}. Pentane, C_5H_{12}, has three isomeric forms: *n*, $CH_3CH_2CH_2CH_2CH_3$; iso,



and neo,

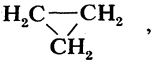


Even relatively simple molecular formulas permit enormous numbers of differently branched structures; for example, it can be calculated that $C_{20}H_{42}$ can exist in 366,319 isomeric forms and that $C_{40}H_{82}$ has 62,491,178,805,831 isomers. No chemist has been sufficiently interested to attempt to isolate the complete set beyond C_9H_{20}, for which all 35 different structures are known.

Positional. A noncarbon atom or group, symbolized as X, can be substituted for hydrogen that is attached to any carbon. The *n*-butane formula, for example, is C_4H_{10}, and C_4H_9X is the empirical formula for any substitution compound. The X may be attached to any carbon atom, and, thus, the substitution compounds derived may be $CH_3CH_2CH_2CH_2X$ and $CH_3CH_2CHXCH_3$; because X occupies different positions, each of these compounds differs from the other.

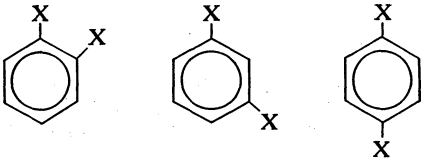
Functional. Ethyl alcohol has the same formula as methyl ether, C_2H_6O, but their structures are quite different, being CH_3CH_2OH and CH_3OCH_3, respectively.

Ring versus unsaturated. The formula C_3H_6 can belong to the cyclopropane ring,



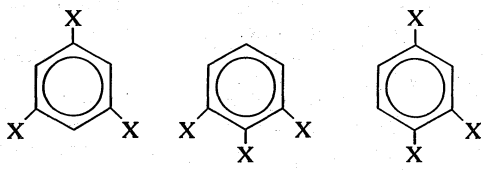
or to the unsaturated (containing one or more double or triple bonds between carbon atoms) propylene, $CH_3CH{=}CH_2$, which contains a double bond. A hydrocarbon containing a triple bond is isomeric with members of three other classes of compounds: those containing two double bonds, those containing two rings, or those containing one ring and one double bond.

Ring. The derivatives of benzene, C_6H_6, which is planar, with two hydrogens substituted (disubstituted), with the formula $C_6H_4X_2$, give three isomers:



ortho- *meta-* *para-*

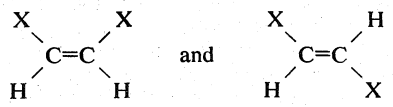
The trisubstituted $C_6H_3X_3$ can also take three forms:



sym- *vic-* *asym-*

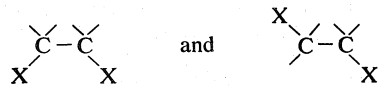
Geometrical. The double bond also holds its substituents in a flat plane. The substituted ethylene, $C_2H_2X_2$, in which each X is attached to a different carbon atom, can exist in two isomeric forms; in one the X groups are on the same side of the double bond (*cis-*) and in the other on the opposite sides (*trans-*), thus:

Cis- and *trans-* forms



cis- *trans-*

The two forms are distinct because rotation of the two halves around the double bond is restricted, unlike the case of the single bond, about which there is a possibility of free rotation. A similar geometric relationship can exist in substituted ring compounds:



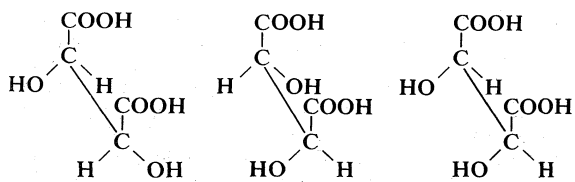
cis- *trans-*

Positional isomers are available for the above examples in which each X is located on the same carbon atom.

Iso-, syndio-, atactic. The pendant, side-chain groups attached to a linear-chain carbon polymer backbone can all be oriented in one direction (iso-); alternately, in each of two directions (syndio-); or randomly (atactic).

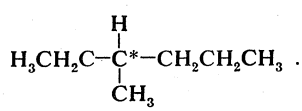
Optical. Optical isomerism occurs when two molecules differ only in that one is a mirror image of the other; *i.e.*, when they are enantiomorphs and exhibit what is called right- and left-handedness. The structures of enantiomorphs cannot be superimposed as a consequence of their asymmetry. The two forms of tartaric acid, first investigated by the French scientist Louis Pasteur, constitute a molecular example, while a third form is symmetrical and, therefore, not optically active. The structural formulas for these are:

Right- and left-handedness in chemical compounds

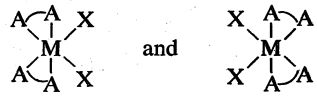


(i) (ii) (iii)

Pasteur was able to identify right- and left-handed crystals of sodium ammonium tartrate under the microscope and to pick the types apart. Most properties of enantiomorphic pairs are identical, but they always differ in that one turns the plane of polarized light clockwise (dextrorotatory), while the other turns it counterclockwise (levorotatory). Pasteur was able to correlate the optical activity of the tartrates with the asymmetry of the crystals. Only a very few examples of the spontaneous separation of optical isomers on crystallization have been discovered to date. The asymmetry that gives rise to optical activity, however, is present in every compound containing a carbon atom with four different groups attached, the simplest such hydrocarbon being hexane with a methyl group attached to the third carbon,



The starred carbon atom is said to be the asymmetric centre. Among inorganic compounds, octahedral *cis-* $M(A{-}A)_2X_2$ exists in two optical isomers: (see ISOMERISM).



An extraordinary fact is that only the left-handed enantiomorphic amino acids in which the α-carbon is in general an asymmetric centre have been found to occur in plant and animal proteins. A man constructed of right-handed proteins could make use of the optically inactive oxygen and water but would not be capable of digesting ordinary food or begetting children with an ordinary wife. No one knows why proteins are constructed of left-handed amino acids.

TIME-DEPENDENT PROPERTIES

Motions preserving molecular integrity. The rather static world of experience is, at the molecular level, only the blur of ceaseless, dynamic motion. Gas molecules in a room atmosphere have speeds of 10^4–10^5 centimetres per second (200–2,000 miles per hour) and collide with each other about 10^{10} times per second. Since a typical frequency of rotation is 10^{11} times per second, such a molecule completes ten rotations between collisions. Molecular bonds typically vibrate with a frequency of about 10^{13} per second, and, therefore, bonds vibrate about 100 times during each molecular rotation and about 1,000 times between collisions in the gases in a room atmosphere.

Molecular motion in liquids

Molecular motion in the liquid phase is more difficult. Translation is impeded because the molecules are tightly packed. Even rotations will be relatively rare unless the molecules are almost spherical or are nearly cylindrical about one axis. Rotations tend to be transformed into oscillations or twisting motions and the translations into a rapid back and forth bouncing off nearest neighbours. The frequency of collision is great, about 10^{12} per second, because the molecules are so close; the bond vibration frequencies, however, will still be higher by almost tenfold.

Translation ceases on solidification; molecules oscillate and, where space permits, rotate about fixed equilibrium positions.

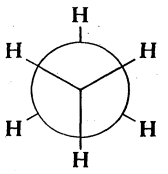
Flexibility is, then, a chief characteristic of molecules, and it is restrained only by powerful forces. The various types of molecular motions that contribute to this flexibility are summarized below.

Vibration. A diatomic (of two atoms) molecule can vibrate in only one way. This vibration consists of an in-phase motion of the two atoms toward and away from their centre of mass along the internuclear axis (comparable to the motions of a pair of weights connected by a spring). By in-phase is meant that the atoms pass through their mean positions and reach their turning points simultaneously. Such vibrations along bond axes are called stretching motions. The amplitude (magnitude or extent) of such motion is typically about 10 percent of the bond length.

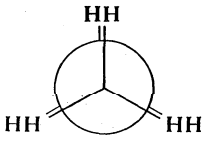
A bent, triatomic (of three atoms) molecule, such as water, HOH, can execute three fundamental modes of motion: a concerted compression and stretching of the two O—H bonds: an asymmetrical motion by which one O—H stretches while the other compresses; and a bending motion in which the angle opens and closes. If such a triatomic unit is held by its centre atom in a more complex system, this last scissoring motion will be accompanied by various twisting, wagging, and rocking torsions as well. Planar (in one plane) molecules undergo out-of-plane deformations of terminal atoms; cyclic molecules undergo simultaneous stretching apart of the ring atoms (called breathing). Bending requires less energy than stretching.

Rotation. Detailed consideration of the rotational motions of groups within molecules led in 1937 to the recognition of several new possibilities. Viewing the ethane molecule, C_2H_6, for example, down its central C–C axis,

reveals two extreme positions for the hydrogen atoms:



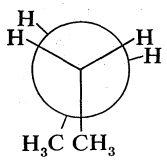
staggered, and



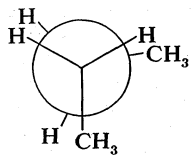
eclipsed. The onset of repulsions when the hydrogens pass each other creates slight barriers to completely free rotation about the C–C bond. The exact cause of the restricted rotation is not known, but the repulsions are lowest in the staggered form. Rotation from one staggered configuration (called rotamer) to another is extremely rapid at room temperature.

Rotamers

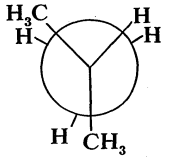
The barrier to free rotation is raised when one hydrogen on each ethane carbon is replaced by a CH_3 group to give *n*-butane. Six rotameric forms can now be drawn, with the barrier decreasing from the form in which the two CH_3 groups are eclipsed to the most favourable staggered form with the two groups farthest apart. In the structural forms below, rotamers (ii) and (iii) and (iv) and (v) constitute enantiomorphic pairs:



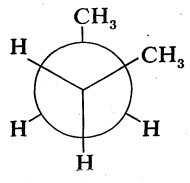
(i)



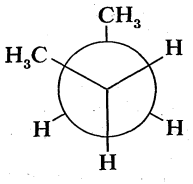
(ii) eclipsed



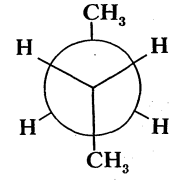
(iii)



(iv)



(v) staggered



(vi)

Rotation at room temperature is too rapid to permit isolation of the individual rotamers.

Another form of motion, called pseudorotation, was postulated in 1960 to explain the rapid interchange of the positions of atoms in trigonal bipyramidal molecules. A similar twisting of CH_2 units is now thought to be responsible for the easy interconversion of conformation in cyclic hydrocarbons; *e.g.*, in cyclohexane (see above). Such pseudorotations have very low energy requirements and take place easily and rapidly at room temperature. The individual conformers of cyclohexane (boat, chair, and twist, or skew-boat) cannot be isolated.

Inversion. Ammonia, NH_3, and its substituted derivatives, called amines, are pyramidal molecules. If the three groups attached to nitrogen are different, two forms with nonsuperimposable mirror images should exist. No such optical isomer has ever been isolated, however, because pyramidal molecules of this type undergo a rapid motion by which the pyramid becomes inverted. In this vibration the nitrogen atom oscillates through the plane of the three attached groups, much as an umbrella can turn inside out. The inversion frequency of ammonia is over 10^{10} times per second. Substitution of bulky groups for hydrogen slows the inversion rate somewhat.

Fluxional molecules. Several organometallic molecules are now known in which a metal atom rapidly changes its point of attachment from one to another of the carbon atoms of a hydrocarbon ring or chain. In the first recognized example, mercury (Hg) in $(C_5H_5)_2Hg$ was shown in 1954 to exchange its bonding position about the five carbons of the C_5H_5 ring. The bromine–magnesium (BrMg) unit in $CH_2{=}CHCH_2MgBr$ appears bonded to one, then another of the terminal carbon atoms. Such nonrigid structures have been described as "ring whizzers" or "chain swingers" or, more simply, as fluxional.

Nonrigid structures

It should be recognized that, despite the possibility of more complex flexible torsions, some types of which can occur simultaneously, molecular vibrations involving each bond and group of bonds are always in progress. Each of the motions described above occurs with a characteristic frequency, energy, and rate.

Motion in which molecular integrity is disrupted. *Dissociation.* Stretching movements of sufficient energy will disrupt the bond between two vibrating atoms, with several possible results, depending upon the final disposition of the electrons in the broken bond.

In heterolytic fission (breaking into differing parts) of a bond, both electrons remain with one fragment, making it an ion with one excess negative charge, while the other fragment, lacking an electron, becomes an ion with one positive charge. In water, hydrogen is a typical positive ion (H^+); molecules releasing H^+ are called acids. Hydroxide ion (OH^-) is a typical negative ion in water; molecules releasing OH^- are called bases. In water these ion products of heterolytic fission are separated and stabilized by oppositely charged centres in the medium, the ionic molecules, like other charged particles, obeying the physical law of being attracted by opposite and repelled by similar charge.

Homolytic fission (separation into two neutral atoms or groups) of bonds leaves each fragment with a balance of charges, but each possesses an odd number of electrons. Such fragments, called radicals, are usually reactive since they can readily form new electron pair bonds with other odd-electron species. Homolytic fission of bonds occurs when molecules are subjected to heat or radiation.

Addition or coordination compounds are made up of constituent molecules that can exist independently and that return to their independent state upon dissociation of the complex. One example is the complex formed reversibly with oxygen by hemoglobin, which transports oxygen through the bloodstream. Dissociation of the complex in the tissues releases oxygen molecules for oxidation of foodstuffs and frees hemoglobin for return to the lungs and repetition of the cycle.

Exchange reactions. Change in position among easily displaced atoms or groups in gases and liquids is common and can be assumed to take place at some rate among most linkages. Readily changed hydrogen linkages represent an extreme case; the exchange of hydrogen between H_2O molecules in liquid water is particularly fast.

Interconversion of isomers. The pure substance acetaldehyde (C_2H_4O) furnishes an example of tautomerism, a condition in which isomers change into each other easily and ordinarily exist together in equilibrium. It has the chemical and physical properties expected for a mixture of two different structures, one with a carbon–carbon double bond ($\diagdown C{=}C \diagup$) and an oxygen–hydrogen (OH) group and the other with a carbon–carbon single bond ($-\overset{|}{\underset{|}{C}}-\overset{|}{\underset{|}{C}}-$) and a carbon–oxygen ($\diagdown C{=}O \diagup$) double bond. The two isomers undergo rapid interconversion within the molecule through exchange of hydrogen between carbon and oxygen in the structure to give the keto ($CH_3{-}CH{=}O$) and enol ($CH_2{=}CH{-}OH$) forms. These changes take place too rapidly to permit separation from the mixture under ordinary conditions.

Isomerization. In many cases the interconversion of isomers is so easy that the individual isomers cannot be isolated. Optical isomers capable of fast, reversible transformation into a symmetric arrangement, for example, will always exist as an equal mixture of dextrorotatory and levorotatory forms; *e.g.*, the enantiomorphic rotamers of ethane. Tautomeric exchange of hydrogen is

within the molecule and fast. Just as fast is the migration of metal atoms in the constantly changing fluxional molecules. At the other extreme lie hydrocarbon and oxygen–silicon isomers, which are stable because the rupture of several strong, inert bonds is necessary to rearrange the structure; for example, mirror image (enantiomorphic) crystals of quartz have retained their identity for eons. Between the two extremes are many molecular rearrangements of stable substances that can be brought about under mild conditions by heat or chemicals.

MOLECULAR STRUCTURE AND GENERAL PROPERTIES OF MATTER

The bulk materials of experience are made up of tremendous numbers of atoms or molecules, and the macroscopic properties of materials can usually be generalized from the properties of the constituent units. Behaviour of molecular materials can be understood through knowledge of molecular structure and intermolecular forces. In this way, the nature of known materials is learned and possibilities for new materials realized.

In the following section the physical and chemical properties of molecules as they can be understood from molecular structure are dealt with.

Physical properties. Those characteristics of molecules generally classified as physical are due to molecular size and shape, forces between molecules, and their interaction with electromagnetic radiation and fields.

Size and shape effects. A collection of molecules fits into a regular crystalline arrangement according to the same rules of geometry as the packing of large objects. Molecular crystalline arrangements conform to various shapes known in solid geometry. Bulk crystalline form can also reflect the shape of the molecules making up the crystal, as demonstrated by Pasteur's identification of right- and left-handed crystal shapes composed of enantiomorphic molecules, solutions of which rotate the plane of polarized light in opposite directions. Long-chain molecules often form fibrous crystals, and sheet-like molecules form layer lattices. Other bulk properties are similarly related to molecular characteristics.

Structure of rubber

Rubber, for example, is an aggregate of very long molecules intertwined in a random way. Unvulcanized rubber is sticky because the molecules tend to pull away from each other. Vulcanized rubber has been heated with sulfur, which forms bridges of sulfur-atom chains linking the rubber molecules together; small amounts of sulfur (few cross-links) lead to soft rubber, larger amounts of sulfur to much harder materials. Rubber's elasticity arises from the reversible stretching of the randomly twisted and coiled molecules, under stress, into a more ordered, chainlike arrangement.

Size and shape are important in other respects; for example, in diffusion of molecules through channels in rigid mineral forms or in nonrigid, biological membranes. Only a single dimension may be critical, as in the case of a rodlike molecule passing through an orifice or of a long-chain structure passing through a membrane channel that its branched isomer cannot penetrate.

Intermolecular forces (see following discussion) determine whether a substance will be found in gas, liquid, or solid phase. Strong attractive forces result in solidification, even for the lightweight lithium hydride, while weak attractive forces between the much heavier radon atoms are easily disrupted, and radon exists as a gas at room temperature. When intermolecular forces are relatively constant through a series of molecules, their size and shape can determine the phase. The saturated hydrocarbons are an example: methane, CH_4, is a gas; isooctane, C_8H_{18}, is a liquid; and triacontane, $C_{30}H_{62}$, is a waxy solid. Shape is also important. The chain molecule decane, $C_{10}H_{22}$, is a liquid at room temperature, but the flat, fused-ring naphthalene, $C_{10}H_8$, of similar mass, packs efficiently into a crystal and is solid under the same conditions. Attaching one carbon–hydrogen group (CH_3) to naphthalene can disrupt the layer packing and make the methyl derivative a liquid at room temperature.

Intermolecular forces. At very short distances, all atoms and molecules attract each other. In the absence of other more powerful forces, weak attractions account for liquefaction and, as temperature is lowered, eventual soldification of subtances. At the other extreme are the very powerful forces of electrical attraction between oppositely charged ions. Crystalline lattices are typical of ionic aggregates. Between the two extremes are electrically neutral molecules with, nevertheless, concentrations of negative charge in one part of the molecule and of positive charge in another part; such molecules are electric dipoles (just as bar magnets are magnetic dipoles) and are called polar, as compared with nonpolar molecules, in which the charges are evenly distributed. The oppositely polarized ends of two molecules will attract one another. Molecules consisting of two atoms of the same element, $A-A$, are nonpolar. On the other hand, all heteronuclear (containing different kinds of atoms) diatomic molecules, $A-B$, are polar, since the electrons joining A and B will inevitably lie closer to one atom or the other, thereby distorting the electrical balance of the system. The presence of the electric dipole can be experimentally verified, since in an electric field the molecules, like the magnets, will align themselves with the electron-rich end toward the positive pole. The extent of the imbalance, called the dipole moment, is a vector quantity (*i.e.*, it has both magnitude and direction), and vector addition can be used to relate the measured dipole moment of a complex molecule to the individual contributions of the bonds that make it up. Linear $A-B-A$ molecules made up of polar bonds between oxygen and carbon, O=C=O, for example, have no net polarity because the two equal and oppositely directed dipolar C=O vectors cancel each other. The polarity of the water molecule, on the other hand, was used by a Dutch scientist, Peter Debye, to show that the H—O—H system is bent, since only in that shape could the two dipolar O—H vectors add to give a net dipole moment for the molecule.

Polar molecules

The electron-rich ends of polar molecules attract the electron-poor ends of their neighbours. This attraction, while much weaker than the force between charged ions, imparts important properties to polar water molecules. The dipole vectors in the water (H_2O) molecule produce an attraction called hydrogen bonding in liquid water and ice. These intermolecular hydrogen bonds are only 5 to 10 percent of the strength of the covalent O—H bonds in the water molecules themselves but account for most of water's unique properties, including its liquid state under mild terrestrial conditions, while the combination of hydrogen with the 20 elements most resembling oxygen are all gases under the same conditions, despite the much greater mass of most of them. In freezing, water expands into a structure that permits the most effective hydrogen bonding of its molecules, and as a consequence ice is less dense than liquid water and floats on its surface. Water is one of the very few materials known in which the solid form does not sink in the liquid. The electron-rich end of the water molecule is drawn even more strongly to positively charged ions than is the opposite end to negative ions. Water is thus an excellent solvent for compounds, such as salts, which give ions, as well as for other polar molecules such as alcohol or sugar. Nonpolar molecules, such as those found in oils, do not mix with water. In general, "like dissolves like" as far as molecular polarity is concerned, but oil and water can be made to mix in the presence of soaps or detergents, the molecules of which are constructed with highly polar, water-soluble groups to which are attached long oil-soluble hydrocarbon chains. Washing action combines the two unlike types of materials into an emulsified mixture with the soap molecules.

Soap structure

Resonance with radiation. Each of the different types of molecular flexibility considered under time-dependent phenomena has a characteristic rhythm of oscillation or frequency. That molecules can oscillate only with certain allowed frequencies is the result of quantum-mechanical laws that govern the behaviour of minute particles. Similarly, molecules can absorb energy only

in certain characteristic amounts. The absorption can stimulate more rapid rotations or vibrations or excite bonding electrons to higher energies. Radiation is described by frequencies or energies characteristic of its position in the electromagnetic spectrum (listing of the wavelengths and frequencies of electromagnetic radiation).

Molecular spectroscopy is based upon the absorption of electromagnetic radiation by molecules. Just as dishes in a cupboard can be made to rattle when certain notes are sung, particular types of motions in certain regions of a molecule will resonate with radiation of selected frequencies or energies. In order of decreasing energy, electronic excitation is generated by ultraviolet or visible light, molecular vibrations by infrared radiation, and rotations by far-infrared and microwave radiation.

Chemical properties. In general, the overall chemical behaviour of molecules is determined by the nature of the chemical bonds and the atoms they join. Subtle changes can arise in more complex systems, however, from the way certain portions of the structure are positioned. Reaction rates and whether a given reaction will take place at all are often governed by these spatial, or steric, factors. Chemical reactions in living organisms are almost always controlled by detailed molecular structure. This is evident in the requirement for particular isomers, such as for left-handed amino acids in food, versus the right-handed structure, which cannot be digested. Chemical processes in biology are apparently governed by the ability of sometimes highly complex molecular aggregates to fit together in space. The double helix structure of deoxyribonucleic acid, DNA, is a famous example: the cylindrical form brings complementary portions of the two individual strands together in a hydrogen-bonding interaction between matched base groups. For the interwound strands to be held together by hydrogen bonding, the sequence of the bases on the two strands must be complementary; this sequence is the information used in constructing new strands. The order on one of the strands is used alone in the production of a slightly different polymer, similar to DNA, called ribonucleic acid, RNA, and RNA serves as the pattern upon which amino acids are brought together and incorporated into new linear protein polymers. Viruses carry RNA units, which invade cells and cause destructive changes in their functioning. Changes in the order of the units in DNA, which can be brought about by radiation with enough energy to disrupt chemical bonds or by other molecular agents (mutagens), produce mutations. Biochemists have succeeded in synthesizing a number of biologically active materials in the laboratory, including an exact copy of the DNA in a virus, in 1967, and a yeast gene (an element that transmits hereditary characters), in 1970.

Structure of DNA

Through investigations into the genetic materials of creatures lower on the evolutionary ladder, the molecular background of human life is revealed. Work continues on the possibility of correcting hereditary defects, by supplying missing genes, and even of designing DNA and RNA molecules that could lead to forms of life that are not now known.

STRUCTURE DETERMINATION

Information about molecular structure, once limited to inferences derived from gross physical properties and chemical reactions of compounds, has been greatly expanded through the techniques of spectroscopy and diffraction. Some of these techniques are now used routinely by the practicing chemist and can give, within minutes, data from which the nature of molecular structure can be inferred. Other techniques, especially diffraction, are capable of determining the dimensions of molecules, but long periods of time are required to collect the data and laborious computations to interpret it. Such methods are used by a relatively few expert chemists.

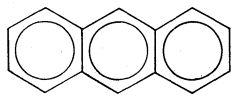
Precise methods. *Microwave.* Rotational characteristics can, in principle, provide information about molecular dimensions, since the frequency of a body's rotation is governed by its moment of inertia (its tendency to continue in motion), which in turn depends upon the masses and the distances between them. With knowledge of the precise masses of atoms, precise distances can be calculated from rotational data. In practice, the microwave technique (a microwave is electromagnetic radiation of between one and 100 centimetres in wavelength) has been limited to relatively simple polar molecules in the gas phase. It has shown, for example, that the two hydrogen atoms in the water molecule are about one angstrom (10^{-10} metre) from the oxygen atom and that the angle formed by the three atoms is 104.5°.

Electron diffraction. Electron beams passing through a sample of molecules are diffracted (scattered). Such radiation cannot penetrate deeply into matter, but the pattern of diffraction of electrons by gaseous samples can be used to obtain molecular dimensions. Molecules of a gas are relatively far apart and so no intermolecular diffraction effects occur; instead, it is the relative dispositions of the atoms in space within each molecule that determine the diffraction pattern. In using this method a molecular structure is assumed, the intensities of the diffracted beams that it would cause are calculated, and the observed values are compared with those predicted to determine the validity of the assumed model. In practice, the technique is used for fairly simple, somewhat symmetrical, gaseous molecules, but different structural arrangements cannot always be distinguished. Electron-diffraction patterns of small molecular fragments boiled off the surfaces of refractory salt crystals, such as sodium chloride (NaCl), have been studied at high temperatures.

X-ray and neutron-beam diffraction. X-rays are scattered far less than electrons and can penetrate deep into crystalline solids. Each atom of the regular arrangement within a crystal can diffract X-rays, and the manner of diffraction within molecular crystals depends upon the shape and size of the molecules and on their positions and orientation in the crystal. The X-ray diffraction patterns are compared with prior calculations based on plausible assumptions, in a trial and error process ideally suited to a computer. This technique has been developed into the most effective method available for the determination of detailed molecular structure. Because hydrogen atoms have far less scattering power than heavier atoms, the positions of hydrogen atoms in complex molecules usually cannot be obtained from X-ray data. Neutrons, on the other hand, are scattered by all atoms to a similar extent, and therefore the positions of hydrogen atoms can be readily established, even when very much larger atoms are present. To obtain neutron beams of the necessary intensity, however, requires an atomic pile, and neutron-diffraction studies can be carried out only in few laboratories. X-ray diffraction requires only that the molecules be in a crystalline, not glassy, solid. Liquids and gases can be studied as crystals at low temperatures.

Limitations of X-ray diffraction

Many special techniques have been devised to reduce the labour of calculating structures from diffraction data. Structures of simple molecules, which once required months of computation, can be solved with great accuracy in a matter of days, using computers. An example is the electron-density-distribution map of the molecule anthracene, $C_{14}H_{10}$, from X-ray data, which verified the chemist's intuitive concept represented in its structural formula



(see Figure 1). The structure of DNA, proposed in 1953 from X-ray data, was a breakthrough toward understanding the fundamental chemistry of life (see Figure 2).

Inferential methods. In recent years there has been rapid growth in experimental methods based on some physically measurable property that varies according to molecular structure. This category includes spectroscopic techniques, in which a beam of radiant energy is dispersed by a sample and its component waves then arranged in order of their wavelengths; these techniques are commonly used to give information that can be re-

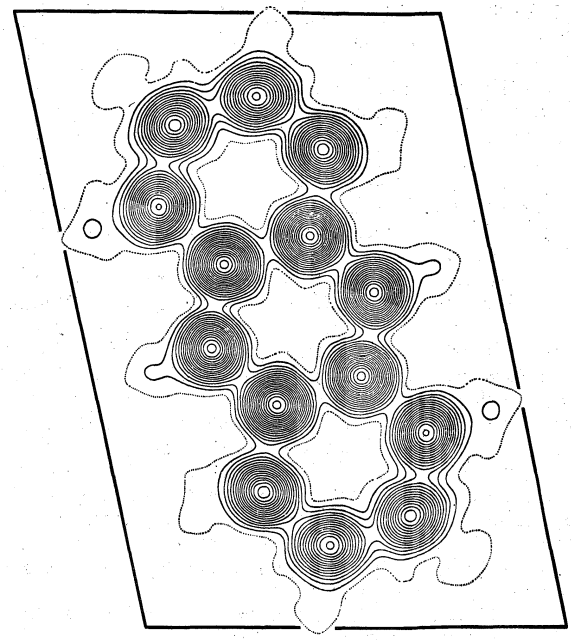


Figure 1: Electron-density map of anthracene, $C_{14}H_{10}$, drawn from X-ray-diffraction data. Contour lines are drawn at density intervals of one-half of an electron per angstrom (see text).
From E. Braude and F. Nachod (eds.), *Determination of Organic Structures by Physical Methods (1955)*; Academic Press

lated to molecular structure. In most cases it is not possible to make theoretical predictions about the relationship between molecular structures and the measured property from established principles. The practicing chemist, therefore, relies upon assumed similarities to measurements on compounds of known structure. The information thus obtained for the solution of new problems is inevitably circumstantial; a choice must often be made between a number of apparently reasonable interpretations. In this respect, especially, additional data from other physical methods can be of great value.

From J. Robertson, *X-ray Diffraction*, vol. 1, (1955); Academic Press

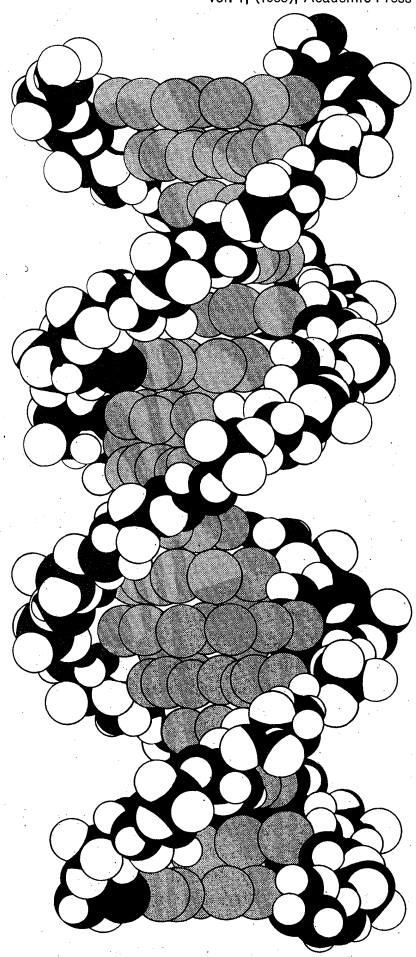


Figure 2: DNA double helix molecule.

Ultraviolet and visible (electronic) spectra. Stable compounds of low molecular weight are electronically excited only by radiation in the higher energy ultraviolet region of the spectrum; and water, carbon dioxide, ethyl alcohol, and other compounds of low weight are thus colourless. Colour in a compound indicates the presence of easily excited electrons, and certain functional groups or types of structure (chromophores) provide colour. In series of compounds with chains of alternating double and single bonds (conjugated), such as 1,3-butadiene, $H_2C{=}CH{-}CH{=}CH_2$, and 1,3,5-hexatriene, $CH_2{=}CH{-}CH{=}CH{-}CH{=}CH_2$, the absorption of radiation, corresponding to electronic excitation, progressively falls in required radiant energy through the ultraviolet into the visible region, with the compound with eight double bonds showing colour. Double bonds isolated in a molecular structure can thus be distinguished from those that are conjugated. Samples are usually studied in dilute solution.

Significance of colour in compounds

Infrared (vibrational) spectra. Absorption of infrared radiation stimulates the vibration of bonded atoms and groups to more violent and longer excursions from their equilibrium positions. By analogy with a pair of weights connected by a spring, strong bonds require more energy to stimulate stretching and bending, while increasing the masses of the bonded atoms lowers the energy required. Because energies and distances between nuclei are relatively constant values for pairs of bonded atoms, despite the great variety of molecular structures into which they are incorporated, it has been possible to assign characteristic infrared group-absorption frequencies to some pairs of atoms. Tabulations of such frequencies are exceedingly useful in the rapid identification of these groups from their infrared spectra, and shifts of frequencies are often used to deduce the nature of adjacent groups in the molecule. The complete infrared spectrum of a complex molecule can be considered to a first approximation as the superposition of the absorptions of the various pairs of atoms and functional groups in the molecule. Such a spectrum will be characteristic of the particular molecular structure; an unknown sample can be identified by comparing its infrared spectrum with that of a known compound. Small changes can be easily detected. Infrared spectroscopy is in general use in all chemical laboratories. Samples can be studied in any phase.

Mass spectrometry. Traditional methods of deducing molecular structure by disrupting it chemically, identifying the fragments, and mentally constructing them into a plausible structure for the original molecule find a counterpart in the study of compounds by fragmenting them under electron bombardment in the mass spectrometer. The masses of the (usually positively) charged fragments so produced are analyzed by electric- and magnetic-field separation, and extremely accurate measurements, which allow the type and number of atoms in each fragment to be specified, can be obtained. The original structure can be reconstructed by trial and error, on a computer. Minute samples can be studied by this technique.

Nuclear magnetic resonance. Except for mass and electric charge, nuclear properties were once of little interest to the chemist. Today it is known that interactions of nuclei with their chemical environment give rise to a broad range of phenomena, and molecular structures are studied with many techniques based upon energy changes in the nucleus. Nuclear-magnetic-resonance spectroscopy, which measures the energy necessary to reorient the magnetic dipole moment of the nucleus in an applied magnetic field, is the most widely used of such methods. The resonance energy required depends upon the particular environment of the atom, typically hydrogen,

in the molecular structure. Thus, the three types of hydrogen atoms in ethyl alcohol, CH_3CH_2OH, can be distinguished, and the signal response is in the ratio 3:2:1, which tells the relative numbers of the structurally distinct hydrogens. Neighbouring hydrogen atoms affect the signal in a way that allows further inferences to be drawn about structure in the surrounding region of the molecule. Nuclear magnetic resonance spectra are measured in solution.

The prodigious progress in molecular-structure analysis since World War II has resulted in large part from the widespread use of combinations of physical techniques; each yields some information about molecular structure and is reinforced by complete and precise determinations of key examples. The foregoing techniques for the identification of the components of molecules can be adapted to micro samples; a complex molecular structure can be solved in a matter of hours by information collected by a battery of methods from a sample barely large enough to see.

In the middle of the 19th century, the gifted British physicist James Clerk Maxwell wrote:

> A molecule is the smallest possible portion of a particular substance. No one has ever seen or handled a single molecule. Molecular science, therefore, is one of those branches of study which deal with things invisible and imperceptible by our senses, and which cannot be subjected to direct experiment.

The basic concepts have not changed. What has changed is that increasingly indirect physical effects have been found to be understandable in terms of molecular structure. The insights obtained from the application of these experimental techniques to problems of molecular structure have provided the major stimulus to mid-20th-century chemistry as a science and are beginning to have similar effect in the world of biology.

BIBLIOGRAPHY

Introductory works: O.T. BENFEY, *From Vital Force to Structural Formulas* (1964); F.C. NACHOD and R.F. BRILL, "Molecular Structure," in the *Kirk-Othmer Encyclopedia of Chemical Technology*, vol. 9 (1952), the status of knowledge on this subject in the early 1950s; J.D. WATSON, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA* (1968), an informative and readable account describing the work of the Nobel prize winners, Watson and Crick.

Textbooks (advanced university level): W.S. BREY, JR., *Physical Methods for Determining Molecular Geometry* (1965); E. CARTMELL and G.W.A. FOWLES, *Valency and Molecular Structure*, 3rd ed. (1966); C.A. COULSON, *Valence*, 2nd ed. (1961); G. KARAGOUNIS, *Einführung in die Elektronentheorie organischer Verbindungen* (1959; Eng. trans., *Introductory Organic Quantum Chemistry*, 1962); F.C. NACHOD *et al.* (eds.), *Determination of Organic Structures by Physical Methods*, 4 vol. (1955–71; vol. 5 in prep.); LINUS PAULING, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3rd ed. (1960); and with ROGER HAYWARD, *The Architecture of Molecules* (1964); R.M. SILVERSTEIN and G.C. BASSLER, *Spectrometric Identification of Organic Compounds*, 2nd ed. (1967).

(F.C.N./J.J.Z.)

Molecular Weight

A molecule is a group of atoms bonded together strongly enough to establish a separate identity and act together as a unit and, consequently, to have a constant mass and atomic composition. Measurements of molecular weight were first made on simple gases of known atomic composition, which at very low densities could be considered to be separable—*i.e.*, discrete—molecules.

Historic development of atomic-molecular theory

Although the idea of molecular weight is accepted today, it evolved only slowly. As modern atomic-molecular theory was developed and the existence of discrete molecules asserted, the Italian physicist Amedeo Avogadro postulated (1811) that equal volumes of different gases would contain the same number of molecules if the temperature and pressure were the same. Therefore, if equal volumes of two gases were weighed under identical conditions of temperature and pressure, the weights would compare exactly as the masses of single molecules. Stanislao Cannizzarro, an Italian chemist, revived Avogadro's concept in 1858 and showed its significance by analyzing existing measurements of gas densities. He was able to rationalize the molecular weights and chemical formulas of the well-known gases, and molecular weights came to be accepted as one of the fundamental properties of any chemical compound.

The number of molecules in one millilitre (ml) of gas is about 10^{19} at room temperature and atmospheric pressure. Although gas densities vary, a value of 0.001 gram per millilitre is common. The mass of a single gas molecule would then be about 10^{-22} gram. Because this mass is not directly measurable, other ways of determining molecular weight are required, and a scale of relative values has been built around the concept of a mole (*i.e.*, the amount of a substance in grams that corresponds to the sum of the atomic weights of all the atoms that make up the molecule; the atomic weight of the atom of an element is defined as its mass as compared with the mass of the atom of carbon-12). A mole of a substance is considered to contain exactly the same number of molecules as there are atoms in 12 grams of the pure carbon isotope carbon-12: $6.02252 \pm 0.00028 \times 10^{23}$ atoms of carbon. This number is called Avogadro's number or, in some literature, Loschmidt's number.

The molecular weight of a substance is defined as the weight in grams of one mole of the substance; that is, the actual weight in grams of 6.02×10^{23} molecules. Molecular weights are expressed in grams per mole (or daltons, a term common in the field of biological chemistry, although it has not been approved by the International Union of Pure and Applied Chemistry). Molecular weights range into the millions, as shown by the Table of representative molecules.

Approximate Molecular Weights

molecule	structure	molecular weight (in grams/mole)
Hydrogen	H_2	2
Sucrose (sugar)	$C_{12}H_{22}O_{11}$	342
Insulin	51 amino-acid residues	5,700
Polyesters	$(O(CH_2)_5CO)_n$	10,000–20,000
Hemoglobin	587 amino-acid residues	65,000
Synthetic rubbers	$(CH_2—CH=CH—CH_2)_n$	100,000
Myosin		600,000
Polystyrene	$(C_6H_5CHCH_2)_n$	50,000,000
Nucleic acids	phage DNA	130,000,000 and greater
Dextrans	polysaccharides	up to 500,000,000

All molcules over this range of molecular weights can be at least conceptually separated from one another for some measurement of molecular weight. For many solids with molecules that cannot be separated into a gaseous state, it is possible to find a solvent that will dissolve the substance and thus separate the molecules so that their weights can be determined.

This simple procedure is subject to two frequent chemical complications: (1) The solution of the dissolved molecules often results in chemical reactions that produce associated or dissociated species rather than a simple, unchanging molecular species. (2) Solids such as sodium chloride (NaCl) have a neat geometrical arrangement of sodium ions and chloride ions in the solid state, but they have a disordered arrangement of ions and no clearly defined molecular species of NaCl in the solution state (in which the ions dissociate and move about in accordance with electrochemical factors). In both these cases the simplest chemical formula that represents the proper chemical composition is assumed, and the molecular weight is assigned to that formula. If alcohol, for example, has a structure for which the simplest formula is C_2H_5OH, its molecular weight is found by multiplying the number of atoms in this formula by their atomic weights as found in the atomic-weight tables: that is, 2 atoms of carbon $\times$ (12.011) + 1 atom of oxygen $\times$ (15.999) + 6 hydrogen atoms $\times$ (1.008) = 46.069. This

weight of alcohol—*i.e.*, 46.069 grams—then constitutes one gram mole. Often when dealing with known simple materials such a calculation is sufficient. For unknown substances or known substances in a totally new environment, one of the methods discussed below is used.

PRINCIPLES OF MOLECULAR-WEIGHT DETERMINATION

The three methods of determining the molecular weight of uncharged molecules are based respectively on thermodynamic theory, on kinetic theory of transport phenomena, and on known spatial arrangements of atoms in the solid state. A fourth method, mass spectrometry, can be used on species that can be ionized; *i.e.*, given an electrical charge either by adding or by removing one or more electrons from the structure.

Thermodynamic methods

The thermodynamic methods of molecular-weight determination rely on the equation of state for gases or the equation of state for solutions. The laws concerning the physical behaviour of gases are among the most useful in science, but they are based upon the concept of an ideal gas in which each molecule is, in effect, a point without volume, without any attraction or repulsion toward other points, and with perfect resilience in collisions. The molecules of a real gas have volume and are attracted to or repelled by one another; they often have electrical charges and magnetic properties, and they have a true mass that makes them react to the gravitational field. They may react chemically under certain conditions, and they spin and tumble about. Their internal structure is such that they can absorb energy or lose it in ways that may or may not affect their motion. In an ideal gas the volume (V), pressure (P), temperature (T), and a constant of proportionality (R) are related in such a way that the volume is directly proportional to the temperature and to the constant and is inversely proportional to the pressure, thus: $V = RT/P$. This relationship is true with R only if one mole of gas is considered. If the gas weighs W_2 and its molecular weight is M, then the number of moles is W_2/M, and the equation may be rewritten pressure times volume is equal to constant times temperature times weight divided by mass, or: $PV = RT(W_2/M)$. Molecules in a solution may be considered to behave somewhat like gases, and an equation of state can be derived for solutions following the same sort of reasoning as with gases, with the additional factors that dominate the liquid state. If π represents osmotic pressure (defined below) as analogous to gas pressure, and C_2 is the concentration of dissolved molecules, analogous to the weight of gas molecules, then the van't Hoff relation for the equation of state for solutions is written $\pi = RT(C_2/M)$. Although ideal gases and ideal solutions are never encountered, the ideal is approached with increasing dilution. Furthermore, if measured quantities are extrapolated to infinite dilution, the use of ideal-gas or ideal-solution laws is usually justified.

Transport and other methods

Transport phenomena include the movement of a variety of properties through a system; *e.g.*, heat. The transport methods of molecular-weight determination depend on the relationship of the size of the molecule to the diffusion, or spreading of concentration, the viscosity (*i.e.*, a liquid's internal resistance to flow), or sedimentation (*i.e.*, the settling of finely divided particles from a liquid) of the molecules, whether in the gaseous or dissolved states. Other transport methods rely on the heat conduction or electrical mobility. If a simple shape can be assumed for the molecule, there is a direct relationship of size, mass, and density. Historically these methods were used for determining Avogadro's number, but today they are seldom used for molecular-weight determinations of small molecules except under special experimental conditions.

The geometrical arrangements of the molecules and their constituent atoms can sometimes be clearly determined in the crystalline state (*e.g.*, when X-rays are passed through a thin slice of a crystal to fall on a photographic plate, the two-dimensional pattern on the plate can be interpreted in terms of the three-dimensional structure of atomic nuclei that deflected the X-rays); in such instances, it is easy to calculate the average volume occupied by the molecule. By multiplying the average molecular volume by the macroscopic crystal density, the average mass of a molecule can be determined. The molecular weight can then be determined quite accurately using Avogadro's number. This method is obviously limited to a few well-studied compounds and is not used for routine molecular-weight determinations.

If the molecule can be ionized (*i.e.*, given an electrical charge by either adding or removing electrons), the ionized sample can be fed into a mass spectrometer, in which the heavier molecules are separated from the lighter. The ratio of mass to charge can be accurately determined (see MASS SPECTROMETRY). Since the ions are frequently of low charge, it is possible to determine ion mass. Using Avogadro's number, the molecular weight can be determined. This method is extremely accurate but is restricted to low molecular weights. Because of its speed, it is used extensively in the petroleum industry.

Application of methods

For an unknown compound the thermodynamic and transport methods discussed above are usually used, unless the molecule is of low molecular weight or the substance is gaseous. In the latter cases the mass spectrograph may be preferred. The thermodynamic and transport types of measurement can be simply and inexpensively performed. From an operational point of view, molecular-weight measurements by the thermodynamic and transport methods generally require that: (1) the substance be used in a solution, (2) the solution be sufficiently dilute that van't Hoff's law is obeyed, and (3) the substance be a pure chemical compound, not a mixture of different molecules.

Frequently, the substance is not a pure chemical compound; petroleum samples, for example, are often mixtures of several hydrocarbons, the boiling points of which are close together. Furthermore, polymer samples, such as plastics, rubbers, or wood, are always mixtures of differing molecular weights (see below *Molecular-weight distributions*). Even pure proteins often occur with small traces of other similar components. The best representation of any mixture is a frequency count of all components, as is routinely done with petroleum samples, using the mass spectrometer. With samples of higher molecular weight, however, such as polymers and proteins, the same techniques are not applicable; average molecular weights must be given. In averaging, the different molecules are counted either by number or by weight. In the first case, the molecular weight is called a number-average molecular weight and, in the second, a weight-average molecular weight.

METHODS OF MOLECULAR-WEIGHT DETERMINATION

Gaseous-state methods. *Density.* The measurement of gas densities to determine the molecular weight of a gas is an application of the ideal-gas law, which can be used with any volatile liquid or solid. Sensitivity of the method is low because large vessels must be weighed to determine the mass and volume of the gas. Real molecules are not simple points in space as the ideal-gas law assumes, and, accordingly, a pressure-dependent term must often be added to the ideal-gas law. At one atmosphere pressure, the correction is less than 1 percent for very simple gases.

Effusion. The general theory of diffusion (the expansion of a gas as a result of its internal pressure, brought about by the incessant motion and collisions of its molecules) indicates that the rate of diffusion will be inversely proportional to the mass of the diffusing particle; *i.e.*, the more massive particle diffuses more slowly. Effusion of a gas through a small hole can be used to measure the ratios of molecular weight of any two gases. The rates of effusion R_1 and R_2, of two species, for example, are related to the square roots of their molecular weights (M_1 and M_2) inversely and therefore can be expressed numerically as $\frac{R_1}{R_2} = \sqrt{\frac{M_2}{M_1}}$, all other variables being constant. In the mass spectrometer, the rate at which molecules enter or leave the ion chamber (a process of effusion) is measured, a technique that allows analysis of mixtures by molecular weights.

Formula for effusion methods

Mass spectrometry. The mass spectrometer is so sensitive that it can distinguish between isotopes (atoms of the same element that differ in atomic mass) in molecules of low molecular weight. Consequently, in many reactions the changes in molecular weight of the molecules involved can be followed through isotope or heavy atom exchange. The amount of material required for analysis is small (less than 0.1 millilitre of gas). Mass spectrometers are tied directly into refinery operations to make automatic molecular-weight determinations of petroleum fractions. The limited range of the mass spectrometer is its only major disadvantage in this application. It is suitable only for a range of molecular weight (for a unit electrical charge) of up to several thousand grams per mole, and the mass spectrometer is generally used for molecular weights no greater than 500 grams per mole. Another disadvantage is that, after the sample is volatilized inside the instrument, only fragments of the original molecule remain to be measured.

Solution methods using colligative-property measurements. Solution methods are in general use because they can be inexpensive, and they cover a very large molecular-weight range, about 100 grams per mole to greater than 100,000,000 grams per mole. They also can vary from simple thermodynamic methods to transport methods that are based on thermal conduction and mass transport. Solution methods usually are separated into those that count numbers of molecules (colligative properties) and those that count weights of molecules.

Osmotic pressure. When molecules of the substance to be measured are placed in a solvent, they affect the energy of the solvent molecules so that the latter no longer have the same chemical potential (or internal pressure) as the pure solvent. If the pure solvent is separated from the solution by a membrane that is impermeable to all but solvent molecules, these will migrate through the membrane from a higher to a lower concentration of solvent; *i.e.*, from the pure solvent side to the solution side. Pressure applied to the solution will inhibit the migration, and the exact pressure that must be applied to bring about equilibrium between solution and solvent is called the osmotic pressure. This is a direct measure of the effect of the substance whose molecular weight is sought upon the chemical potential of the solvent, and experiments show that osmotic pressure (π), as was stated earlier, equals the product of the concentration of dissolved molecules (C_2), the temperature (T), and the gas constant (R) from the ideal-gas law, divided by the molecular weight (M) of the dissolved molecules. Rearranging this, the equation may be written: $M = RTC_2/\pi$. As with the gas equation of state, all of the solution measurements are conducted at low concentrations or are extrapolated to zero concentration, in order to minimize the discrepancy between an ideal state and the actuality, in which molecules affect one another in various other ways as well.

Automatic osmometers

In recent years automatic osmometers have become commercially available. They are not more accurate than simple osmometers, using capillary tubes, but they are faster. With corrections for diffusion through the membrane, osmotic-pressure measurements have accurately measured the molecular weight of sucrose ($M = 342$ grams per mole). Osmometers have been used to measure M as high as 1,000,000 grams per mole.

Vapour pressure. The vapour pressure of a pure solvent is always lowered when the substance to be measured is added to it, because the concentration of solvent molecules at the surface is reduced. The solvent molecules have fewer chances at the surface to exert a pressure. Such small pressure differences, however, are generally so difficult to measure that the technique is not used routinely for molecular-weight determinations.

Vapour-pressure osmometer. Differences in vapour pressure can be measured by thermistors, devices that employ semiconductors to measure minute changes in temperature. If a thermistor has a drop of solution and a drop of solvent, with only a small air gap separating the drops, the different rates at which vapour is transported between the drops cause a temperature difference that the thermistor can sense. This temperature difference can be related to the molecular weight of the solute. The technique needs to be calibrated with known compounds, but the measurement has become common in organic- and polymer-chemistry laboratories. The range is limited to a few thousand grams per mole in molecular weight.

Boiling point and freezing point. When a compound is dissolved in a solvent, the chemical potential of the solvent is reduced, and the freezing point of the resulting solution becomes lower than that of the pure solvent. The difference in temperature is inversely proportional to the molecular weight of the solute. Again, the phenomenon may be explained in terms of the concentration of solvent molecules. A solution also has a higher boiling point than the pure solvent, and the difference in temperature is, as with the freezing point, inversely proportional to the molecular weight (the greater the difference the smaller the molecular weight). The elevation of the boiling point, however, is less sensitive than the lowering of the freezing point because of the higher heat of vaporization. The method is used in teaching because it is inexpensive, using only a thermometer and a water bath. Its range, except in the most experienced hands, is less than 5,000 grams per mole.

Other solution methods. *Equilibrium sedimentation.* If a sample solution is centrifuged (spun in a container shaped like a test tube) at extremely high speed, the dissolved molecules may collect in the region at the bottom of the tube. The process of concentrating the solution near the bottom of the tube, however, sets off an osmotic driving force that tends to make the dissolved molecules diffuse back into now less concentrated solvent. If the experiment is carried out carefully, the two forces are balanced and a very simple determination of molecular weight is possible through the osmotic equation of state (as explained above). The range of molecular weights that can be determined is enormous because the buoyancy of the medium and the speed of the centrifuge rotation can both be altered. In many ways measurement of solutions with the centrifuge, using the product of mass and buoyancy, are analogous to measurement with the mass spectrometer of gas, using a ratio of mass and charge. The method is capable of great accuracy and is fast. Its primary application has been in biological chemistry.

Light scattering. Fluctuations of concentration in a solution were analyzed by Einstein and found to be proportional to the change of osmotic pressure with concentration. To make a concentration fluctuation requires working against the osmotic forces in the solution, *i.e.*, the tendency for the ratio of solute to solvent molecules to become uniform throughout the solution. If a light beam strikes a solution in which there are fluctuations of concentration, it will be scattered. The Dutch chemist Peter Debye during World War II showed how the intensity of scattered light could be measured and applied to the analysis of the molecular weights of synthetic rubber through the osmotic equation of state. There is no limit in range for this method and it can be very precise, but extreme care is necessary to get meaningful results. Most polymer and biological laboratories use this method.

Solid-state measurements. In the technique of X-ray crystallography, extremely thin sections of solid sample are bombarded with X-rays, which are deflected by the pattern of nuclei in the sample to emerge in a pattern that can be photographed. This technique is so complex that it is used only for exceedingly important materials that cannot be dissolved in any useful solvent. In principle, a large molecule that is completely homogeneous could be analyzed by viewing its entire length by high-resolution electron microscopy. In practice, estimates of molecular weight can be made for any dispersed particle thought to be a single molecule if the densities can be guessed from other experiments.

X-ray crystallography

Relative methods using size. *Sedimentation rate and diffusion constant.* The sedimentation velocity and free diffusion methods assume that certain laws are obeyed, and, from this assumption, the molecular weight can be determined. If both diffusion and sedimentation mea-

surements are made, the need for the assumption is eliminated. These methods are seldom used for measuring molecular weight.

Intrinsic viscosity. A molecule turning over as it is moved in the flow of a solution has friction properties; a viscometer is used to make intrinsic viscosity measurements from which the molecular weight can be calculated. Einstein solved the problem very generally for spherical particles. The method is also used extensively for determining the molecular weight of polymers.

Gel permeation chromatography (GPC). The most recent development in routine methods of molecular-weight determination is gel permeation chromatography. Gels are semisolids consisting of two components—a liquid dispersed through a network of solid material. In chromatography, a solution is passed through a medium that has a stronger affinity for one of the dissolved compounds, which settles out first and thus is separated. Gels that have very well defined holes have been made of polystyrene and dextrans; when a solution passes through a column of the gel, the smaller molecules can diffuse into the holes and therefore are retained longer in the gel column. The larger molecules, not able to get into the holes, move through the column faster. Using calibration with known materials, the technique has worked effectively throughout a molecular-weight range of from a few hundred to many million grams per mole. This technique is fast becoming a standard method in the polymer industry because of its speed. Even porous glass with uniform holes has been used successfully.

MOLECULAR-WEIGHT DETERMINATIONS OF MACROMOLECULES

Very large molecules—that is, molecules with molecular weights greater than 20,000 grams per mole and perhaps up to 1,000,000,000 grams per mole—have become extremely important in industry and biochemistry. Plastics, rubbers, thickeners, paints, wood, paper, proteins, nucleic acids, polypeptides, and polysaccharides are of this type. As recently as the 1920s, these molecules were considered to be aggregates of smaller molecules rather than single large molecules, and their molecular-weight determination has been an extremely important development.

Gels. Gels exist as macromolecules, but their molecular weight is immeasurably large. A gel is not really a dissolved molecule but rather a separate continuous phase (a phase is a state of matter such as solid, liquid, gas). From a molecular point of view, a gel is analogous in three dimensions to the coupling of many short segments of threads in a two-dimensional mesh fabric. Such gels can be held together as a mesh by covalent bonds or hydrogen bonds. If the bonding forces are covalent, the gel cannot be handled by any of the above methods, and the molecular weight tends toward infinity, but the gel will have a particle size almost large enough to be observed by the unaided eye. Vulcanized rubber is typical of this type of gel. Gelatin is typical of a protein gel that can be dissolved in special solvents or at higher temperatures and become individual chains that are no longer in a gel structure. These single chains can then be analyzed for their molecular weight.

Effect of bonding in gels

Molecular-weight distributions. Although biological processes tend to produce relatively uniform macromolecules, industrial synthetic polymerization processes produce a rather broad range of groups of molecules. In terms of molecular weights, the proteins are said to be monodisperse in molecular weight, while the polymers are polydisperse.

Petroleum, a natural product, can be treated in a refinery to separate fractions that are more or less uniform, and thus the molecular weights of the components can be determined. A detailed accounting of all of the molecular weights in a truly polydisperse system, however, cannot be obtained. Instead, a statistical description, called the molecular-weight distribution (MWD), may be obtained.

In polymers, the existence of small amounts of other components of lower or higher molecular weight is important. Industrially, optimum processability and strength are related to the type of molecular-weight distribution. The presence of dimers or trimers (molecules consisting of two or three, respectively, identical subunits), or degraded material from a biological macromolecule, gives insight into many kinds of biochemical processes; consequently, the fast elucidation of the MWD has been a goal of macromolecular chemistry. In recent years this goal has been achieved by gel permeation chromatography and less routinely by a process called boundary spreading in the ultracentrifuge (extremely high speed of rotation and, consequently, better separation of particles). Previously, time-consuming separation by fractionation schemes was required. In both these faster techniques, however, the molecular-weight distribution may be reported incorrectly, because the analysis assumes that only linear molecules are present in the sample whereas the sample may also contain branched and complicated ring and other types of molecular structures.

Industrial techniques

Branched molecules. Both natural and synthetic polymers have nonlinear, branched molecules. For the problem of samples that are mixtures of linear and branched polymers, as well as for many other problems, the equilibrium-sedimentation method is useful because it offers the greatest versatility in determining a true molecular-weight distribution.

Insoluble polymers. Commercial polymers may have such high melting points that the operations of solution chemistry are difficult to carry out. One popular product has never had an accurate molecular-weight description because the solutions must be handled at 300° C (about 600° F). New special-purpose polymers are often high melting. To determine their molecular weights, experimenters have used correlations of molecular weight with viscoelastic properties and mechanical strength found in other, lower melting polymer systems. For example, cellulose is difficult to characterize because it must be dissolved in chemically active solvents, and such solvents may alter the cellulose. This is the case with many polymers.

End-group analyses. If the polymer has identical subunits, the ends of the molecule can often be analyzed chemically because they are different from all the other chemical units in the chain and will react with reagents. The molecular weights of food additives, foaming agents, and other intermediate-molecular-weight polymers are determined by chemical analysis of the number of ends. If the ends can be counted, the number of chains is known, and the average mass per chain, or molecular weight, can easily be determined.

Nonideal solutions. All macromolecules, as a result of their size, are affected in many ways and therefore have large additive terms, in addition to the van't Hoff relation, for molecular weight. To avoid such factors as much as possible, macromolecules must be measured in very dilute solutions and the resultant molecular weight extrapolated to infinite dilution. All early work in this field is now of limited value because this effect was not fully understood. Sometimes a poor solvent, which then ensures a more ideal solution, is used to minimize the extrapolation. Biological macromolecular systems and synthetic polyelectrolytes are always nonideal because of the charges on the molecules, and they must be analyzed for their molecular weight with special experimental care.

Biological macromolecular systems

Association of macromolecules occurs in biological macromolecules, such as insulin, and in synthetic polymers such as polyvinyl chloride. Like any other association equilibrium with similar small-molecular-weight systems, the problem can often be avoided by using a more active solvent or by studying the change of molecular weight with dilution or temperature or both.

BIBLIOGRAPHY. J.R. PARTINGTON, *An Advanced Treatise on Physical Chemistry*, vol. 1 (1949), presents original references to early work in molecular weights and basic equations. See also the same author's *A History of Chemistry*, vol. 4 (1964), for a detailed account of the historical development. The development of molecular weight as a concept is carefully woven into the fabric of the whole kinetic-molecular hypothesis in JAMES B. CONANT and LEONARD K. NASH

(eds.), *Harvard Case Histories in Experimental Science*, vol. 1 (1957). Other recommended works include: E.A. MOELWYN-HUGHES, *Physical Chemistry*, 2nd rev. ed. (1961), an excellent general reference book in physical chemistry; F.W. BILLMEYER, *Textbook of Polymer Science* (1971), a readable account of all aspects of polymer science; and C. TANFORD, *Physical Chemistry of Macromolecules* (1961), an excellent physical-chemical introduction to biomacromolecules.

(D.McI.)

Molière

Although the sacred and secular authorities of 17th-century France often combined against him, the comic genius of Molière emerged finally to win him eventual acclaim as the greatest of all French writers. Comedy had a long history before Molière, who employed most of its traditional forms, but he succeeded in inventing a new style that was based on a double vision of normal and abnormal seen in relation to each other—the comedy of the true opposed to the specious, the intelligent seen alongside the pedantic. An actor himself, Molière seems to have been incapable of visualizing any situation without animating and dramatizing it, often beyond the limits of probability; though living in an age of reason, his own good sense led him not to proselytize but rather to animate the absurd, as in such masterpieces as *Tartuffe*, *L'École des femmes*, *Le Misanthrope*, and many others. It is testimony to the freshness of his vision that the greatest comic artists working centuries later in other media, such as Charlie Chaplin, are still compared to Molière.

Molière, oil painting by the 17th-century artist Pierre Mignard. In the Musée Condé, Chantilly, France.

By courtesy of the Musee Conde, Chantilly, France; photograph, Giraudon

Beginnings in theatre. Molière was born (and died) in the heart of Paris. The registers showed that he was baptized on January 15, 1622, as Jean-Baptiste Poquelin. His mother died when he was ten; his father, one of the appointed furnishers of the royal household, gave him a good education at the Collège de Clermont (the school that, as the Lycée Louis-le-Grand, was to train so many brilliant Frenchmen, including Voltaire). Although his father clearly intended him to take over his royal appointment, the young man renounced it in 1643, apparently determined to break with tradition and seek a living on the stage. That year he joined with nine others to produce and play comedy as a company under the name of the Illustre Théâtre. His stage name, Molière, is first found in a document dated June 28, 1644. He was to give himself entirely to the theatre for 30 years and to die exhausted at the age of 51.

A talented actress, Madeleine Béjart, persuaded Molière to establish a theatre, but she could not keep the young company alive and solvent. In 1645 Molière was twice sent to prison for debts on the building and properties. The number of theatregoers in 17th-century Paris was small, and the city already had two established theatres, so that a continued existence must have seemed impossible to a young company. From the end of 1645, for no fewer than 13 years, the troupe sought a living touring the provinces. No history of these years is possible, though municipal registers and church records show the company emerging here and there: in Nantes in 1648, in Toulouse in 1649, and so on. They were in Lyons intermittently from the end of 1652 to the summer of 1655 and again in 1657, at Montpellier in 1654 and 1655, and at Béziers in 1656. Clearly they had their ups as well as downs. These unchronicled years must have been of crucial importance to Molière's career, forming as they did a rigorous apprenticeship to his later work as actor-manager and teaching him how to deal with authors, colleagues, audiences, and authorities. His rapid success and persistence against opposition when he finally got back to Paris is inexplicable without these years of training. His first two known plays date from this time: *L'Étourdi ou les contretemps* (*The Blunderer*, 1762), performed at Lyons in 1655, and *Le Dépit amoureux* (*The Amorous Quarrel*, 1762), performed at Béziers in 1656.

First success

The path to fame opened for him on the afternoon of October 24, 1658, when, in the guardroom of the Louvre and on an improvised stage, the company presented Corneille's *Nicomède* before the king, Louis XIV, and followed it with what Molière described as "one of those little entertainments which had won him some reputation with provincial audiences." This was *Le Docteur amoureux* ("The Amorous Doctor"); whether it was in the form still extant is doubtful. It apparently was a success and secured the favour of the King's brother Philippe, duc d'Orléans. It is difficult to know the extent of the Duc's patronage, which lasted seven years, until the King himself took over the company known as "Troupe du roi." No doubt the company gained a certain celebrity and prestige, invitations to great houses, and subsidies (usually unpaid) to actors, but not much more.

From the time of his return to Paris in 1658, all the reliable facts about Molière's life have to do with his activity as author, actor, and manager. Some French biographers have done their best to read his personal life into his works, but at the cost of misconstruing what might have happened as what did happen. The truth is that there is little information except legend and satire. The fact that authors like Montaigne, Plutarch, Julius Caesar, and Seneca may have been in his library (according to a legal inventory of 1708), for example, does not mean that his plays should be read with the doctrines of such authors in mind.

Although unquestionably a great writer, Molière was not an author in the usual sense: he wrote little that could be called literature or even that was meant to be published—some poems and a translation of the ancient Latin writings of Lucretius, incomplete. His plays were made for the stage, and his early prefaces complain that he had to publish to avoid exploitation. (Two of them were in fact pirated.) He left seven of his plays unpublished, never issued any collected edition, and never (so far as is known) read proofs or took care with his text. Comedies, in his view, were made to be acted. This fact was forgotten in the 19th century. It took such 20th-century actors as Louis Jouvet, Charles Dullin, Jean-Louis Barrault, and Jean Vilar to present a new and exact sense of his dramatic genius.

Nor was he at all a classical author, with leisure to plan and write as he would. Competition, the fight for existence, was the keynote of Molière's whole career. To keep his actors and his audiences was an unremitting struggle against other theatres. He won this contest almost single-handed. He held his company together by his technical competence and force of personality.

Molière's first Paris play, *Les Précieuses ridicules* (*The Affected Young Ladies*), prefigured what was to come. It centres on two provincial girls who are exposed by valets masquerading as masters in scenes that contrast, on the one hand, the girls' desire for elegance coupled with a lack of common sense and, on the other, the valets' plain

speech seasoned with cultural clichés. The girls' fatuities, which they consider the height of wit, suggest their warped view of culture in which material things are of no account. The fun at the expense of these affected people is still refreshing and must have been even more so for the first spectators.

Molière's Paris theatres

Les Précieuses, as well as *Sganarelle* (first performed in October, 1660), probably had its premiere at the Théâtre du Petit-Bourbon, a great house adjacent to the Louvre. The Petit-Bourbon was demolished (apparently without notice), and the company moved early in 1661 to a hall in the Palais-Royal, built as a theatre by Richelieu. Here it was that all Molière's "Paris" plays were staged, starting with *Dom Garcie de Navarre, ou le prince jaloux* in February 1661, a heroic comedy of which much was hoped; it failed on the stage and succeeded only in inspiring Molière to work on *Le Misanthrope*. Such failures were rare and eclipsed by successes greater than the Paris theatre had known.

Scandals and successes. The first night of *L'École des femmes* (*The School for Wives*), December 26, 1662, caused a scandal as if people suspected that here was an emergence of a comic genius that regarded nothing as sacrosanct. Some good judges have thought this to be Molière's masterpiece, as pure comedy as he ever attained. Based on Paul Scarron's version (*La Précaution inutile*, 1655) of a Spanish story, it presents a pedant, Arnolphe, who is so frightened of femininity that he decides to marry a girl entirely unacquainted with the ways of the world. The delicate portrayal in this girl of an awakening temperament, all the stronger for its absence of convention, is a marvel of comedy. Molière crowns his fantasy by showing his pedant falling in love with her, and his elephantine gropings toward lovers' talk are both his punishment and the audience's delight.

From 1662 onward the Palais-Royal theatre was shared by Italian actors, each company taking three playing days in each week. Molière also wrote plays that were privately commissioned and thus first performed elsewhere: *Les Fâcheux* (*The Impertinents*, 1732) at Vaux in August 1661; the first version of *Tartuffe* at Versailles in 1664; *Le Bourgeois Gentilhomme* at Chambord in 1670; and *Psyché* in the Tuileries Palace in 1671.

Marriage to Armande Béjart

On February 20, 1662, Molière married Armande Béjart. It is not certain whether she was Madeleine's sister, as the documents state, or her daughter, as some contemporaries suggest. There were three children of the marriage; only a daughter survived to maturity. It was not a happy marriage; flirtations of Armande are indicated in hostile pamphlets, but there is almost no reliable information.

Molière cleverly turned the outcry produced by *L'École des femmes* to the credit of the company by replying to his critics on the stage. *La Critique de L'École des femmes* in June 1663 and *L'Impromptu de Versailles* in October were both single-act discussion plays. In *La Critique* Molière allowed himself to express some principles of his new style of comedy, and in the other play he made theatre history by reproducing with astonishing realism the actual greenroom, or actors' lounge, of the company and the backchat involved in rehearsal.

The quarrel of *L'École des femmes* was itself outrun in violence and scandal by the presentation of the first version of *Tartuffe* in May 1664. The history of this great play sheds much light on the conditions in which Molière had to work and bears a quite remarkable testimony to his persistence and capacity to show fight. He had to wait five years and risk the livelihood of his actors before his reward, which proved to be the greatest success of his career. Most men would surely have given up the struggle: from the time of the first performance of what was probably the first three acts of the play as it is now known, many must have feared that the Roman Catholic Church would never allow its public performance.

Undeterred, Molière made matters worse by staging a version of *Dom Juan, ou le festin de Pierre* with a spectacular ending in which an atheist is committed to hell—but only after he had amused and scandalized the audience. *Dom Juan* was meant to be a quick money raiser, but it was a costly failure, mysteriously removed after 15 performances and never performed again or published by Molière. It is a priceless example of his art. The central character, Dom Juan, carries the aristocratic principle to its extreme by disclaiming all types of obligation, either to parents or doctors or tradesmen or God. Yet he assumes that others will fulfill their obligations to him. His servant, Sganarelle, is imagined as his opposite in every point, earthy, timorous, superstitious. These two form the perfect French counterpart to Don Quixote and Sancho.

Setbacks and harassment by the authorities

While engaged in his battles against the authorities, Molière continued to hold his company together single-handedly. He made up for lack of authors by writing more plays himself. He could never be sure either of actors or authors. In 1664 he put on the first play of Jean Racine, *La Thébaïde*, but the next year Racine transferred his second play, *Alexandre le Grand*, to a longer established theatre while Molière's actors were actually performing it. He was constantly harassed by the authorities. These setbacks may have been offset in part by the royal favour conferred upon Molière, but royal favour was capricious. Pensions were often promised and not paid. The court wanted more light plays than great works. The receipts of his theatre were uncertain and fluctuating. In his 14 years in Paris, Molière wrote 31 of the 95 plays that were presented on his stage. To meet the cumulative misfortunes of his own illness, the closing of the theatre for seven weeks upon the death of the Queen Mother, and the proscription of *Tartuffe* and *Dom Juan*, he wrote five plays in one season (1666–67). Of the five, only one, *Le Médecin malgré lui* (*The Doctor in Spite of Himself*, 1914), was a success.

In the preceding season, however, *Le Misanthrope*, almost from the start, was treated as a masterpiece by discerning playgoers, if not by the entire public. It is a drawing-room comedy, without known sources, constructed from the elements of Molière's own company. Molière himself played the role of Alceste, a fool of a new kind, with high principles and rigid standards, yet by nature a blind critic of everybody else. Alceste is in love with Célimène (played by Molière's wife, Armande), a superb comic creation, equal to any and every occasion, the incarnate spirit of society. The structure of the play is as simple as it is poetic. Alceste storms moodily through the play, finding no "honest" men to agree with him, always ready to see the mote in another's eye, blind to the beam in his own, as ignorant of his real nature as a Tartuffe.

The church nearly won its battle against Molière: it prevented public performance, both of *Tartuffe* for five years and of *Dom Juan* for the whole of Molière's life. A five-act version of *Tartuffe* was played in 1667, but once only: it was banned by the President of Police and by the Archbishop on pain of excommunication. Molière's reply was to lobby the King repeatedly, even in a military camp, and to publish a defense of his play called *Lettre sur la comédie de l'Imposteur*. He kept his company together through 1668 with *Amphitryon* (January 13), *George Dandin* (Versailles, July 18), and *L'Avare* (September 9). Sooner or later so original an author of comedy as Molière was bound to attempt a modern sketch of the ancient comic figure of the miser. The last of his three 1668 plays, *L'Avare*, is composed in prose that reads like verse; the stock situations are all recast, but the spirit is different from Molière's other works and not to everyone's taste. His miser is a living paradox, inhuman in his worship of money, all too human in his need of respect and affection. In breathtaking scenes his mania is made to suggest cruelty, pathological loneliness, even insanity. The play is too stark for those who expect laughter from comedy; Goethe started the dubious fashion of calling it tragic. Yet, as before, forces of mind and will are made to serve inhuman ends and are opposed by instinct and a very "human" nature. The basic comic suggestion is one of absurdity and incongruity rather than of gaiety.

His second play of 1668, *George Dandin*, often dismissed as a farce, may be one of Molière's greatest creations. It centres on a fool, who admits his folly while suggesting that wisdom would not help him because, if

things in fact go against us, it is pointless to be wise. As it happens he is in the right, but he can never prove it. The subject of the play is trivial, the suggestion is limitless; it sketches a new range of comedy altogether. In 1669, permission was somehow obtained, and the long run of *Tartuffe* at last began. More than 60 performances were given that year alone. The theme for this play, which brought him more trouble than any other, may have come to him when a local hypocrite seduced his landlady. Of the three versions of the play, only the last has survived; the first (presented in three acts played before the King in 1664) probably portrayed a pious crook so firmly established in a bourgeois household that the master promises him his daughter and disinherits his son. At the time it was common for lay directors of conscience to be placed in families to reprove and reform conduct. When this "holy" man is caught making love to his employer's wife, he recovers by masterly self-reproach and persuades the master not only to pardon him but also to urge him to see as much of his wife as possible. Molière must have seen even greater comic possibilities in this theme, for he made five acts out of it. The final version contains two seduction scenes and a shift of interest to the comic paradox in Tartuffe himself, posing as an inhuman ascetic while by nature he is an all-too-human lecher. It is difficult to think of a theme more likely to offend pious minds. Like Arnolphe in *L'École des femmes*, Tartuffe seems to have come to grief because he trusted in wit and forgot instinct.

The long run of *Tartuffe*

Last plays. The struggle over *Tartuffe* probably exhausted Molière to the point that he was unable to stave off repeated illness and supply new plays; he had, in fact, just four years more to live. Yet he produced in 1669 *Monsieur de Pourceaugnac* for the King at Chambord and in 1670 *Le Bourgeois Gentilhomme*.

Le Bourgeois Gentilhomme treated a contemporary theme—social climbing among the bourgeois, or upper middle class—but it is perhaps the least dated of all his comedies. The protagonist Jourdain, rather than being an unpleasant sycophant, is as delightful as he is fatuous, as genuine as he is naïve; his folly is embedded in a bountiful disposition, which he of course despises. This is comedy in Molière's happiest vein: the fatuity of the masculine master is offset by the common sense of wife and servant.

Continuing to write despite his illness, he produced *Psyché* and *Les Fourberies de Scapin* (*The Cheats of Scapin*, 1677) in 1671. *Les Femmes savantes* (*The Blue-Stockings*, 1927) followed in 1672; in rougher hands this subject would have been (as some have thought it) a satire on bluestockings, but Molière has imagined a sensible bourgeois who goes in fear of his masterful and learned wife. *Le Malade imaginaire* (first performed 1673; Eng. trans., *The Imaginary Invalid*), about a hypochondriac who fears death and doctors, was Molière's last play. It is a powerful play in its delineation of medical jargon and professionalism, in the fatuity of a would-be doctor with learning and no sense, in the normality of the young and sensible, the young lovers, as opposed to the superstition, greed, and charlatanry of other characters. During the fourth performance of the play, on February 17, 1673, Molière collapsed on stage and was carried back to his house in the rue de Richelieu to die. As he had not been given the sacraments or the opportunity of formally renouncing the actor's profession, he was buried without ceremony and after sunset on February 21.

Molière's collapse on stage and death

Molière as actor and as playwright. Molière's acting had been both his disappointment and his glory. He aspired to be a tragic actor, but contemporary taste was against him. His public seemed to favour a tragic style that was pompous, with ranting and roaring, strutting and chanting. Molière had the build, the elasticity, the india-rubber face, as it has been called, of the born comedian. Offstage he was neither a great talker nor particularly merry, but he would mime and copy speech to the life. He had the tireless energy of the actor. He was always ready to make a scene out of an incident, to put himself on a stage. He gave one of his characters his own cough and another his own moods, and he made a play out of actual rehearsals. The characters of his greatest plays are like the members of his company. It was quite appropriate that he should die while playing the part of the sick man that he really was.

The actor in him influenced his writing, since he wrote (at speed) what he could most naturally act. He gave himself choleric parts, servants' parts, a henpecked husband, a foolish bourgeois, and a superstitious old man who cursed "that fellow Molière." (The comparison with Charlie Chaplin recurs constantly.) Something more than animal energy and a talent for mime was at work in him, a quality that can only be called intensity of dramatic vision. Here again actors have helped to recover an aspect of his genius that the scholars had missed, his stage violence. To take his plays as arguments in favour of reason is to miss their vitality. His sense of reason leads him to animate the absurd. His characters are imagined as excitable and excited to the point of incoherence. He sacrifices plot to drama, vivacity, a sense of life. He is a classical writer, yet he is ready to defy all rules of writing.

To think of Molière as a cool apostle of reason, sharing the views of the more rational men of his plays, is a heresy that dies hard; but careful scrutiny of the milieu in which Molière had to work makes it impossible to believe. The comedies are not sermons; such doctrine as may be extracted from them is incidental and at the opposite pole from didacticism. Ideas are expressed to please a public, not to propagate the author's view. If asked what he thought of hypocrisy or atheism, he would have marvelled at the question and evaded it with the observation that the theatre is not the place for "views." There is no documentary evidence that Molière ever tried to convey his own opinions on marriage, on the church, on hell, or on class distinctions. Strictly speaking, his views of these things are unknown. All that is known is that he worked for and in the theatre and used his amazing power of dramatic suggestion to vivify any imagined scene. If he has left a sympathetic picture of an atheist, it was not to recommend free thought: his picture of the earthy serving man is no less vivid, no less sympathetic. Scholars who have tried to make his plays prove things or to convey lessons have made little sense of his work and have been blind to its inherent fantasy and imaginative power.

The heresy of Molière as a propagandist

Since the power of Molière's writing seems to lie in its creative vigour of language, the traditional divisions of his works into comedies of manners, comedies of character, and farce are not helpful: he does not appear to have set out in any instance to write a certain kind of play. He starts from an occasion in *Le Mariage forcé* (1664; *The Forced Marriage*, 1762) from doubts about marriage expressed by Rabelais's character Panurge, and in *Le Médecin malgré lui* he starts from a medieval fable, or *fabliau*, of a woodcutter who, to avoid a beating, pretends he is a doctor. On such skeleton themes Molière animates figures or arranges discussion in which one character exposes another or the roles are first expressed and then reversed. It is intellectual rhythm rather than what happens, the discussion more than the story, that conveys the charm, so that to recount the plot may be to omit the essential.

His unique sense of the comic. The attacks on Molière gave him the chance in his responses to state some aesthetic home truths. Thus, in *La Critique de L'École des femmes*, he states that tragedy might be heroic, but comedy must hold the mirror up to nature: "You haven't achieved anything in comedy unless your portraits can be seen to be living types . . . making decent people laugh is a strange business." And as for the rules that some were anxious to impose on writers: "I wonder if the golden rule is not to give pleasure and if a successful play is not on the right track."

The attacks on *L'École des femmes* were child's play in comparison with the storm raised by *Tartuffe* and *Dom Juan*. The attacks on them also drew from the poet a valuable statement of artistic principle. On *Dom Juan* he made no public reply since it was never officially condemned. The documents in defense of *Tartuffe* are two placets, or petitions, to the King, the preface to the first

Concept of comedy

edition of 1669 (all these published over Molière's own name), and the *Lettre sur la comedie de l'Imposteur* of 1667. The placets and preface are aesthetically disappointing, since Molière was forced to fight on ground chosen by his opponents and to admit that comedy must be didactic. (There is no other evidence that Molière thought this, so it is not unfair to assume that he used the argument only when forced.) The *Lettre* is much more important. It expresses in a few pregnant lines the aesthetic basis not only of *Tartuffe* but of Molière's new concept of comedy:

> The comic is the outward and visible form that nature's bounty has attached to everything unreasonable, so that we should see, and avoid, it. To know the comic we must know the rational, of which it denotes the absence and we must see wherein the rational consists. . .incongruity is the heart of the comic. . .it follows that all lying, disguise, cheating, dissimulation, all outward show different from the reality, all contradiction in fact between actions that proceed from a single source, all this is in essence comic.

Molière seems here to put his finger on what was new in his notion of what is comic: a comedy, only incidentally funny, that is based on a constant double vision of wise and foolish, right and wrong seen together, side by side. This is his invention and his glory.

A main feature of Molière's technique is a mixing of registers, or of contexts. Characters are made to play a part, then forget it, speak out of turn, overplay their role, so that those who watch this byplay constantly have the suggestion of mixed registers. The starting point of *Le Médecin malgré lui*, the idea of beating a man to make him pretend he is a doctor, is certainly not subtle, but Molière plays with the idea, makes his woodcutter enjoy his new experience, master the jargon, and then not know what to do with it. He utters inanities about Hippocrates, is overjoyed to find a patient ignorant of Latin, so that he need not bother about meaning. He looks for the heart on the wrong side and, undeterred by having his error recognized, sweeps aside the protest with the immortal: "We have changed all that." The miser robbed of his money is pathetic, but he does not arouse emotions because his language leads him to the absurd ". . . it's all over . . . I'm dying, I'm dead, I'm buried." He demands justice with such intemperance that his language exceeds all reason and he threatens to put the courts in the court. Molière's *Misanthrope* is even more suggestive in his confusion of justice as an ideal and as a social institution: "I have justice on my side and I lose my case!" What to him is a scandal of world order is to others just proof that he is wrongheaded. Such concision does Molière's dramatic speech achieve.

A French genius. When Voltaire described Molière as "the painter of France," he suggested the range of French attitudes found in the plays, and this may explain why the French have developed a proprietary interest in a writer whom they seem to regard in a special sense as their own. They stress aspects of his work that others tend to overlook. Three of these are noteworthy.

First, formality permeates all his works. He never gives realism—life as it is—alone, but always within a pattern and a form that fuse light and movement, music and dance and speech. Modern productions that omit the interludes in his plays stray far from the original effect. Characters are grouped, scenes and even speeches are arranged, comic repartee is rounded off in defiance of realism.

Second, the French stress the poetry where foreigners see psychology. They take the plays not as studies of social mania but as patterns of fantasy that take up ideas, only to drop them when a point has been made. *Le Misanthrope* is not considered as a case study or a French *Hamlet* but as a subtly arranged chorus of voices and attitudes that convey a critique of individualism. The play charms by its successive evocations of its central theme. The tendency to speak one's mind is seen to be many things: idealistic or backbiting or rude or spiteful or just fatuous. It is in this fantasy playing on the mystery of self-centredness in society that Molière is in the eyes of his own people unsurpassed.

A third quality admired in France is his intellectual penetration in distinguishing the parts of a man from the whole man. Montaigne, the 16th-century essayist who deeply influenced Molière, divided qualities that are acquired, such as learning or politeness or skills, from those that are natural, such as humanity or animality, what might be called "human nature" without other attributes. Molière delighted in opposing his characters in this way; often in his plays a social veneer peels off, revealing a real man. Many of his dialogues start with politeness and end in open insults.

Molière opposed wit to nature in many forms. His comedy embraces things within the mind and beyond it; reason and fact seldom meet. As the beaten servant in *Amphitryon* observes: "That conflicts with common sense. But it is so, for all that."

MAJOR WORKS

Les Précieuses ridicules (first performed 1659, published 1660; trans. by B.H. Clark, *The Affected Young Ladies*, 1915); *L'École des femmes* (1663; trans. by the Earl of Longford, *The School for Wives*, 1948; and by M. Malleson, 1954); *Le Tartuffe, ou l'imposteur* (first version 1664, present version 1669; trans. by M. Malleson, *The Imposter*, 1950); *Dom Juan, ou le festin de pierre* (1665; trans. by J. Ozell as *Don John; or, The Libertine*, 1665 and rev. and augmented by O. Mandell, 1963); *Le Misanthrope* (first performed 1666, 1667; adapted by W. Wycherly, *The Plain-Dealer*, 1677; trans. by M. Malleson, 1955); *L'Avare* (1669; trans. by H. Fielding as *The Miser*, 1733; by M. Malleson, 1950; and by K. Cartledge, 1962, with the same title); *Le Bourgeois gentilhomme* (1670; trans. by M. Malleson, *The Prodigious Snob*, 1952); *Les Femmes savantes* (1672; trans. by V. Beringer and M. Down, *The Blue-Stockings*, 1927); *Le Malade imaginaire* (1674; trans. as *The Imaginary Invalid* by B.H. Clark, 1925; by M. Malleson, 1959; and B. Briscoe, 1967; and as *The Hypochondriac* by H. Baker and J. Miller, 1961).

BIBLIOGRAPHY

Editions: Collected editions of Molière's works include those by C. VARLET DE LA GRANGE, 8 vol. (1682); by M.A. JOLY, 6 vol. (1734); by E. DESPOIS and P. MESNARD in the "Grands Écrivains de la France Series," 13 vol. (1873–1900); by RENE BRAY in the "Belles Lettres Series," 8 vol. (1935–52); by GUSTAVE MICHAUT, 11 vol. (1949); by ROBERT JOUANNY in the "Garnier Series," 2 vol. (1960); and by GEORGES COUTON in the "Pléiade Series," 2 vol. (1971). Among editions of particular plays those of *L'Avare* by CHARLES DULLIN (1946), of *Le Malade imaginaire* by PIERRE VALDE (1946), of *Tartuffe* by F. LEDOUX (1953), and of *Le Misanthrope* by GUSTAVE RUDLER (1947) deserve special mention.

Biography: Earlier literature is superseded by GUSTAVE MICHAUT, *La Jeunesse de Molière* (1922), *Les Debuts de Molière à Paris* (1923), and *Les Luttes de Molière* (1925). See also JOHN PALMER, *Molière: His Life and Works* (1930, reprinted 1968); GUSTAVE MICHAUT (ed.), *Molière raconté par ceux qui l'ont vu* (1932); GEORGES MONGREDIEN, *La Vie privée de Molière* (1950); and RENE BRAY, *Molière, homme de théâtre*, new ed. (1963). Official documents have been collected in MADELEINE JURGENS and ELIZABETH MAXFIELD-MILLER, *Cent Ans de recherches sur Molière, sur sa famille et sur les comédiens de sa troupe* (1963).

Theatrical history: H.C. LANCASTER, *A History of French Dramatic Literature in the Seventeenth Century*, pt. 3 (1936, reprinted 1966); ANTOINE ADAM, *Histoire de la littérature française au XVII^e siècle*, vol. 3 (1956); P. MELESE, *Le Théâtre et le public à Paris sous Louis XIV* (1934); THEODORE VAN VREE, *Les Pamphlets et libelles littéraires contre Molière* (1933); and B.E. and G.P. YOUNG (eds.), *Le Registre de La Grange*, 2 vol. (1947). On particular plays, see ANTOINE ADAM, "La Genèse des 'Précieuses ridicules,'" *Revue des sciences humaines* (1939); JACQUES ARNAVON, *Le Misanthrope de Molière* (1930) and *L'École des femmes de Molière* (1936); and RENE JASINKI, *Molière et le Misanthrope* (1951).

General criticism: P.F. SAINTONGE and R.W. CHRIST, *Fifty Years of Molière Studies: A Bibliography, 1892–1941* (1942); W.G. MOORE, *Molière: A New Criticism* (1949); JACQUES GUICHARNAUD, *Molière, une aventure théâtrale. Tartuffe, Dom Juan, Le Misanthrope* (1963).

(W.G.Mo.)

Mollusca

The mollusks are the most numerous, varied, and widely distributed of animal groups after the insects and verte-

brates. The phylum Mollusca includes an immense variety of forms, including the snails and slugs (class Gastropoda), clams and oysters (Bivalvia), and octopuses and squids (Cephalopoda). There are also four smaller classes: Aplacophora (the worm-like solenogasters), Monoplacophora (represented by only a single living genus), Amphineura (chitons), and Scaphopoda (tusk shells). Despite the variety of body plans in the phylum, mollusks have in common a soft body without an internal skeleton, a slippery, mucus-secreting skin, and in most, but not all, an external calcareous shell.

In this article the biology and classification of mollusks are considered in general terms. For more detailed information on the three major classes, see GASTROPODA; BIVALVIA; CEPHALOPODA.

General characteristics of mollusks

The largest mollusks are giant squids of 12-metre (40-foot) arm span; the smallest are snails a millimetre long. Large gastropods are numerous, and the clams *Tridacna* are the giants among the bivalves. Mollusks live on rocky, sandy, and muddy bottoms. They can burrow into soft or calcareous rock or into wood, or can become free-swimming (pelagic). Both snails and bivalves have spread in some numbers to freshwaters, and pulmonate gastropods —the most numerous of all in species—live chiefly upon land, particularly in forests, but also in dry areas. Some gastropods are internal parasites of other animals. Mollusks, in short, are worldwide in distribution and have penetrated to most habitats that will support animals. They go back to the earliest ages as fossils in the rocks.

Based on photographs by courtesy of (B) Dr. R. Catala, Aquarium de Noumea, and (C) Marineland of the Pacific; from (A,F) Gusta Iagersten, *Life in the Sea* (1964); Basic Books, Inc. and Tidens Forlag, (D) J.E. Morton, *Molluscs* (1957 & 1967); Hutchinson Publishing Group, (E) J.E. Morton and M.C. Miller, *The New Zealand Sea Shore*, (G,H) L.H. Hyman, *The Invertebrates*, vol. VI (1967); used with permission of McGraw-Hill Book Co.

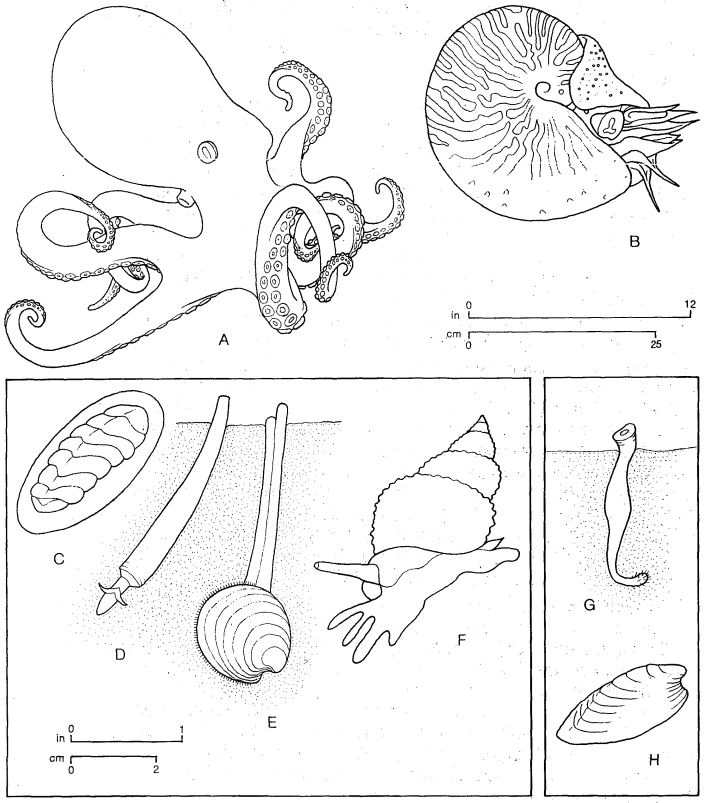


Figure 1: *Body plans of mollusks.*
(Above) Cephalopods: (A) Octopus; (B) Nautilus. (Below) (C) Amphineura: *Ischnochiton heathiana*. (D) Scaphopoda: *Dentalium*. (E) Bivalve: *Dorsina subrosea*. (F) Aplacophora: *Chaetoderma loveni*. (G) Gastropod: *Buccinum undatum*. (H) Monoplacophora.

Mankind has a strong interest in the mollusks for economic, as well as scientific, reasons. Many gastropods and even more of the bivalves serve as choice protein foods. The cephalopods (squid, cuttlefish, and octopus) are famous culinary items in many countries. Bivalve shells are used for many ornamental purposes, and a few species produce valuable pearls. Highly coloured or sculptured gastropod shells and the iridescent mother-of-pearl of bivalve shells, are used as ornaments, for buttons, or as primitive currency. Some freshwater snails are medically important as carriers of the intermediate larval stages of parasitic worms.

NATURAL HISTORY

Habitat and feeding habits. Most gastropods of the subclass Prosobranchia are marine, but a few live on land or in freshwater. All opisthobranchs are marine, many being residents of the littoral zone, a few on the bottom to depths of over 2,000 metres. Others, including both nudibranchs and shelled forms, are pelagic. The pulmonates include several families of freshwater snails, but the large majority are land snails, widely distributed through every sort of terrestrial habitat. Gastropods feed by a variety of methods: some scraping algae from the substrate with the tonguelike radula, others browsing on attached (sessile) or free-floating seaweeds, still others feeding on detritus or carrion. A number of gastropods are predatory, feeding on a variety of animals from sponges and coelenterates to other mollusks and live fish.

Bivalves inhabit marine and freshwater from the coastal (littoral) high tide line to about 6,000-metre depths. The largest number of species and individuals is found in shallower waters. Nearly all bivalves are filter feeders; that is, they extract microplankton from water pumped across the gills. Some bottom-dwelling bivalves apparently feed primarily on plant and animal macrodetritus, dead material that sinks from nearer the surface.

Among cephalopods, the octopuses are littoral in distribution, inhabiting rocky areas, where they feed on crustaceans, fishes, and bivalves. Squid are pelagic swimmers or bottom dwellers (often to great depths) also predatory on crustaceans and fishes, and even on other squid.

The smaller molluscan classes have been less studied than the three better known classes. The Aplacophora are found in bottom mud at depths of 18–4,000 metres or closely associated with other marine animals such as corals, anemones, and polychaete worms. It is inferred from the gut contents of preserved specimens that the burrowing species feed by filtering micro-organisms from the mud and those living on other animals feed on their hosts.

The chitons are most numerous on rocks of the intertidal zone, but one family is found commonly to depths of 4,000 metres. Most chitons feed by scraping algae and other organisms from the rock surfaces with the radula.

Little is known about the ecology of *Neopilina*, the only living monoplacophoran, all of which have been taken by bottom dredges from depths of 2,700–5,700 metres. The animals apparently feed on microscopic animals gleaned from the ooze of the ocean floor.

Tusk shells (Scaphopoda) are bottom dwellers on sand or mud to depths of 5,000 metres. They feed on small organisms, from diatoms to minute crustaceans and mollusks, captured by filamentous tentacles (captacula).

Locomotion. Few mollusks except the cephalopods are fast-moving, and some (such as boring bivalves) are virtually stationary as adults. Gastropods move by wave action of the broadly spread sole. The rear edge is moved forward first and an undulatory wave moves along the sole, carrying the animal forward. Most bivalves burrow by thrusting the large, bladelike foot into mud or sand to form an anchor, toward which the body is drawn by the foot retractor muscles. Many bivalves, including pearl oysters, mussels, and some scallops, do not burrow but remain attached to firm surfaces by byssus threads. The byssus is a bundle of tanned protein fibres secreted at the base of the foot. The nonsessile scallops (*Pecten*) and file shells (*Lima*) can perform spasmodic swimming motions by clapping the valves to expel water.

Most cephalopods alternate between walking and swimming. Slow movements are accomplished by drawing the body along on the arms. Swimming is accomplished, especially in the beautifully streamlined pelagic squids, by expelling a jet of water from the mantle cavity.

The chitons (Polyplacophora) cling to rocks or other hard substrates by a broad flat foot, on which they move for feeding. The closely related solenogasters (Aplaco-

phora) burrow in mud, thrusting the anterior end forward by alternate expansions and contractions. Other solenogasters move slowly among colonies of hydroids and corals in a manner similar to that of terrestrial snakes. The monoplacophorans have never been observed alive but apparently creep across soft bottom ooze on the broad foot, in a manner like that of chitons. The fourth small class of mollusks, the tusk shells, burrow into mud or sand using the foot as do the burrowing bivalves.

Reproduction. Reproduction in mollusks is primitively by separate sexes, eggs and sperm being released into the water and fertilized externally. This still prevails in Aplacophora, Polyplacophora, Monoplacophora, Scaphopoda, and primitive Gastropoda and Bivalvia. In the Gastropoda the first larva is called a trochophore, succeeded after torsion (see below) by a veliger with the strongly ciliated swimming lobes known as a velum. Bivalves have a larva with comparable wheel-like swimming lobes, often called a rotiger. Amphineura, Scaphopoda, and bivalves of the superfamily Nuculoidea have yolky larvae called vitellariae, of shorter free-swimming existence. Most prosobranch gastropods practice internal fertilization between separate sexes and produce capsulate spawn, from which swimming larvae may or may not be released. The opisthobranchs and pulmonates have complexly hermaphroditic reproductive systems, and fertilization is most often reciprocal. Two individuals, each with both male and female organs, join and exchange sperm.

The hallmark of all the Gastropoda is the developmental modification of torsion. During early life, in the free-swimming larva (when present), the visceral mass is rotated through 180° with respect to the head. The result is to bring the mantle cavity from its ventral, back-facing position to a dorsal forward-facing aspect. The right and left paired (mantle) pallial and pericardial organs are transposed. The visceral nerve loop is twisted to a figure eight. The illustrations of larvae before and after torsion (Figure 2) show the facility offered after torsion for withdrawing the vulnerable head and velum (the ciliated flap that provides locomotion) into the mantle cavity, followed by the foot bearing the horny covering (operculum).

From J.E. Morton, *Mulluscs*, 4th ed. (1967); Hutchinson Publishing Group Ltd.

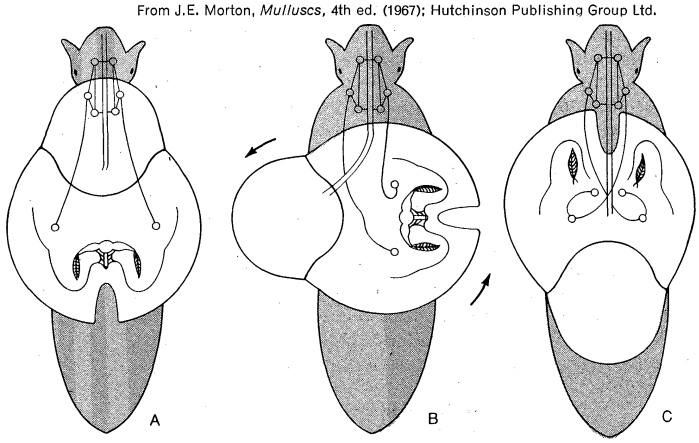


Figure 2: *Torsion.*
(A) Before torsion. (B) 90° of torsion. (C) Torsion completed (180°).

In addition to those with separate sexes, some bivalves exhibit various forms of simultaneous or successive hermaphroditism. A few, such as oysters, incubate the larvae between the gills of the mother. The Cephalopoda copulate by the male transferring spermatophore packets to the female by a modified "hectocotylus" arm; the eggs are allowed to float freely in the ocean or are attached to the bottom. In octopuses they are guarded by the mother and in the paper nautilus (*Argonauta*) are carried in a special "shell," secreted by an arm of the female.

FORM AND FUNCTION

Body plan. Mollusks are highly diverse as to body plans, and there is no standard molluscan shape or life form. Each of the seven classes has a different and rather characteristic pattern.

The majority of gastropods have the familiar snail shape. The mantle cavity and visceral mass are enclosed in a single helicoid shell piece, almost always coiled to the right. Marine gastropods generally close this by the operculum, a hard plate at the back of the foot. The shell spire is reduced to a shallow cone in many families, such as the limpets. In the more advanced gastropods the shell has become progressively reduced or lost. The whole subclass of opisthobranchs has evolved toward the naked slug form, culminating in the nudibranchs, which have unparalleled beauty of colour, often with graceful gill appendages.

Body plan of bivalves

The bivalves show an entirely different pattern, with a two-piece shell hinged dorsally by a ligament and closed or relaxed by adductor muscles. The whole body is withdrawn between the flaps of the mantle, which line the shell. The head is lost as such, and feeding is by ciliary means, with the large gills responsible for drawing in a water current and filtering food particles from it. The foot is large and laterally compressed. In burrowing bivalves, such as cockles, wedge-shells, venus shells, and clams, the edges of the mantle are fused, leaving only a gape for the foot and posterior openings that are drawn out into inhalant and exhalant siphons. Many surface-dwelling bivalves have brightly coloured shells with a lavish development of nacre or mother-of-pearl over the inner surface. The oyster family have lost the byssus and the foot and are cemented by one shell valve to the substratum. The cephalopods are the largest, most mobile and most metabolically active of the mollusks. The long lineages of shelled fossil forms are represented by only one surviving genus, *Nautilus*, in many ways more primitive than any other living cephalopods. In the other modern forms the body is naked, with a chambered internal shell plate in cuttlefish or a horny bar or rod (pen) in squids. These decapod (ten-armed) forms have a circlet of eight short arms, equipped with suckers, and two long, retractile tentacles with suckers at the expanded tips. A spiral, chambered internal shell is found in *Spirula*. The squids are streamlined for fast swimming. Giant squids of the genus *Architeuthis* may reach 12 metres (40 feet) in length. The octopods have no internal skeleton, and the body is soft and spherical, with eight tentacular arms of equal length.

Of the four small classes of mollusks, three, the Aplacophora, Polyplacophora and Monoplacophora, are undoubtedly primitive. All are marine. In the chitons (Polyplacophora) the shell is not of one piece, like a limpet's, but of eight, movably articulating and surrounded by a fleshy or scaly girdle that holds the animal close to the ground. This can be lifted up at any point, to take in a water current to the numerous gills lying in a shallow mantle cavity at the sides of the foot. The mouth is in front and the anus, renal, and genital pores at the back. The exposed shell surface is studded by numerous light-sensitive and tactile organs. Different from chitons, but evidently related to them, are the deep-water, wormlike Aplacophora, aberrant in structure, with the shell entirely lost.

Body plan of Monoplacophora

The Monoplacophora have a one-piece limpet-like shell. They are regarded as the most archaic mollusks and were long believed extinct. Consequently, the discovery of a living member of the family Tryblidiidae, *Neopilina galatheae*, at 5,000 metres depth off the Pacific Coast of Mexico by Henning Lemche of the Galathea Expedition (1950–53) was one of the greatest events in the history of malacology. This "living fossil" has a shell up to four centimetres long, and a broad flat foot. The mouth lies in front, the anus behind, a condition considered primitive. The mantle cavity forms shallow lateral grooves, as in chitons, but has five digitately branched gills. Figure 4 shows the internal structure of *Neopilina*. There are six pairs of renal organs and two pairs of gonads, with eight shell muscles on each side.

The class Scaphopoda contains the small elephant's tusk shells (*Dentalium*) and relatives. They are specialized in structure, with a tapered shell, open at both ends, a cylindrical foot, and fine tentacular filaments called captacula, attached in a cluster at each side of the mouth. These

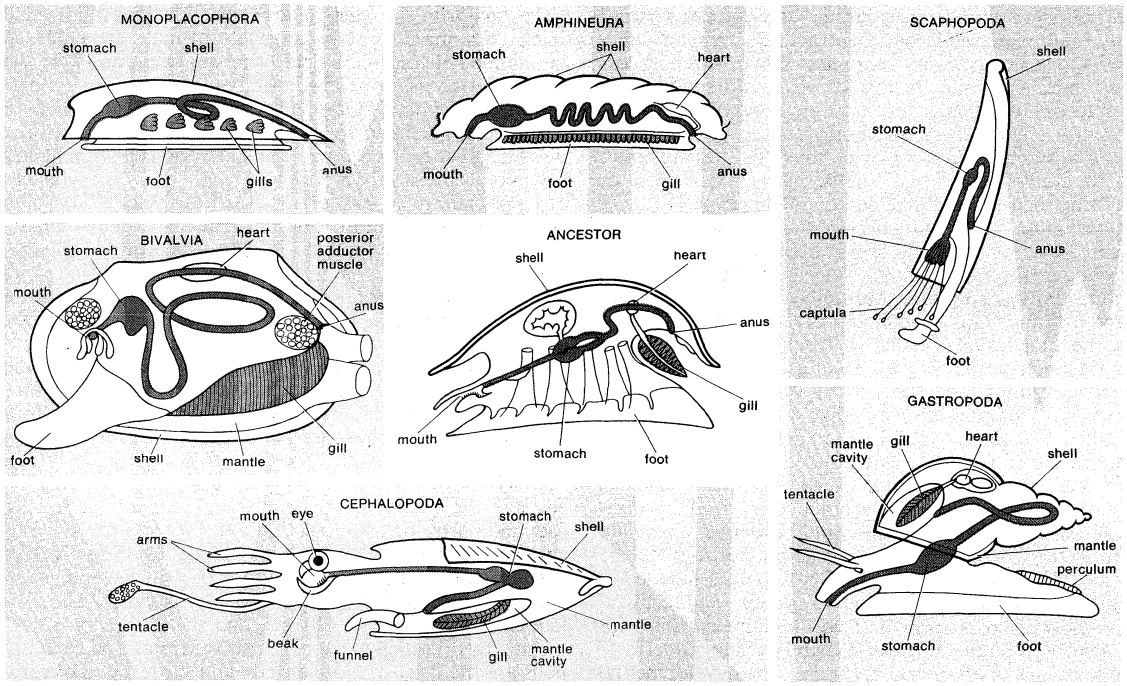


Figure 3: Body diagrams of the six classes of mollusks and a hypothetical generalized ancestor. Dark gray indicates the digestive tract; lighter gray represents the respiratory system. Scaphopoda, which have no specialized respiratory system, respire directly through the fleshy mantle.

emerge from the aperture of the broad end of the shell when the foot is retracted and are used to capture prey and carry it to the mouth for swallowing. The narrow end of the shell projects from the sand and through it the animal passes inhalant and exhalant water currents.

Internal structure. With the great diversity of the mollusks, an account of the internal structure can best be given from an idealized mollusk, combining the general features of the phylum (Figure 3).

Plan of the generalized mollusk

Early mollusks must have had two components, both of them variously surviving in all modern forms: (1) the head-foot, with a flat muscular crawling surface and snout held close to the ground, and (2) the viscero-pallium, consisting of visceral mass, the mantle or pallium (the fleshy outer fold of the integument) and mantle cavity. Attachment to the ground was by mucus, and locomotion by cilia, or a backward muscular wave. Muscles from the broad pedal base extended up to insertion on the shell, to withdraw the animal or pull the shell firmly against the ground. In the idealized mollusk the mouth leads to a muscular buccal mass containing the odontophore, or "tongue," bearing a radula, a chitinous sheet studded with microscopic teeth. This is rolled forward over the ground, abrading and sweeping up particles of food. The glands of the alimentary canal include salivary glands, esophageal pouches (in different ways modified among higher forms), and a bulky digestive gland forming most of the visceral mass. The brown lobes of this gland have several functions, to take up particulate food by engulfing, to complete digestion intracellularly, or (in most higher and carnivorous mollusks) to secrete extracellular enzymes, as well as to store protein, fats, glycogen, and calcium salts. The stomach is primitively a flasklike sac, with its neck continued toward the intestine. Here it contains a mucous or fecal rod, rotated by cilia, which in bivalves and a few gastropods becomes the muco-protein crystalline style. Its functions are to serve as a capstan winding in food and as a store of starch-digesting enzymes.

The mantle cavity forms the commanding structural feature of almost all mollusks. Lying within the muscular fold of the pallium, which adds to the growing shell edge, the mantle cavity is posterior in primitive mollusks or extends as shallow grooves along the sides of the foot. Originally a cloacal space into which the anus and excretory organs discharged, it appears to have taken on a respiratory function by folds of its vascular wall, which have now become complex gills. Cilia upon the gills pump a continuous water current through the cavity and between the gill filaments. This serves for respiration and also—in the bivalves and a few of the gastropods—to bring microscopic particulate food.

A sense organ called the osphradium is set strategically to sample the entering water current at the threshold of the mantle cavity before it reaches the gill and provide chemical information. Into the exhalant current pass the feces, the renal and genital products, and sometimes secretions such as mucus and coloured protective substances (purple or brown). The hypobranchial gland secretes an abundance of mucus, binding and consolidating waste particles.

The gills or ctenidia have a characteristic structure in the majority of mollusks. They are almost always two in number, as in bivalves and gastropods. Some primitive forms have them multiplied as in chitons (up to 70 pairs) and the primitive cephalopod *Nautilus* (2 pairs). *Neopilina* has five pairs, but their truly ctenidial nature is doubtful. The ctenidial filaments have well-developed ciliary tracts. As well as maintaining the water current, they serve in bivalves for filtering and retaining food particles, for conveying them to the mouth, and sometimes for sorting particles by size.

Much of the body space is a hemocoele; *i.e.*, an expanded series of venous sinuses, virtually restricting the true coelom (body cavity) to the pericardium. Venous blood is returned to the two gills and thence by way of the two auricles to the ventricle, from which anterior and posterior aortas distribute it to the tissues. The blood has only sparse stellate or amoeboid corpuscles. In gastropods and cephalopods the plasma contains the pigment hemocyanin; some muscular parts such as the buccal mass locally have myoglobin. Hemocoelic blood has an important role as a fluid skeleton, giving to the body turgidity without loss of flexibility. It massively fills the foot, providing a firm base for the action of the complex muscles. Tentacles, siphons, and other appendages are erected by blood flow.

Excretion from the blood takes place by hydrostatic pressure, passing into the pericardium a filtrate (minus proteins) from the blood in the heart. The renal organ walls are themselves glandular and special excretory glands may develop upon the auricle walls. Marine mollusks excrete ammonia or urea; terrestrial gastropods frequently produce uric acid.

In the early mollusks the reproductive system was simple, a single sac, possibly producing both eggs and sperm, opening into the pericardium and using the renal organs as gonoducts. Chitons, primitive gastropods, scaphopods, and most bivalves and cephalopods have the sexes separate. Some prosobranch and all opisthobranch and pulmonate gastropods are hermaphroditic. In these gastropods and in cephalopods fertilization is internal. The female genital ducts have a glandular role, investing the eggs with albumin, and protective capsules, or jelly mass. They also store outgoing and incoming sperm. Even in hermaphrodite species, the vast majority practice cross-fertilization, copulating and exchanging sperm.

The nervous system shows an enormous range of elaboration between chitons and primitive gastropods and advanced cephalopods with their complex behaviour and powers of learning.

Primitively, there was evidently a nerve ring around the esophagus, with dorsal cerebral centres linked at the sides to centres in the viscera and foot. The former gave rise to the visceral loop, supplying the pallial organs and visceral mass, and carrying visceral ganglia; the latter produced the nerves of the foot, primitively a ladderlike system of ganglionic cords. The sense organs include, in addition to the cerebral tentacles, paired sacs called otocysts or statocysts, in or close to the foot, and paired eyes. The otocysts apparently function in the perception of balance and motion. The eyes range from simple pigmented cups without image formation to the complex eyes of the cephalopods, virtually as intricate as those of higher vertebrates. For further information on these organs, see MECHANORECEPTION and PHOTORECEPTION.

EVOLUTION AND PALEONTOLOGY

Paleontology. The distribution of the molluscan classes in time may be briefly summarized. Chitons make an early appearance in the Ordovician, 500,000,000–430,000,000 years ago. The Monoplacophora extend from the Cambrian, about 500,000,000 years ago, while among the Gastropoda the true limpets Pleurotomariidae and Bellerophontidae with one or two other families go back to Cambrian times, 550,000,000–500,000,000 years ago. The scaphopods begin in the Ordovician, as do the bivalves of the Nuculidae, Arcidae, and some of the Pteriacea. The anisomyarian lamellibranchs were on the whole predominant in the Mesozoic (220,000,000–65,000,000 years ago), being superseded by siphonate and burrowing eulamellibranchs in the Tertiary (65,000,000–25,000,000 years ago). The oldest cephalopods, the Nautiloidea, were initiated in the Upper Cambrian (500,000,000 years ago), reaching a maximum in the Silurian and declining through Devonian to the Triassic, about 200,000,000 years ago. The ammonoids came in with the Devonian and persisted until complete extinction in the Cretaceous, about 100,000,000 years ago. The earliest of the coleoid cephalopods were the shelled belemnoids, from Triassic to Eocene; the squids and octopods were evidently delayed in maximum development until Tertiary and recent times.

Evolutionary trends. The fossil record gives little clue as to how the six classes diverged from the molluscan base. Useful as it may be as a concept, the structure of the presumed "archi-mollusk" must still be largely inferred from comparative morphology. In the present picture, the phylum shows three hugely successful stocks: the Gastropoda, Bivalvia, and—less diverse but in most ways the zenith of molluscan evolution—the Cephalopoda. There is as well a primitive pool of smaller classes, of which only the Scaphopoda shows any distinctive line of advance. The Polyplacophora and Aplacophora are so different from each other as probably to deserve separate classes. It is chiefly from the Polyplacophora and Monoplacophora (*Neopilina*), along with the earliest Gastropoda, that something can be gleaned about the archaic condition.

Significance of *Neopilina*

Neopilina is of inestimable value as a living fossil. Its undoubtedly primitive features include the restricted pallial cavity, the gut structure (especially the stomach), the larger extent of coelom, the relations between gonads and renal organs, and the ladderlike nervous system. The claims for segmental character, however, with implications of annelid origins, must be taken cautiously. Five pairs of gills, two pairs of gonads, two pairs of auricles, six pairs of renal organs, with variable multiplication of shell muscles, give a strange segmental arithmetic. They would appear to represent not true metamerism (segmentation), but an example of the widespread tendency for repetition of many organ systems in archaic mollusks. *Nautilus*, for example (and perhaps other primitive cephalopods), has two pairs of pallio-pericardial organs, ex-

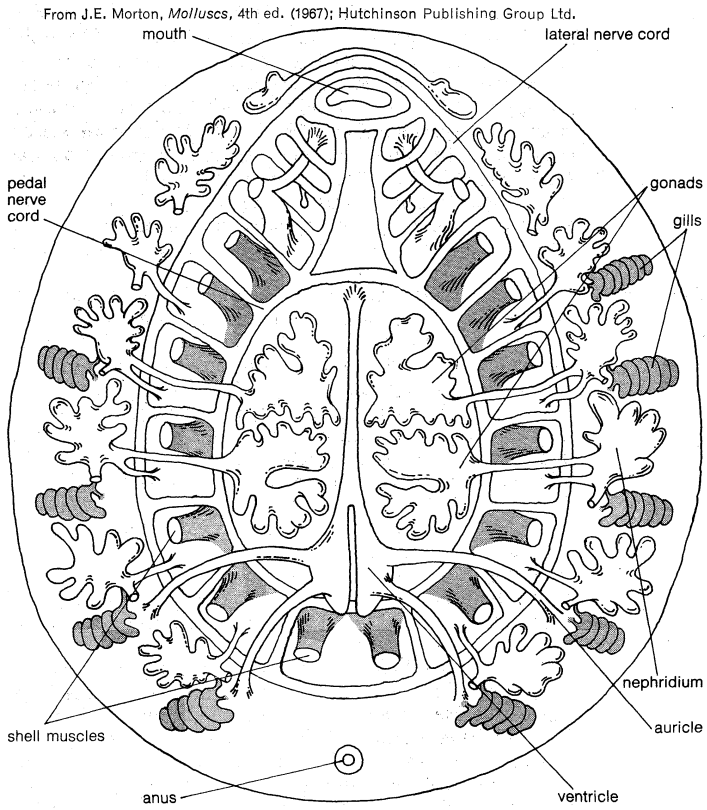


Figure 4: Monoplacophora, internal anatomy of *Neopilina*.

plicable on functional grounds. Chitons have up to 70 gill pairs, eight shell plates, and a varying number of shell muscles.

The relation of mollusks to other phyla is still obscure. In spite of the similar spiral cleavage and trochophore larva, the connection with annelids is not strong. The segmented worms have a fixed metamerism, with a locomotor wave very different from that of the mollusks. The backwardly directed wave of the molluscan foot suggests stronger ties with Turbellaria, a class of flatworms. Several workers have reviewed the evidence for a distant turbellarian relationship. They conclude that the rhythmic contractility of a flatworm presumably became confined to the molluscan ventral surface, and the dorsal parts became visceralized, covered by a mantle shell, and acquired a radial growth pattern.

Evolution of gastropods

It is, on the whole, the Gastropoda that keep closest to the basic molluscan pattern. They have, however, evolved a multifarious diversity of shapes and forms with different habitats.

The most characteristic feature of gastropods is torsion; moreover, all have a spiral coiling of the visceral mass. The early *Bellerophon*, of the extinct order Amphigastropoda, had a paired pallial complex and a bilaterally symmetrical flat spiral. Most living gastropods present an asymmetry of the pallio-pericardial organs, with the suppression of the right members after torsion. Early patterns remain in the prosobranch order Archaeogastropoda, in which *Pleurotomaria, Haliotis*, and the keyhole limpets have paired gills and pallial complex.

The full evolutionary potential of the gastropods was

realized in the order Pectinibranchia. In this group glandular genital ducts, a penis, and internal fertilization made possible early colonizations of the land and freshwater. Most other prosobranchs remained marine and the orders Mesogastropoda and Neogastropoda radiated into large and diverse groups, each evolving toward carnivorous life.

The prosobranchs are outstripped in adaptive novelty only by the Opisthobranchia. Here, as shown in several parallel lines, there is an evolutionary tendency toward early reversal of torsion, loss of the shell, and then loss of the mantle cavity. The plan typified by the slugs is typical of this large group, with the body variously and beautifully coloured or ornamented with appendages. The opisthobranchs have evolved two successful pelagic offshoots in the swimming pteropods, Thecosomata and Gymnosomata, parallel but not closely related. The land Pulmonata are the largest subclass of mollusks, but by comparison with the opisthobranchs, their structural evolution has been conservative. Their impressive achievements are physiological: adaptations for terrestrial life and associated water economy; adaptations for dormancy to avoid cold and desiccation; and specialization of reproductive and courtship behaviour.

Evolution of bivalves

By contrast with the gastropods, the Bivalvia are specialized in a restrictive manner. The mantle cavity has been converted to ciliary feeding by ctenidia and there is an efficient system of adductor muscles for rhythmically controlled opening and closing of the shell. The Protobranchia, undoubtedly primitive, but perhaps polyphyletic, constitute an archaic stock with the gills still relatively small and posterior in position, showing links with the original molluskan condition. The foot is broad and plug-like with a flattened distal expanse, comparable to a sole. The protobranchs have relatively huge labial palps, the tentacles of which are used by some species for food gathering. This may well have been the original bivalve habit, the mouth being withdrawn from the ground, then the head and buccal mass, leading to feeding on micro-organisms.

Subsequent evolution of bivalves is marked by great emphasis on pallial feeding and sorting mechanisms. A large group, the Heterodonta, most of which appeared in the Tertiary, have become adapted for burrowing deeply into a soft substrate, the mantle edges becoming fused and extended into siphons. Another large assemblage, the Anisomyaria (pearl oysters, wing shells, scallops, true oysters, and file shells) either live attached at the surface, by the byssus threads or by firmly cementing the shell to the substratum. The scallops have become free-living and "swim" by clapping the shell valves.

Evolution of cephalopods

The Cephalopoda must have arisen from a very early molluscan stock, and primitively were distinguished by pointed, caplike external shells. The first advance must have been the lightening of the shell, by withdrawal of the viscera from the apex and enclosure of air in the chambers of the shell. In modern nautiloids these chambers are not sealed-off buoyancy tanks but are under control like a fish's swim bladder. It has recently been shown that living *Nautilus* varies its specific gravity by admitting or expelling small quantities of water under active osmotic control. Along with shell buoyancy, new agility was acquired in cephalopods with muscular expulsion of water from the mantle cavity. In the shelled forms, such as the modern *Nautilus*, this is controlled largely by the funnel. In naked modern forms the muscular mantle is itself strongly contractile. The foot early lost its original functions and developed into tentacular processes, later taking the form of ten or eight suckered arms. Of the huge assemblage of ammonoid species none is living, and of nautiloids only one genus. Nautiloids, ammonoids, and the early shelled coleoids provide one of the most detailed evolutionary sequences of the past, but the cuttlefish, squids, and octopus, with their behavioral specializations, are a success story of relatively modern date.

CLASSIFICATION

Distinguishing taxonomic features. The six classes of Mollusca are defined primarily on the basis of overall body plan (symmetry and the presence of torsion in the adult or larva), the type of shell, and modifications of the internal organs.

Annotated classification. This classification is a consensus of recent views, especially those of L.H. Hyman, C.M. Young, K.M. Wilbur, and J.E. Morton.

PHYLUM MOLLUSCA

Unsegmented metazoans with body originally divisible into head-foot and viscero-pallium, the latter covered with a mantle or pallium that secretes a calcareous shell. No true endoskeleton, body cavity a hemocoele, with true coelom generally reduced to gonocoele and pericardium. External body surface frequently ciliated and mucous. A fine-toothed radula, lying in the buccal mass, is characteristic of most mollusks.

Class Monoplacophora

Extinct forms from Lower Cambrian through Devonian; one living genus, *Neopilina*, with four species. Bilaterally symmetrical, with ventral foot and posterior anus; one-piece oval shell; mantle cavity with five pairs of lobate branchiae. Dorsal and ventral coelomic cavities well developed. Two pairs of auricles; 2 pairs of gonads, discharging through 2 of the 6 pairs of renal organs. Primitive nervous system, with long pallial and pedal cords. Deep ocean bottom; known only from eastern Pacific.

Class Polyplacophora or Amphineura (chitons)

Elongated, flattened, bilaterally symmetrical, with posterior anus; mantle extensive, covering dorsum and sides; broad foot for adhesion to substrate; eight articulating dorsal shell plates, bordered by spiculose or scaly band; numerous pairs of gills (ctenidia). Dorsal posterior heart, with ventricle and lateral auricles; primitive nervous system, with longitudinal ganglionic cords. Marine littoral; length 1–33 cm (0.5–13 in.).

Class Aplacophora (solenogasters)

Wormlike; mantle surrounding body, except (in some) for ventral groove of foot; naked, except for minute calcified spicules; mantle cavity small and posterior. Marine; in deep water ooze or on sessile coelenterates. Length 2.5 cm (1 in.).

Class Gastropoda (snails and slugs)

Asymmetrical, with well-developed head and—at least primitively—a broad foot. Shell in one piece, coiled in a helical spiral; sometimes absent in adults. Viscero-pallium showing torsion of 180°; pallio-pericardial complex reduced and one-sided. Marine, freshwater, and terrestrial (see GASTROPODA).

Subclass Prosobranchia

Mostly marine snails. Generally aquatic; torsion pronounced and retained; visceral nerve loop crossed in figure 8. Head with paired tentacles and eyes at base. Spiral shell closed by an operculum. Mantle cavity primitively with two ctenidia, usually reduced to one (post-torsionally on the left). Heart posterior to mantle cavity. Solitary gonad opens primitively through right kidney, but later forms develop a glandular genital duct. Usually, where marine, a free-swiming veliger larva. Adult length from less than 1 to 60 cm.

Subclass Opisthobranchia (sea slugs)

Hermaphroditic, with shell reduced and becoming internal. Torsion usually reversed, with mantle cavity moving back along right side, and visceral loop shortening and uncrossing. Mantle cavity ultimately lost. Various adaptive gills, never ctenidia. Eventual return to bilateral sluglike external symmetry, with great range of form, colour, and locomotion. Usually with a (reduced) free-swimming veliger larva. Marine; length to 43 cm (16 in.).

Subclass Pulmonata (land snails)

Hermaphroditic, generally terrestrial, but some freshwater and primitive members marine. No ctenidium; mantle cavity generally vascularized as a lung. Small contractile pallial aperture. Detorsion seldom complete, but nervous system concentrated so as to lose all trace of crossing of nerves. Shell and visceral mass generally spiral, but some assume naked sluglike form. Length 0.5–20 cm (0.2–8 in.).

Class Scaphopoda (tusk shells)

Bilaterally symmetrical, mantle and shell elongated to a tapered tube open at each end. Foot pointed and cylindrical, used for digging. No ctenidia; heart, or distinct blood vessels. Head without eyes but carrying bunches of slender tentacles (captacula), for taking foraminiferans as food. Radula well-developed. Sexes separate without special genital ducts. Fertilization external. Marine, worldwide; length 2–15 cm (0.8–6 in.).

Class Bivalvia (clams, oysters and relatives)

Bilaterally symmetrical with head rudimentary, lacking radula, buccal mass or head sense organs. Ciliary feeders, collecting particles with enlarged and highly specialized

ctenidia. Large labial palps used for particle sorting. Mantle lobes completely enclose laterally compressed body; shell consists of two separately calcified valves and dorsal ligament. Foot strongly muscular, compressed and adapted for burrowing. Sexes separate or united (hermaphroditic). Genital ducts simple, paired, with fertilization external. Usually long larval life. Marine and freshwater; worldwide; length 0.5–170 cm (0.2–52 in.). See BIVALVIA.

Subclass Protobranchia

Small and primitive; ctenidia with flat, nonreflected filaments, foot with flat ventral surface; body organization primitive. Feeding chiefly by tentacular prolongations of labial palps.

Subclass Lamellibranchia

The majority of bivalves, including mussels, oysters, clams and scallops. Ctenidia much larger in relation to palps and form feeding organs; filaments elongated and narrow, reflected to form two-sided lamellae. Attached by a byssus; or burrowing, with siphons.

Class Cephalopoda (octopuses, squids, cuttlefishes)

Bilaterally symmetrical mollusks with a circle of tentacles around the head. Foot modified as a funnel, from which a propulsive water jet is produced. Nervous system and sense organs, especially eyes, highly developed. Marine, worldwide (see CEPHALOPODA).

Subclasses Nautiloidea and Ammonoidea

Shelled cephalopods of vast number and diversity. All extinct, except living *Nautilus*, with chambered shell, numerous nonsuckered retractile arms, two pairs of gills.

Subclass Coleoidea

Cuttlefish, squids, octopus, and extinct belemnoids. Modern cephalopods, with mantle naked, shell internal or lacking. Head with eight sucker-bearing arms and in squids and cuttlefish (Decapoda) a pair of long tentaculate arms in addition. In Octopoda, eight long tentaculate arms of equal length.

Critical appraisal. The above classification shows certain changes from the traditional division of the mollusks. With the discovery of the primitive class Monoplacophora, there is much to be said for raising also to the level of a class the small aberrant Aplacophora, which, except in a basic archaism of structure, are very remote from the chitons (Polyplacophora). In the Gastropoda, the older practice of separating Streptoneura (Prosobranchia) from Euthyneura (Opisthobranchia and Pulmonata) has been discontinued. It is better to regard the pulmonates and opisthobranchs as separate assemblages coequal in rank and diverging widely from an early point where they seem to have arisen close together from prosobranchs. With the use of the continental class name "Bivalvia," the old term Lamellibranchia has been lowered to a subclass name. It expresses admirably the type of gill structure of the vast majority of bivalves, as contrasted with the diversity of archaic forms gathered into the Protobranchia. Some would raise the septibranch bivalves (Cuspidariidae and Poromyidae) to a separate subclass; but they appear to follow neatly upon certain families of the higher Lamellibranchia. The old division of Cephalopoda into Tetrabranchia and Dibranchia has long disappeared. There is as yet no information on the gill number of any ancient shelled cephalopod, save for the possibly atypical genus *Nautilus*.

BIBLIOGRAPHY

Books for the general reader: R.T. ABBOTT, *Sea Shells of the World* (1962), a field guide designed for the shell collector; V. FRETTER and A. GRAHAM, *British Prosobranch Molluscs* (1962), an important work on the biology and identification of European prosobranchs; J.E. MORTON, *Molluscs*, 4th ed., (1967), a broad discussion of the biology of the entire phylum; R.D. PURCHON, *The Biology of the Mollusca* (1968), gives a deeper coverage of many topics and an excellent bibliography; K.M. WILBUR and C.M. YONGE (eds.), *Physiology of Mollusca*, 2 vol. (1964–66), a somewhat more technical account of the biology of the group; and C.M. YONGE, *Oysters* (1960), a readable description of the biology of this important species.

Books and papers of a technical nature: D.R. CROFTS, "Muscle Morphogenesis in Primitive Gastropods and Its Relation to Torsion," *Proc. Zool. Soc.* 125:711–750 (Nov. 1955), a clear description of the torsion process; E.J. DENTON and J.B. GILPIN BROWN, "On the Buoyancy of the Pearly Nautilus," *J. Mar. Biol. Ass. U.K.*, 46:723–759 (1966), provides insight into the mechanism of bouyancy in *Nautilus*; A. GRAHAM, "Molluscan Diets," *Proc. Mala. Soc. Lond.* 31:144–159 (June 1955), discussion of the origins of mollusks and the evolution of feeding behaviour; J.E. MORTON, "Evolutionary Trends in Molluscan Classification," *Proc. Linn. Soc.*, 174:53–72 (1963), discusses the modern taxonomic arrangement; L.H. HYMAN, *The Invertebrates*, vol. 6, *Mollusca I* (1967), a detailed but rather technical account of the biology of the Aplacophora, Polyplacophora, Monoplacophora, and Gastropoda; H. LEMCHE, "A New Living Deep-Sea Mollusc of the Cambro-Devonian Class Monoplacophora," *Nature*, 179:413–416 (Feb. 23, 1957), the first description of *Neopilina;* and with K.G. WINGSTAND, "The Anatomy of *Neopilina galatheae*," *Galathea Rep.*, 3:9–72 (1959), a full description of the anatomy of this remarkable animal; H.W. LISSMANN, "The Mechanism of Locomotion in Gastropod Molluscs," *J. Exp. Biol.*, 21:58–69 (Jan. 1945), on gastropod locomotion; J. THIELE, *Handbuch der systematischen Weichtierkunde* (1931–35), a classic German work on the biology of the Mollusca; M.J. WELLS, *Brain and Behaviour in Cephalopods* (1962), a useful book on the behaviour of these advanced mollusks; C.M. YONGE, "The Monomyararian Condition in the Lamellibranchia," *Trans. R. Soc. Edin.*, 62:443–478 (1953), deals with an important group of lamellibranch bivalves, and "On the Primitive Significance of the Byssus in the Bivalvia and its Effects in Evolution," *J. Mar. Biol. Ass. U.K.*, 42:113–125 (1962), a survey of the structure and evolution of the bivalve byssus; J.Z. YOUNG, "Learning and Discrimination in the Octopus," *Biol. Rev.*, 36:32–96 (Feb. 1961), an excellent discussion of learning in the octopus.

(J.E.Mo.)

Moltke, Helmuth von

Chief of the Prussian and German general staffs, Helmuth von Moltke was the military architect of the victories over Denmark (1864), Austria (1866), and France (1870–71) that enabled Otto von Bismarck to transform a loose league of German states into an empire dominated by Prussia.

Early career. Helmuth Karl Bernhard von Moltke was born on October 26, 1800, at Parchim, in Mecklenburg (since 1952 in the German Democratic Republic's district of Schwerin). His father, a man of unstable character, belonged to the nobility of Mecklenburg, his mother to an old family of the free city of Lübeck. The Moltkes were impoverished; and the young Moltke, whose health was not too good, had an unhappy start to life. Since his father had emigrated to Holstein (then a Danish possession) in 1805 and had taken Danish nationality, the boy completed his education with the Royal Cadet Corps in Copenhagen and joined a Danish infantry regiment. After a visit to Berlin in 1821, however, he decided to transfer to the Prussian Army, and in 1822 he obtained a commission as a second lieutenant in the Prussian Life Guards, being posted to Frankfurt an der Oder.

In October 1823 Moltke was sent to the General War College for a three-year course; but, since his health was deteriorating, he went in the summer of 1825 to Bad Salzbrunn for convalescence, during which he studied modern languages. On his return to Frankfurt in the summer of 1826 he took up writing, largely to improve his financial position, and published a novel, *Die beiden Freunde* (1827; new ed., 1957). In May 1828 he was transferred to the General Staff's Topographical Bureau in Berlin.

Continuing his literary work, Moltke published his *Darstellung der innern Verhältnisse und des gesellschaftlichen Zustandes in Polen* (1832; Eng. trans., *Poland: an Historical Sketch*, 1885) and also accepted a publisher's offer for a translation of Edward Gibbon's *Decline and Fall of the Roman Empire*; although he completed most of it, it was never published.

Moltke was attached to the Prussian General Staff in March 1832 and promoted to the rank of first lieutenant a year later. Toward the end of 1835 he was sent to Turkey to advise Sultan Mahmud II on the modernization of the Turkish Army; and in 1836 he was authorized by Berlin to enter the Turkish service. After some work in Istanbul and travels in the Balkans, he went in 1838 to Armenia, where Turkish forces were preparing an offensive against the Egyptian invaders of Syria. The failure of the offensive, in 1839, was due largely to the Turkish commander's disregard of Moltke's advice.

Mission in Turkey

Returning to Germany at the end of 1839, Moltke re-entered the Prussian service. He published a selection of his letters from Turkey, *Briefe über Zustände und Begebenheiten in der Türkei,* in 1841. In April 1842 he married an English girl, Marie Burt (1825–68), stepdaughter of his sister Augusta. An essay on the considerations that should govern the choice of routes for new railways (1843) is more significant of his future than his next book, *Der russisch-türkische Feldzug in der europäischen Türkei 1828–1829* (1845; Eng. trans., *The Russians in Bulgaria and Rumelia,* 1854).

Late in 1845 Moltke was appointed personal aide-de-camp to the invalid Prince Henry of Prussia (1781–1846), who was living privately in Rome. Though this appointment ended with Prince Henry's death, it gave Moltke the opportunity to begin work on a splendid topographical map of Rome and its vicinity (published in 1852) and to write his "Wanderungen um Rom" (published in his *Wanderbuch,* 1879; Eng. trans., *Notes of Travel,* 1880). Moreover, when the warship bringing Prince Henry's body back to Germany reached Gibraltar, Moltke left it and made his own way home across Spain, recording his impressions in his "Tagebuchblätter aus Spanien" (also published in the *Wanderbuch*).

Working again for the Prussian General Staff, Moltke started thinking about the major problems confronting Prussia: the unification of Germany and the security of a country surrounded by potential enemies yet lacking natural frontiers. He was convinced that only the army could provide a satisfactory solution to these problems; and he expressed his delight at the suppression of the revolution of 1848–49 by writing to his brother that the curtain had come down on Prussia's worst enemy—democracy.

A colonel from 1851, Moltke was in 1855 appointed personal aide-de-camp to Prince Frederick William, the future king of Prussia and German emperor (Frederick III). This post involved more travels: to Scotland and thence to England; to Russia (*Briefe aus Russland,* published 1877; Eng. trans., *Letters from Russia,* 1878), and to France. Even so, apart from the period of his Turkish service and his brief stay in Rome, Moltke's career up to 1856 had been mostly dull routine or concerned with the affairs of his princely masters rather than the profession of arms. Only in Turkey had he commanded troops.

Career as chief of the General Staff. Moltke was selected as chief of the Prussian General Staff in 1857 and confirmed in that office in September 1858. Thus began the era of the great triumvirate—Otto von Bismarck (chancellor), Moltke, and Albrecht von Roon (1803–79; minister of war from 1859)—that within 13 years was to change the map of Europe.

Moltke entered upon his new official duties at a time when a technical revolution was changing the whole conception of war. The rearmament of the German infantry with the breech-loading needle gun had been proceeding since 1848 and was almost complete. Breech-loading guns for the artillery were on the way but were not finally introduced until 1861. Much more significant, however, was the rapid development of railways.

Military use of railways

Moltke was the first military man to appreciate the important role that railways could play in the deployment, movement, and supply of armies on a great scale. Hitherto, the movement of troops had been restricted by the paucity and seasonal unreliability of road communications, and the aim of strategy of every great commander had been to bring the strongest possible force in the best possible condition onto a small battlefield where he could survey the entire army and almost every soldier could hear him. Moltke saw that the advent of railways had changed this. Many more men and much more equipment could now be deployed much faster on vastly wider fronts. In place of battlefields of a few square miles, there would be long battlefronts of perhaps hundreds of miles. The limited flank attack by a few battalions would give place to wide turning movements by many divisions. Supplied by railways, troops would be able to keep the field in all weather throughout the year.

At the same time, Moltke appreciated that a new technique and a much more highly trained body of staff officers would be required to realize his new conception of warfare. The mounted staff officer with his necessarily confined outlook would give way to one with a much wider view, capable of compiling intricate railway-movement tables for vast numbers of men, animals, and equipment and of arranging for daily trainloads of supplies. He saw, too, that changes were necessary in another direction. Whereas, previously, commanders had kept a very tight hold on their subordinates and had been able always to give short and explicit orders, it was clear to Moltke that this system would not work in an army of perhaps millions locked in battle along a front that might extend for hundreds of miles. He therefore instituted the system of "general directives" in place of rigid "operation orders." In these directives the recipient was given a long-term task in general terms but was allowed considerable latitude and was expected to use his discretion and initiative in carrying it out. When Moltke joined the General Staff, its "chairborne" officers were held in little esteem by the "real" soldiers. He built up the new system of the Prussian General Staff, which later became the model for all armies organized on modern lines.

The Seven Weeks' War (1866). Some of Moltke's theories and methods were tested in practice by the Prussian forces in the short German–Danish War over Schleswig-Holstein in 1864; but a far more instructive opportunity of testing them came with the Seven Weeks' War, which Prussia, under Bismarck's guidance, launched against Austria and certain other German states in the summer of 1866. This war, for which initially Prussia deployed some 256,000 men in three armies on a front of 260 miles (nearly 420 kilometres), culminated in Prussia's overwhelming victory at Königgrätz (Sadowa). Moltke's contribution to victory was rewarded by Prussia with a donation of money that enabled him to buy an estate at Kreisau in Silesia (since World War II in the Wrocław Province of Poland); and the military result vindicated his system in the eyes of some older and hitherto recalcitrant generals. The war also served to expose deficiencies in the functioning of the system, among them ill-trained staff officers and an inadequate intelligence service. With the newly secured cooperation of his subordinates, he remedied these shortcomings before the final proving of his notions in 1870–71. He was also responsible for the official Prussian history of the Seven Weeks' War, *Der Feldzug von 1866 in Deutschland* (1867; Eng. trans., *The Campaign of 1866 in Germany,* 1872).

The Battle of Königgrätz

The Franco-German War (1870–71). Prussia's triumph of 1866 excited the envy of France, whose emperor Napoleon III was tempted to seek an increase in prestige at Prussia's expense or through intervention in German affairs. To eliminate this challenge, Bismarck envisaged war against France; and Moltke and Roon, having profited from the lessons of 1866, were able to tell him, at the end of 1869, that in their judgment the Prussian Army was capable of defeating the French and that the time seemed ripe. Consequently, Bismarck, in the next few months, provoked Napoleon III into hostilities: the Franco-German War began in July 1870, Moltke deploying some 384,000 men in three armies.

The Battle of Sedan

The Germans' great victory at Sedan on September 2, 1870, brought the fall of the Second Empire in France and was soon followed by the proclamation, at Versailles on January 18, 1871, of a new German Empire. France sued for peace in February and accepted Bismarck's terms in May 1871. Moltke, whose military machine had been far more efficient than France's, was created Graf (count) in October 1870, after Sedan, and appointed field marshal in June 1871, after the peace treaty.

The last years of office. Moltke was chief of the General Staff for 17 more years from 1871. For part of this time he was occupied with planning for the eventuality of Germany's having to fight a war on two fronts—against Russia in the east as well as against France again in the west. In August 1888, however, the old man at last resigned his post, not least because of his lack of sympathy with the manners and ideas of his new sovereign, the emperor William II. He had, however, already chosen his own immediate successor, Alfred von Waldersee.

Personality. On resigning office in 1888, Moltke retired to Kreisau. He died, during a visit to Berlin, on April 24, 1891. A tall, spare figure, he had a tanned face that usually wore an expression of grave austerity. His acute intelligence was obvious to all who met him; but, though he was a considerable linguist, he was habitually so taciturn that he was described as being "silent in seven languages" (he knew at least German, Danish, French, English, Italian, and Turkish, besides any Slavic or Iberian languages that he might have picked up). No indiscreet or unkind word is recorded as having passed his lips, and to his military colleagues he became "the Golden Man," without enemies or detractors. His married life was affectionate and happy but childless. As a writer, he is sometimes reckoned among the masters of 19th-century German prose.

BIBLIOGRAPHY. There are editions of MOLTKE'S *Gesammelte Schriften und Denkwürdigkeiten*, 8 vol. (1891–93); of his *Militärische Werke*, 4 vol. (1892–1912); and of his *Briefe an seine Braut und Frau und an andere Anverwandte*, 2 vol. (1894; Eng. trans., *Letters to His Wife and Other Relatives*, 2 vol., 1896). Important selections from the two former collections can be read in English: *The Franco German War of 1870–71*, 2 vol. (1891; rev. ed., 1893); *Projects for the Campaign of 1866 Against Austria* (1907); and *Military Correspondence, 1870–71* (1922–23). The fullest modern study is by E. KESSEL, *Moltke* (1957). For a shorter biography, see E. VON NASO, *Moltke, Mensch und Feldherr* (1963). For the English reader there is still F.E. WHITTON, *Moltke* (1921). W. GORLITZ, *Der deutsche Generalstab: Geschichte und Gestalt, 1657–1945* (1950; Eng. trans., *History of the German General Staff, 1657–1945*, 1953), provides a rather good background, explaining Moltke's part in both the development and the implementation in war, of the general staff system.

(C.N.B.)

Mommsen, Theodor

A scholarly masterpiece as well as a great work of literature, Theodor Mommsen's *Römische Geschichte* (*The History of Rome*) made him the most popular of all German scholars of his time and spread his fame throughout the world. It earned him the Nobel Prize for Literature for 1902. Mommsen believed that the historian is perhaps more of an artist than a scholar. Yet anyone attempting to write a history of Rome based on original research obviously must master both the methods and materials of classical philology; he must also have a thorough knowledge of Roman law. Mommsen met these conditions and made a unique contribution to the development of a systematically historical treatment of antiquity covering encyclopaedically every aspect of the Classical period.

Education

Career. He was born on November 30, 1817, the son of a Protestant minister, in Garding, Schleswig-Holstein, and he grew up in Oldesloe (now Bad Oldesloe). Mommsen received his basic classical training in the senior classes of the Gymnasium (secondary school) Christianeum in Altona, then part of the duchy of Holstein. From 1838 to 1843 he studied jurisprudence at the University of Kiel; inasmuch as the study of jurisprudence in Germany at the time was largely a study of Roman law, this had an essential influence on the direction of his future research. He owed his idea of the close interrelationship between law and history not so much to his teachers as to the writings of Friedrich Karl von Savigny, one of the founders of the historical school of jurisprudence. After he had received his master's and his doctor's degrees, a research scholarship granted by his sovereign, the king of Denmark, allowed him to spend three years—from 1844 to 1847—in Italy. During this time Italy became his second home and the Archaeological Institute in Rome one of the headquarters from which he pursued his research. By that time Mommsen had already conceived the plan for the *Corpus Inscriptionum Latinarum*, a comprehensive collection of Latin inscriptions preserved since antiquity on stone, iron, and other enduring materials, arranged according to the basic principles of philological methodology. Having been prepared for this field by the young Kiel professor, Otto Jahn, he soon became a master of epigraphy—the study and interpretation of inscriptions—under the guidance of Bartolomeo Borghesi, the learned statesman of San Marino. Within the next several decades Mommsen made the corpus of Latin inscriptions into a source work that was essential in complementing the one-sidedly literary tradition and that, for the first time, made a comprehensive understanding of life in the ancient world possible.

Mommsen, drawing by Tycho Mommsen, 1844.
In the collection of Hellmut Mommsen.
By courtesy of Hellmut Mommsen

Career as a politician and journalist

When he returned from Italy, Mommsen found his country in a state of mounting unrest. As a native of Schleswig he was a subject of the Danish king, but he considered himself German, wanted to remain German, and looked forward to German unity. For him freedom meant not only the independence of the German states from foreign influence but also the freedom of the German citizen to adapt himself to any sort of constitution except that of despotism or a police state. A liberal, he considered the republic the ideal state, yet he was quite content with a constitutional monarchy so long as it was not a cover for some sort of pseudo-constitutional autocracy. Mommsen's political activities began with his editorship of the *Schleswig-Holsteinische Zeitung* for the provisional government established during the revolution of 1848. Yet journalism was not much to his taste; he was happy when, at the end of 1848, he was offered a professorship in civil law at the University of Leipzig. Nevertheless, he remained politically minded as long as he lived—as a thoughtful and critical observer as well as an active politician. (He was a deputy in the Prussian Landtag from 1873 to 1879 and in the German Reichstag from 1881 to 1884.) He continued to devote time and energy to politics, but it is doubtful that he thereby served his country's and his own best interests. While he was an acknowledged authority in his field of scholarship, in politics he remained a camp follower, who achieved no more than many others. Moreover, he more than once jeopardized his career by his political activities. Because of his participation in an uprising in Saxony in May 1849, he lost his professorship and almost landed in prison.

After his dismissal from his post in Leipzig, Mommsen in 1852 accepted a professorship in jurisprudence in Zürich. The grief he expressed about being an "exile" showed how deeply he felt himself to be a German. In 1854, however, he was offered a professorship in Prussia at the University of Breslau. It was at this time that he married Marie Reimer, daughter of a bookseller. During a long and happy marriage his wife bore him 16 children.

History of Rome

The historian and his works. During the years he spent at Leipzig, Zürich, and Breslau, Mommsen wrote the first three volumes of the *Römische Geschichte*, up to the Battle of Thapsus, 46 BC. This work embodied the new historical method applied to the history of Rome. Mommsen critically examined hitherto unques-

tioned traditions and rejected the attitude of the Enlightenment, which had idealized the classical age. He readily acknowledged himself to be a disciple of the historian Barthold Georg Niebuhr, who introduced rigorous criticism of sources into historiography, however much their methods of research and presentation differed and despite the fact that he went considerably beyond his great predecessor in demythologizing Roman history. In Mommsen's view it was important that the ancients should come down to earth from the Olympian heights upon which they appeared to the mass of the public. This modern style was not to everyone's taste, for, in bringing the past to life, he used the political and sociological vocabulary of the 19th century. When he speaks of the squirearchy and the cloth exerting their "malignant" influence even in ancient Rome, it is Mommsen the liberal politician speaking. Nevertheless, his *Römische Geschichte* is not a politically tendentious work but a piece of scholarship of the highest rank, which gains from its distinction of style.

The philologist is regarded as the preserver of verbal tradition, but as a philologist Mommsen was more than that: he was an artist, and he proved his artistry in his treatment of language. He disliked any incongruous mixture of prose styles and, in the *Römische Geschichte* and *Romisches Staatsrecht* ("Roman Constitutional Law"), he created two works both of which attain exemplary unity of form and content yet demonstrate two different styles. Without being a creative poet, he used the means of poetry and enjoyed exercising his poetic talent. An excellent testimony to his abilities is the *Liederbuch dreier Freunde* ("Songbook of Three Friends"), which he published in 1843 together with his brother Tycho and the writer and poet Theodor Storm. Throughout his life Goethe was his ideal not only as a poet but also as "the wisest man of the century." His perfect command of English, French, and Italian did much to make his journeys of research successful; he quoted Shakespeare in his letters almost as often as Goethe.

To many critics Mommsen's glorification of the dictator Caesar and his disparagement of Caesar's opponents, Pompey and Cicero, seem strangely inconsistent with his political liberalism. He tried to make his critics understand that he had praised Caesar only as a saviour of the decaying state; yet Mommsen's admiration for the autocrat reveals something of his own character. He himself was an autocrat in his own branch of scholarship, adopting a manner that his opponents labelled "caesarism." At the same time, however, he had an unusual need for the fellowship of like-minded men. He held personal contacts to be one of the most important elements of life; indeed, it might be said that he had a genius for friendship. Yet it was mostly a friendship with men who looked up to him. With anyone who considered himself Mommsen's equal, a friendly relationship was not likely to last long.

It was only as a member of the Prussian Academy of Sciences that Mommsen could pursue his project of publishing his collection of Latin inscriptions, and for this reason in 1858 he was offered a post in Berlin. In 1861 he also became a professor in the philosophy faculty at the university; because of his philological and historical interests he chose that faculty rather than that of law. As a teacher of Roman history and epigraphy—especially in his seminars—he trained many students who were later to make their mark in these fields. The main part of his scholarly work was taken up with the continuation of the *Corpus Inscriptionum Latinarum*. He also acted as adviser on many other great scholarly enterprises, such as the *Monumenta Germaniae Historica*, the exploration of the *limes* (Roman border fortifications in southwest Germany), the numismatic work of the Prussian Academy, and the *Thesaurus Linguae Latinae*. Even in old age his mind was open to the new demands of scholarship, as shown by his interest in the new study of papyrology.

Mommsen's historical work was interrupted by his work on inscriptions; thus, the *Römische Geschichte* was never completed. When, several decades later, in Berlin, he set out to complete his history, he abandoned the idea of writing the fourth volume, which was to contain the history of the emperors, because he felt that he would not be capable of writing it in the same brilliant style as his history of the republic. The fifth volume (1885) deals with the history of the Roman provinces in the first three centuries of the empire. No one but Mommsen could have depicted this period in so authoritative a manner, for no one else knew the nonliterary sources—the inscriptions and coins—as did he.

The greatest monument to Mommsen's scholarship, the work which is of even greater significance for scholars than the *Römische Geschichte*, is *Römisches Staatsrecht* ("Roman Constitutional Law"), published between 1871 and 1888. He himself said that if he were to be remembered by anything, it would be by this work. The Romans themselves never codified their constitutional law; Mommsen was the first to do it. His historical approach to classical scholarship led him to systematize the innumerable legal details upon which the Roman constitution was based and to explain this complex body of law through an understanding of its historical development. Only a man who, like Mommsen, was grounded both in law and in the classics would be in a position to investigate the public law of the Romans, and only a man trained to think in historical concepts could understand it.

"Roman Constitutional Law"

In public law, criminal law stands side by side with constitutional law, and Mommsen's last great work, published in 1899, is *Römisches Strafrecht* ("Roman Criminal Law").

When Mommsen, who had already become a mythical figure for his contemporaries, died in Berlin on November 1, 1903, just four weeks before his 86th birthday, he had attained what he had always wanted. The task which he had set himself to fulfill, according to his own almost superhuman standards, he had completed.

MAJOR WORKS

ROMAN STUDIES: *Römische Geschichte*, vol. 1–3 (1854–56; *The History of Rome*, 1862–75), vol. 5 (vol. 4 never written; 1885; *The Provinces of the Roman Empire*, 2 vol., 1909); *Geschichte des römischen Münzwesens* (1860); *Corpus Inscriptionum Latinarum* (1863 and after); *Römische Forschungen*, 2 vol. (1864–79); *Römisches Staatsrecht*, 3 vol. (1871–88); *Römisches Strafrecht* (1899).

OTHER WORKS: *Unteritalische Dialekte* (1850); *Inscriptiones regni Neapolitani Latinae* (1852).

The *Römische Geschichte* has been either completely or partially translated into Danish, English, French, Hungarian, Polish, Russian, and Spanish. English translations include *The History of Rome* by W.P. Dickson, 4 vol. (1862–75) with vol. 5 as *The Provinces of the Roman Empire* (1909); another translation in the Everyman series (1911); and abridgments by C. Bryans and F.J.R. Hendy (1888, new ed. 1959) and by D.A. Saunders and J.H. Collins (1958). There is a French translation of the *Geschichte des römischen Münzwesens* by le Duc de Blacas: *Histoire de la monnaie romaine*, 4 vol. (1865–73); and of the *Römisches Strafrecht* by J. Duqesne: *Le Droit pénal romain* (1907).

BIBLIOGRAPHY. KARL ZANGEMEISTER, *Theodor Mommsen als Schriftsteller: Ein Verzeichnis seiner Schriften*, revised and continued by EMIL JACOBS (1905); *Gesammelte Schriften*, 8 vol. (1905–13). Correspondence: with Ulrich von Wilamowitz-Moellendorff, ed. by FRIEDRICH and DOROTHEA HILLER VON GAERTRINGEN (1935); with Otto Jahn, ed. by LOTHAR WICKERT (1962); with Theodor Storm, ed. by HANS-ERICH TEITGE (1966). LOTHAR WICKERT, *Theodor Mommsen: Eine Biographie*, 3 vol. (1959–69, vol. 4 in prep.), an exhaustive biography.

Shorter studies: CARL BARDT, *Theodor Mommsen* (1903); OTTO HIRSCHFELD, "Gedächtnisrede auf Theodor Mommsen" (1904), in *Kleine Schriften*, pp. 931–965 (1913); LUDO MORITZ HARTMANN, *Theodor Mommsen: Eine biographische Skizze*, including an appendix "Ausgewählte politische Aufsätze Mommsens" (1908); W. WARDE FOWLER, *Theodor Mommsen: His Life and Work* (1909); ALFRED HEUSS, *Theodor Mommsen und das 19. Jahrhundert* (1956); HANS LIEBESCHÜTZ, "Treitschke and Mommsen on Jewry and Judaism," in *Year Book VII of the Leo Baeck Institute of Jews from Germany*, pp. 153–182 (1962); BERTHOLD SUTTER, "Theodor Mommsens Brief 'An die Deutschen in Österreich' (1897)," in *Ostdeutsche Wissenschaft*, 10:152–225 (1963); ALBERT WUCHER, *Theodor Mommsen: Geschichtschreibung und Politik*, 2nd ed. (1968); LOTHAR WICKERT, *Drei Vorträge über Theodor Mommsen* (1970).

(L.F.K.W.)

Monaco

A sovereign principality whose densely clustered hills and headland look southward over the Mediterranean Sea, Monaco (officially Principauté de Monaco) has an area of only about 0.73 square mile (1.9 square kilometres). Many unusual features, however, have made the principality among the most luxurious tourist resorts in the world and have given it a reputation far exceeding its size. Situated astride the resort area of the Cote d'Azur (French Riviera)—the city of Nice lies nine miles to the west, the Italian border five miles to the east—it has a mild Mediterranean climate with annual temperatures averaging 61° F (16° C) and with only about 60 days of rainfall. Monthly average temperatures range from 50° F (10° C) in January to 75° F (24° C) in August. Many of the visitors alternate their hours between the glistening beaches and boating facilities, the international sportscar races, and the world-famous Place du Casino, the gambling centre in the Monte Carlo section that made Monte Carlo an international byword for the extravagant display and reckless dispersal of wealth.

History of the principality. Evidences of Stone Age settlements in Monaco are preserved in its Museum of Prehistoric Anthropology. In ancient times the headland was known to the Phoenicians, Greeks, Carthaginians, and Romans. In 1191 the Genoese took possession of it, and in 1297 the long reign of the Grimaldi family began. The Grimaldis allied themselves with France except for the period from 1524 to 1641, when they were under the protection of Spain. In 1793 they were dispossessed by the French Revolutionary regime, and Monaco was annexed to France. With the fall of Napoleon I, however, they returned; the Congress of Vienna (1815) put Monaco under the protection of Sardinia. The principality was reduced in size in 1848 by the loss of territory to France. The Franco-Monegasque treaty of 1861 restored Monaco's independence, and in 1865 a customs union was established between the two countries. Another treaty with France in 1918 contained a clause providing that, in the event that the Grimaldi dynasty should become extinct, Monaco would become an autonomous state under French protection.

Beginning of Grimaldi rule

Contemporary Monaco. The four sections, or *quartiers,* of Monaco are the town of Monaco, or "the Rock," a headland jutting into the sea on which the old town is located; La Condamine on the west of the bay, with its natural harbour; Monte Carlo, including the casino and numerous hotels; and the newer zone of Fontvieille, in which various light industries have developed.

In Monaco are the Roman Catholic cathedral, which houses a Louis Bréa painting of St. Nicholas and which was built in 1875 on the site of an ancient church; the prince's Genoese and Renaissance palace; and the Oceanographic Museum of Monaco built in 1910.

On the Monte Carlo coast a 1,600-foot (500-metre) beach has been constructed, adding 22 acres to Monaco's territory. In the early 1970s, another strip of ground was being added in the Fontvieille district, upon which a new town was to be built overlooking the sea.

Monaco's population in 1971 was slightly over 23,000. More than half of the people were French citizens, some 4,000 were Italian, and only about 3,500 could claim Monegasque descent. Monaco's refusal to impose income taxes on its residents and on international businesses that have established headquarters there led to a severe crisis with France in 1962. A compromise was reached and French citizens with less than five years residence were taxed at French rates and by which taxes were imposed on Monegasque companies doing more than 25 percent of their business outside the principality.

Monaco has had a constitution since 1911, but, in 1959, Prince Rainier III suspended part of the constitution and dissolved the National Council; in 1961 he appointed instead a national assembly. The crisis of 1962 with France led him to restore the National Council and to grant a new, liberal constitution. The council comprises 18 members, elected by universal sufferage for a term of five years. Government is carried on by a minister of state (who must be a French citizen) and three state councillors acting under the authority of the prince. Legislative power is shared by the prince and the National Council. Since 1819 the judicial system has been based on that of France, and two Parisian judges form a Court of Appeal. There is also a Municipal Council (15 members elected for four years) that administers municipal affairs.

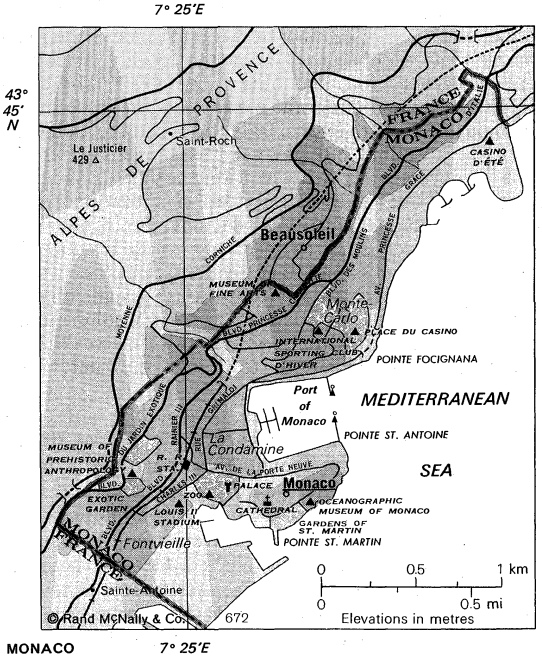
Administration of the principality

In the early 1970s, more than one-half of the government's revenues came from taxes on commercial transactions; the remainder was drawn from franchises on radio, television, and the casino, from state-operated monopolies on tobacco and stamps, from sales taxes, and from the taxes imposed since 1962.

In recent times, Monaco's chief industry has been tourism, and its facilities make it one of Europe's most luxurious resorts. Once a winter attraction, it now draws summer visitors as well to its beaches and expanded mooring facilities. The social life of Monte Carlo centres about the Place du Casino, with its picturesque gardens, and the nearby International Sporting Club d'Hiver, which houses a theatre. The casino was built in 1861, and in 1967 its operations were taken over by the principality. Characteristic of the principality's orientation toward tourism, the minority Monegasque portion of the population is prohibited from gambling activity—but it is also free from taxation.

Tourist attractions

The casino itself contains the home of l'Opéra de Monte Carlo as well as the Foyer de la Dance, which saw the premiers of most of the works by the famous Ballet Russe de Monte Carlo of Sergey Diaghilev. There is also a Monte Carlo national orchestra. The best known of the automobile events are the Monte Carlo Rally and the Grand Prix de Monaco. Radio Monte Carlo broadcasts in 23 languages to Europe and the Middle East, and Télé Monte Carlo's programs are commercially sponsored.



MONACO 7° 25'E

BIBLIOGRAPHY. A standard history of Monaco is L.H. LABANDE, *Histoire de la Principauté de Monaco* (1934); an older one is GUSTAVE SAIGE, *Le Protectorat espagnol à Monaco* (1885). GEORGES REYMOND, *Monaco antique* (1970), is a study of ancient texts and epigraphs.

Monasticism

The word monasticism is derived from the Greek *monachos* "living alone"; but the etymology indicates only one of the elements of monasticism as a force in history and society. The etymological method of arriving at an understanding of monasticism is, at any rate, misleading because a large section of the world's monastics live in cenobite (common life) communities. The term monasticism does, however, indicate what later became a socially

and historically highly significant feature; *i.e.*, living alone in the sense of being unmarried or celibate, though this feature is not directly related to its etymology.

Still, even this aspect of monasticism does not extend beyond the cultures and languages within which was formulated the religious terminology that originated in the eastern Mediterranean; *i.e.*, the Judeo-Christian and Islāmic religions. In the Islāmic world, terms that can be translated by "monk," "monastic," and similar words do not mean "single" in the Arabic and Persian terminologies as in the Greek. Other aspects (*e.g.*, *zuhd*, "asceticism") of the monastic life in Islām provided the etymological and definitional sets denoting "monasticism." None of the many Indic (Sanskrit, Pāli, Apabhraṃśa, Prākrit) terms for monk means "single" or "living alone" in these languages, although monastics within those traditions—Brahmin-Hindu, Buddhist, and Jaina—do indeed live alone or in groupings set off from the rest of their societies, analogously to Jewish, Christian, and Islāmic monastics. The etymologies of the Indian as well as of some of the Arabic and Persian terms connote poverty, certain ecstatic states of mind, dress conventions, and some (by historical rather than semantic connection) imply single, celibate living.

NATURE AND SIGNIFICANCE

Within a cross-cultural perspective, monastics can be seen to have been instrumental in creating, preserving, and augmenting institutions of learning, religious as well as secular, and in transmitting cultural goods, artifacts, and intellectual skills through the generations. Monastic institutions have also had medical, political, and military-related functions, though the latter two have all but disappeared in most societies.

Definition and characteristics

A definition of monasticism that would cover all of its forms, Eastern and Western, must necessarily be wide; particulars must be relegated to the analysis of specific monastic systems. A universal definition might thus be: "monasticism" is a term covering religious institutions, ritual, and belief systems whose agents, members, or participants attempt to practice religious works that are above and beyond those required by the religious teachings of their society or of exceptional individual religious and spiritual leaders in their society; *i.e.*, those who have interpreted radically the tenets that apply to all believers or to the whole society. Beyond such a statement, one can speak only of the major characteristics of the monastic life and its institutions, since none of them is universal. Celibacy is fundamental to the majority of the world's religious orders, but it is by no means universal; asceticism is universal, provided the term is defined widely enough so as to include all supererogatory religious practices. The truly universal characteristic of monasticism follows from its definition: the monastic separates himself from his society, either to be by himself as a hermit or an anchorite (religious recluse) or to join a society of others who have separated themselves from their surroundings with similar intentions; *i.e.*, the full-time pursuit of the religious life in its most radical and usually its most fundamentalistic interpretation.

There is no monasticism in societies that do not have a written and transmitted lore. Nonliterate societies do not have monastic institutions, since the monastic takes his departure from an established corpus, or written body, of religious doctrine, which has undergone criticism and has generated countercriticism as well as a dialectic that presupposes a literate, codified manipulation of the doctrine. The monk and the monastic founders may either support or oppose the official religious tradition, but the presence of such a tradition is essential as the matrix of all monastic endeavour.

PURPOSES OF MONASTICISM

Discovery of the true self. *Overcoming imperfections.* All monasticism has its mainstay in theologically based convictions that the present state of things leaves much to be desired—that life in society cannot generate the spiritual consummation stipulated by the religious world view that was or is recommended overtly or covertly by its founder. In some traditions, especially in those of South Asian provenance, the true "self" is clogged and covered by imperfections—by sin, ignorance, or other theologically suggested impediments. The ego with which the layman and the seeking neophyte identifies is not his true self—it is the latter that has to be discovered, uncovered. The sundry barriers—differently conceived as matter, individuated mind, or as a soul–mind aggregate defiled by sin, ignorance, and perversion—must be broken through, or a veil lifted, so that the true self, the primordial spirit, may shine forth. In most traditions, this breakthrough is held to be unattainable through the ordinarily acceptable good social life. A new, hardened, and disillusioned approach must be sought. The body and the mind, which are part or the whole of the impediment, have to be controlled, disciplined, and chastised—hence either asceticism or a set of psychophysical experimentations that differ radically from the normal acts of life is espoused.

Spiritual perfection. The quest for spiritual perfection is elitistic, even when, as within Christian monastic orders, humility is essential. Withdrawal from society is necessary, since the instrumentalities of perfection cannot normally be acquired and augmented in the surroundings of everyday life. The operational basis is meditation on a set of spiritual concepts that either represent the supreme value or provide crutches to support the body and the mind on their tedious journey toward whatever their supreme consummation may be. Processes of intense contemplation, often accompanied by physical exercise, constitute the mystical practice—*i.e.*, prayer, worship, incantation, propitiation, and various forms of self-abasement or self-inflation. All these are pursued in enormously variegated forms and degrees.

Forms of mystical practice

Emancipation of the self. *Salvation.* Seen cross-culturally, the ultimate purpose of the monastic endeavour is the attainment of a state free from bondage, both bondage and freedom in such endeavours being defined in theological terms. Most languages of cultures with monastic traditions have special terms to denote bondage and freedom; others use terms of common parlance that are then understood by members of society as referring to theologically adumbrated types of bondage and freedom. Thus, the term salvation in the Christian context means deliverance from the powers of evil besetting the person's body, mind, and soul. Varying concepts of salvation, liberation, and emancipation are generated by, or closely related to, the society's identification of an individual, the extensions of his body, mind, soul, and spirit, and his status within a larger universe. The idea of salvation, like other such concepts, presupposes a specific cosmological view against which to frame the answers to the question—formulated or unformulated—"What is it that is bound and that can, should, or must be freed to achieve the most desirable state within or vis-à-vis the totality of things; *e.g.*, the cosmos, God, and other absolutes?" The question has spatial and temporal parameters; in some of the indigenous Indian religions, salvation can be achieved during one's lifetime, but whether this happens or whether the achievement is delayed into another existence is actually accidental to Indian notions of liberation. In the Judeo-Christian and Muslim world, salvation proper cannot be achieved as long as the body continues to exist. "Salvation" and its semantic equivalents, in other words, refer to the present *or* future in the South Asian religions but refer only to the future in the eastern Mediterranean creeds. The life of the monastic is a full-time seeking of salvation, in contradistinction to that of the "part-time" general believer.

Redemption. Redemption as deliverance from the spiritual effect of past transgressions may or may not be identical with salvation, though the terms are in many cases synonymous. In any event, as part of his monastic ambition, the monk seeks redemption from his sins or he intercedes, takes upon himself and works out the redemption of others. This is accomplished through personal sacrifice or many forms of self-mortification. The processes of mortification augment or stabilize the austerities required of the monastic and are part of the picture in all

monastic traditions. The emphasis on the autocentric or the vicarious aspect of the quest for redemption depends entirely on the doctrinal framework with which the monastic identifies. In either case, the dialectic of autocentric and vicarious emphases cuts both ways—in mortifying his own body and mind for the benefit of others, the monk also helps his own advancement along the spiritual path. When a Jaina monk (a follower of a 6th-century-BC Indian religious reformer, Mahāvīra) volunteers to lie upon a bed infested with vermin that suck his blood, he may do so for a client or a patron with a view to diminishing that person's karmic load (involving the doctrine that every deed, good or bad, receives due retribution), but at the same time he practices the monastic virtues prescribed for him as a monk. When a Franciscan monk (a follower of Francis of Assisi, the 12th-13th century Christian monastic leader) serves the poor and the sick, he also improves his own virtues of service and humility, all of which are instruments for his own redemption.

Social and vicarious acts as means of spiritual advancement

Liberation. When liberation from cycles of birth and death constitutes the foundation of a belief system, as in the basic pattern of *saṃsāra* (*i.e.*, an inevitable metempsychotic chain that can be broken only through supererogatory efforts of the meditational order) in the Indian traditions, the monks are the harbingers of the methods of liberation. In India and Tibet as well as in Southeast Asia, the monk has always been the centre of attention of the religious life far more than in the Christian world, in which he was and is marginal to the main ritualistic and ideological thrust. The importance of the monastic life in a religious system is related to its ideological content. Thus, if the state of life after salvation is transformed, but is yet basically a continuous type with the present life, then the monastic has less importance than he does in belief systems in which salvation means a totally different state, one that cancels finitude and eradicates all traces of separate individual existence.

Limited personal goals—*e.g.*, power or wealth—are aimed at by some Hindu monastics, but these are infrequent and atypical for the total monastic situation.

Social and institutional purposes. *Conquest of the spiritual forces of evil.* Social goals are interwoven with those of salvation in most monastic traditions, and there is a vacillating emphasis on one or the other depending on the founders' interpretation of the theological framework. The first Christian hermits of the Egyptian desert (*c.* AD 250–500), Anthony of Egypt, Paul of Thebes, Pachomius of the Thebaid and others, were referred to as the "Desert Fathers." They presaged later monastic institutions and were the inceptors of cenobitism—literally "lying [*i.e.*, eating, sleeping, and living] together"—which was to be fundamental to all later Christian monastic institutions. Though the early hermits, mostly native Egyptian (Coptic) peasants, were inspired by the example of famous recluses, their rigorous monastic asceticism soon projected an orientation toward communal life that was related to paramilitary models encountered at one time or another in virtually all monastic traditions—the community that was viewed as composed of soldiers of the spirit. The Desert Fathers believed that they were to combat the forces of the devil tempting them in the desert. Much of this might have been antedated by the Jewish Qumrān community, which was located near the Dead Sea and is often identified with the Essenes, a religious group that flourished in the Judaean desert between 150 BC and AD 70. The Qumrān ascetics considered themselves to be the true, unpolluted priests of orthodox Judaism, decrying the Jerusalem priesthood that they characterized as defiled, spurious, and unclean, sullied by Hellenism and potentially heretical. This may have been the first instance of a monastic elite versus an urban sacerdotal establishment in which the interpretation of the canonical teachings was under dispute. Rigorous asceticism, communal prayer, and common work were the rule, though celibacy might not as yet have been part of that community (see also CHRISTIANITY BEFORE THE SCHISM OF 1054; BIBLICAL LITERATURE).

The Qumrān community

Improvement of society. By and large, monastic institutions may have aided the progress of civilization, even though they are often, and perhaps rightly, blamed for obstructing and retarding it.

Monasticism as an instrument for the creation, preservation, and transmission of secular and religious traditions was pervasive in all those cultures that gave a special status to the cenobite institution. Its function as a propagating or proselytizing agent of the religious tradition, however, is by no means universal, nor even regionally determined. In the Mediterranean religions, as well as in those that originated in South Asia, there are discrete, uninterrupted, ordered sequences from the totally contemplative, hence nonproselytizing, to the teaching and preaching orders with virtually no scope for contemplation. The role of monks and their orders in the arts, sciences, letters, as well as in the pedagogical and the therapeutic social services, is thus discussed under the headings of the diverse monastic systems (see below).

Institutional centres for religious leadership. In some religions, monasteries become training centres for institutional religious leaders. There is, however, a clear dichotomy between training centres for ecclesiastical and for monastic leaders. Even though the distinction may seem to be blurred in the Roman Catholic and Eastern Orthodox traditions, the fact that monastic training could also be priestly should be viewed in a cross-cultural survey as accidental. In all Indian religions, by contrast, there is a radical and exclusive division between the priestly and the monastic careers and their concomitant institutions. The common denominator lies in the supererogatory status of the monastic life—if churches and seminaries prepare ecclesiastical leaders, teachers, and intellectuals, monasteries may train people to whom the same epithets apply, but with a difference: the monk is usually said to be more radical and less compromising, modally speaking, than the ecclesiastic.

Priestly and monastic training

Other purposes. Apart from the ubiquitous redemptory, spiritual, and social goals of monastic systems, most of them condone peripheral goals of more or less mundane types. Thus, a Tibetan lamasery (monastic religious centre) is not only a centre of spiritual counsel but also a bank, a judicial court, a school, and a gossip centre for the laity. Some of the most specialized operations are found in monasteries, including coaching in wrestling, in some Hindu orders, and the preparation of perfumes by the Muslim Sanūsīyah (a conservative, rational, and missionary order established in the 19th century).

TYPES OF MONASTICISM

Organizational or institutional types. *Eremitic.* A taxonomy of institutional types includes several varieties: first, historically and in terms of simplicity, is the eremitic (religious recluse) type, including the early Christian hermits or anchorites; the actual or legendary *ṛṣis* ("seers") of Vedic India (pre-800 BC); some of the earliest Jaina *śramaṇas* ("ascetics"), particularly the semihistorical founders of Jainism (Mahāvīra and Pārśvanātha); the Taoist recluses of early southwestern China; and, down even to today, sporadic hermits in the various religious areas of the world—such as Gauribala in Ceylon (now Sri Lanka), La Mêre in Pondicherry, India and other Western converts to Asian belief systems without organized monastic trappings. Some of the European and American neomystics should also be included in this class.

Common to all true hermits and eremitic institutions is an emphasis on living alone, on a highly regularized contemplative life (with individually generated, often experimental spiritual disciplines), and on frequently idiosyncratic and even heretical interpretations of underlying scriptural or disciplinary codes. Incipient self-mortification and individual austerities can be traced, but these are incidental to the eremitic style.

Quasi-eremitic. The lauras (communities of anchorites) of early Christianity in Greece and Cyrenaica (exemplified by the Mt. Athos tradition that exists even today), the small-scale *āśrams* ("religious retreats") of monastic Hinduism from at least 300 BC to this day, and perhaps the semiformal congregations of the early Buddhist monks and nuns, preceding the establishment of

the *saṅgha* ("monastic order," or "community"), are best called "quasi-eremitic." Common elements in this type of monastic assembly are loose organizational structures with no administrative links to mother institutions and no external hierarchies. The category forms a transition between the eremitic and the cenobitic; in many cases, such as that of the medieval Indian *panthā* organizations (sectarian), groups may display eremitic and cenobitic features consecutively, either during different annual seasons or on the occasion of special conventions. In the 20th century, some Nepalese followers of Gorakhnāth (8th century AD) live as recluses during most of the time but form themselves into a quasi-military association on certain occasions, such as during the all-Indian monastic assemblies (*kumbhamelā*) every sixth year at certain centres of pilgrimage. During these periods they become organizationally indistinguishable from the most highly structured cenobitic units present at the conventions.

Cenobitic. It is probably not wrong to identify proper "monasticism" with cenobitism. There seems to be a correlation between a formulated rule, or set of rules (*regula* in the Christian orders, *vinaya* and *śīla* as part of the Buddhist canon), and cenobitic institutions; eremitic and quasi-eremitic settings lack formulated rules and give more scope to the individual's self-imposed disciplines. A Christian ascetic, Pachomius of the Thebaid (*c.* AD 290–346), was reputedly the founder of organized cenobitism in the Western world and is said to have built nine monasteries for men and two for women. These monasteries may have been the model for the monasteries founded by the Greek theologian St. Basil the Great (*c.* 330–379), who set down what was the first monastic rule. The basis for all subsequent Eastern Christian (Greek) monastic institutions, it was simple and primitive compared to the *regulae* ("rules") of the orders founded in later centuries in Europe. Avoiding the extreme austerities of the Desert Fathers, St. Basil's rule was strict but not severe. Its asceticism was an instrument in the consummate service of God; it was to be pursued in community life and in obedience. Liturgical prayer and manual and mental work were obligatory. Germinally, the Rule of St. Basil enjoined or implied chastity and poverty, though these were far less explicit than in the later *regulae*. What Basil's rule was for Eastern monachism, St. Benedict's was for early Western monasticism. Benedict of Nursia (*c.* 480–543/547) was a practical Roman, and his *regulae* (enjoining poverty, chastity, obedience, and stability), which was used—for the most part—by most of the paramilitary aristocratic orders, such as the Knights Templars, until the 13th century, and which is the rule of the Benedictine Order today, incorporates instructions on institutionally held property. It also set the model for all subsequent Roman Catholic monastic orders; *i.e.*, the requirement that the individual monk does not own property, but that it is held by the order through its appointed trustees.

Rules of St. Basil and St. Benedict

One-third of the Theravāda (southern, or so-called Way of the Elders) Buddhist canonical literature is *vinaya* ("comportment"), the Buddha's own statement of more than 200 rules for his monks. Compared to these, the later Brahmanic (Hindu) orders, such as the *Saṃnyāsin* order founded by the Hindu reformer Saṅkarācārya (8th century AD), contain hardly any "rules" except an implicit renunciation of worldly desires, a detachment from society, and an indifference toward the "opposites" such as pleasure and pain. The 6th-century-BC founder of Jainism, Mahāvīra, about 80 years senior to the Buddha, formulated the nucleus of the Jaina doctrine and also established the core Jaina order, giving it a very elaborate rule that goes into minute regulations for monastic residence: restricting the itinerant monk's sojourn to one week at a time in a village and one month in a town; requiring that the monk must not sleep more than three hours and that he must spend the day and the rest of the night in expiation, meditation, studying Jaina scripture, and begging for alms. Some scholars believe that the Jaina rule provided the model for all monastic rules in India and thus for the monastic traditions in all the Asian countries that came under India's religious tutelage.

Indian and other monastic orders and groups

The Essenes of Judaism, regardless of whether or not they were identical with the Qumrān settlement, probably had a written rule; certainly, they were highly formalistic, emphasizing ritualistic purity, with ablutions prescribed for the members, and with a fundamentalistic adherence to the letter of the Jewish ritualistic and legal books Leviticus and the Deuteronomy written into their discipline. In modern times, the so-called hippie communes, insofar as they seek religious experience, should be included in a historical list of cenobitic organizations; growing food, preparing and consuming it jointly, and sharing common dormitory facilities are essential elements of the cenobitic structure.

Quasi-monastic. Paramilitary, or quasi-monastic, associations are another type of monastic group. Most Christian orders of this type also had medical or healing commitments; non-Christian monastic orders of this type did not cater to the sick. The Knights Templars, whose order was founded in the 12th century during the Crusades—the most prestigious and most thoroughly defamed aristocratic organization—styled themselves "poor fellow soldiers of Christ and the Temple of Solomon"; their foundational commitment was the protection and the guidance of pilgrims en route to and in the Holy Land. Though the Templars, following the rules of St. Basil and St. Benedict, took vows of poverty, chastity, and obedience, their corporate wealth and the secrecy of their initiation and of all their internal affairs helped bring about their extermination under the French king Philip IV the Fair in the early 14th century. The military model was evident in their hierarchical structure—there were chaplains, knights, and sergeants under a grand master of the temple.

Christian military orders

The Templars were inspired by the Knights Hospitallers (or Order of the Hospital of St. John of Jerusalem), founded in the 11th century. The Hospitallers were probably the first to generate genuine medical and hospital services, at first for pilgrims to Jerusalem. They were the classical nursing order, and their first foundation was the Hospital Saint-Antoine-de-Viennois (*c.* 1100). There were other houses in Spain, Italy, and Germany. Their chief officers were ordained priests, but the majority of the members were nonsacerdotal "hospitallers," or lay brothers and sisters. They adopted the Benedictine rule until 1231, meeting under an elected master, and an annually rotating chapter-general assembling the "commanders"; the order switched to the "modern" rule of St. Augustine in 1247.

The Teutonic Knights (Deutscher Ritterorden), founded in Jerusalem in 1189, had a feudally independent relationship to Rome and the papal administrative bureaucracy (Curia) specially defined by over 100 papal bulls; their grand master, who had the same rights as a prince of the Holy Roman Empire, was assisted by five "grand commanders." The organization was composed of knights (usually noblemen), middle-class priests, and sisters (mostly noblewomen), well trained in hospital services. After the fall of Acre to the Muslims in 1291, the order moved its headquarters to various places in Europe. But in the 13th century and after, the order revived its erstwhile function, when European rulers called upon it to war against the Altaic and the Prussian pagan peoples.

The popular, and wrong, identification of Tibetan monks as "lamas" has obscured the highly segmented institution of the Tibetan Buddhist clergy. Among the Khamba (*khams pa*) in eastern Tibet, men with minimal monastic initiation (*lung*) organized themselves as a military or police force to protect the unarmed higher initiated clergy as well as the monastic territory. These were not defenders of, or fighters for, Buddhism, but were a type of monastic police force, sartorially indistinguishable from the actual clerics. They were very evident in Tibet's confrontation with the Chinese Communists (1959–65).

Buddhist, Muslim, and Sikh military orders

In the Islāmic world, the mystical orders (Ṣūfī) and the partially overlapping dervish (*darvīsh*) assemblies constituted a living critique, as it were, of the formalistic, fundamentalist, and Qur'ān-oriented (scriptural) orthodoxy controlled by the *'ulamā'* ("teachers") from Cairo,

the assemblage of Muslim learned ones who set standards for orthodox faith and practice. The Ṣūfīs' direct approach to divinity through such meditative or ecstatic practices as the *dhikr* (the chanting of the names of Allāh, or God)—that was accompanied by variegated physical expressions such as dances and songs and by the ingesting of drugs, usually of the genus *Cannabis* (such as hashish)—became symbolic representations of the criticism of officially sanctioned behaviour. The Turkish Bektaşi (Bektashi) excelled in poetry and in humorous repartee, which even now is recalled in references to the art of the Bektaşi. The Sanūsīyah (Senussi) order of Ṣūfīs, in Libya and other northeast African countries, not only antagonized the Wahhābīyah (a generic name for fundamentalistic orthodoxy in Islām rather than a term denoting any specific group) but also achieved impressive political and military corporate stature in very recent times, opposing the Italian colonialist forces in Libya with some measure of success. These orders sought communion with Allāh through mystical practices rather than salvation by righteousness. "Not I and God but only God" was one of their mottoes.

Though the religion of Sikhism is historically Hindu, the early "pure" (khālsā) did not encourage monasticism; Gurū Nānak, the founder of Sikhism (1469–1539), was a married man, and so were most of the other nine Gurūs ("teachers"). In the late 17th century, however, the Nirmal-akhāḍā was created in complete analogy to the celibate monastic orders of Hinduism and organized on the same principles. Underlying this development is a universal Hindu tendency to create monastic corollaries to lay teachings; the process has been repeated in India much more recently. The Arya Samaj (a Hindu reform society), founded by Swami Dayananda Sarasvati in 1875, is a good case in point. Although Sarasvati was a monk in the Daśanāmī *sannyāsin* order (Holy Men of the Ten Names), he discouraged monasticism—yet, bending to an all-Hindu cultural pressure, monks have been ordained in his organization since the early decades of this century. An older quasi-monastic organization among the Sikhs is the Nihaṅg Sāhibs, a basically military organization within Sikhism. Created to fight Muslim incursions into the Sikh communities of the Punjab, the Nihaṅg Sāhibs wear robes that are blue and yellow military uniforms that have been unchanged since the 17th century. The Nihaṅg Sāhibs are married, but during their temporary active service as *nihaṅgs* they abstain from sexual intercourse and live in a cenobitic manner.

Mendicant monks and orders. By a strict definition, mendicancy (living by begging) would preclude cenobitism. In actual fact, however, there are many orders that are mendicant and cenobitic at different times. The Hindu and Buddhist official orders are really both. During his training, the neophyte lives in a strictly cenobitic setting; on subsequent peregrinations, he eats the food he obtains by begging, which is part of his advanced discipline, and he eats alone. The Burmese, Thai, and Sinhalese Buddhist clergy could be termed mendicant stationary—the monks wander out in the early morning to collect food in their alms bowls, but they consume it jointly at the house in a cenobitic fashion.

The Christian mendicant orders

In the Christian world, the Franciscans, founded by St. Francis of Assisi (1181/82–1226), with their numerous subsections, and the Dominicans, founded by Domingo de Guzmán (1170–1221), were and are the most powerful statutory mendicant orders. The synthesis of contemplation and the apostolic ministry is strong in these orders; the Dominican motto "to contemplate and to give the fruits of contemplation to others" is significant. The Sanskrit term *parivrājaka* ("walking around") quite literally connotes mendicant status and as a title is carried by a large number of Hindu monastic organizations. It has canonical sanction—the Hindu scriptural definition of a monk is "[one who] having renounced the desire for sons, for wealth, the fear of social opprobrium and the craving for social approval, he sallies forth, begging for food." Here also there is a blend between the contemplative and the preaching life; the various Hindu orders place varying emphases on the one or the other, a distribution that seems quite parallel to that of the Christian orders. The vow of chastity is spelled out for the Hindu mendicants, but poverty and obedience are implied rather than enjoined. The Hindu monastic organization is much looser than either the Buddhist or the Christian, and in this sense it resembles the earliest eremitic and quasi-eremitic types in Judaism and Christianity.

Other organizational or institutional types. Permanent versus temporary membership correlates with different monastic institutional types, though it does not seem to have any bearing on organizational structure. In the Theravāda Buddhist order (*saṅgha*) of Thailand, Burma, and Sri Lanka, men join a monastery for an unspecified period of time, with minimum periods (three months to a year) prescribed in the *Dhammayut*, the smaller and more highly ascetic of the two sections of the Siamese *saṅgha*; the Mahāssaṅghikas, who form the monastic majority, do not specify any such duration. Lifelong monastic views are, in those regions, a matter of individual choice. The order itself does not take any official stance on them. This differs radically from all full-fledged Catholic orders as well as from those Hindu organizations that initiate members by the *virajā-homa* (*i.e.*, the Vedic rite of renunciation); since the initiate is declared dead by this ceremony, he cannot return to the world of the living (*i.e.*, to society). Dispensations, on the other hand, were given, though reluctantly, to Roman and Greek Christian monks and nuns who wanted to leave their orders. In the Hindu monastic code, there can be no such dispensation—monks who leave and return to social life are highly stigmatized.

Hierarchical and status types. *Sacerdotal.* In addition to organizational and institutional forms, a typological survey must also include aspects of monastic status and hierarchy. The first and most important division here is between sacerdotal and nonsacerdotal full-time supererogatory specialists. Most of the canon-based (scriptural) religions of the world distinguish between priests and nonpriestly practitioners. In the case of Greek and Roman Catholic Christianity, the distinction is crucial on the sacerdotal but incidental on the monastic end. Monks who have priestly ordination are full priests and full monks; monks and nuns who do not have it are full monastic members nevertheless, sharing the same vows and the same discipline. Islām does not officially recognize monastic status, nor does it have priests—the *imām* is the leader in prayer, but he has no special vows or ordinations. The dichotomy does not apply in Judaism either.

Separations of sacerdotal and monastic functions in Hinduism

In the case of the religions originating in India, the situation is markedly different: in Hinduism, only a male person born into a Brahmin (highest) caste is entitled to perform sacerdotal, Vedic (scriptural) ritual; this requires no further initiation than that given to all high-caste boys. A Hindu priest (*purohita*) must be married. Celibacy is not incumbent on all Hindu monastic orders, though it is in those of high prestige. But the monk cannot perform any sacerdotal service, even if he was born into a Brahmin family. The fact of monastic ordination cancels his sacerdotal status. Monastic organizations ordain monks in various ways, and the types of ordination are numerous; but monastic and priestly ordinations are totally different and distinctive in type, scope, and intent. The Brahmin priest supports and enhances the mundane well-being of his client and the worldly estate of his society through Vedic and other rituals. The monk, on the other hand, withdraws from the mundane in a radical sense by rejecting sacerdotal commitments, and he recommends such withdrawal to his clients in a long-range perspective.

Secondary and tertiary orders. The notion of secondary and tertiary orders was generated in the Roman Catholic world, though by analogy it could be extended to non-Christian cultures. The triple division of the Franciscans and the Dominicans epitomizes this hierarchical and status type: the first order consists of ordained priests and lay brothers, the second of contemplative nuns; the third order incorporates laymen and laywomen, "tertiaries," who live under abridged, or "minor," vows

that do not always include celibacy. In the Theravāda Buddhist world, these tertiaries have parallels in the *saṅgha*, which can be viewed as the Buddhist analogue to the first orders of Dominicans and Franciscans. Whereas the full-fledged Buddhist monk has over 200 vows, part-time monks (*śramaṇa*s) have less than one-third that number. In Burma, quasi-monastic, but unordained practitioners (*upāsaka*s) may stay at monasteries and participate in the meditative and congregational activities of the monks, for a limited period and with payment of a nominal fee to the bursar of the cloister.

Status of nuns

In all monastic traditions of the world, the status of nuns is considerably lower than that of monks. The only conceivable exception is that of certain famous saintly women in Hindu India, today and in the past; some of them, known for their extreme piety or, more importantly, for their physical-mental (yogic) and mesmeric (hypnotic) powers, have gained high charismatic (spiritually influential) status that may place them, as individuals, above male monastics. Yet there is no truly hierarchical superiority wherein an ever so exalted nun could have disciplinary powers over a monk or even over a male novice. Though the Roman Catholic tradition has refused equal status to nuns, because women cannot obtain sacerdotal ordination, the Indian attitude rests on notions of ritualistic impurity—women, being polluted through the menstrual cycle, never have access to the ritual complex due to their innate defilement, hence their status is much lower—even though some noncanonical texts (*e.g.*, the *Bhagavadgītā*) assert spiritual, though not ritualistic, equality of women and men.

When women postulants approached the Buddha for admission into the order, he was reluctant; finally, when his disciples and sponsors had succeeded in establishing nunneries, the Buddha said that this step augured the decline of the order. This did not discourage women either then or later. Buddhist nunneries are not numerous, and their ratio to male convents does not exceed 1:20 in any of the Buddhist countries. The modal Buddhist monastic attitude toward the nuns is one of embarrassed silence except in Japan, where the general loosening of monastic rules has worked in their favour.

Tertiary orders

Tertiary orders in the Christian world were established by noblewomen who combined piety with pioneering medical knowledge and strong motivation to gear the latter to religious pursuits approaching the monastic in degree of dedication. Though the term tertiary did not originally contain reference to the sex of the membership, its selective denotation was well established by the 13th century, usually referring to women, often of aristocratic background, who led a saintly life in a cenobitic setting but were inspired by humanitarian ideals rather than by that of sheer contemplation. In a very real sense, women belonging to such groups were the first nurses; their tradition has been continued in all the Christian nursing orders extant today and is being emulated by some new non-Christian orders, such as the Hindu Ramakrishna Mission. Where there are male tertiaries, as in some segments of the Dominican and Carmelite orders, the humanitarian connotation dominates.

Though the hierarchical types in the Christian West must be viewed from an organizational or managerial vantage point, there is much variation in Eastern orders. In the religious world derived from the teachings of India, a true hierarchy comparable to the Christian orders is found only in the Tibetan ecclesiastic setting. Contrary to the popular notion, the lamas are not simply high-ranking monks but are viewed as incarnations of a particular aspect of the Buddha or of a teacher who in turn was such an incarnation. Though Tibetan monasteries prided themselves on the presence of one or more lamas, they really stood above and outside the operational hierarchy, and their function was and is advisory rather than executive.

VARIETIES OF MONASTICISM IN THE RELIGIONS OF THE WORLD

Since monastic systems developed mainly in the Mediterranean monotheistic, and in South Asia's theologically more complex situation, their diffusion into other parts of the Western and Eastern world can generally be viewed as a modification of the two historical core areas that were located in two relatively small regions; *i.e.*, the Judeo-Christian-Islāmic and Hindu-Buddhist-Jaina areas.

Religions of the East. *Hinduism.* With some reservations, the religion of the majority of the population of the Indian subcontinent after the decline and exit of Buddhism can be called "Hinduism," to supplant the erstwhile "Brahmanism" or the Vedic religion of pre-Buddhist days. Buddhism, like Christianity, is an export religion in the sense that it did not survive in an ecclesiastically organized fashion in its homeland. Hinduism—on the other hand—has absorbed so many Buddhist traits that it is virtually impossible to isolate the latter in medieval and later Hinduism. The most important Buddhist-inspired element in Hinduism is no doubt its monastic tradition. Where there were hermitages in ancient, pre-Buddhist India (such as the abodes of the Vedic "seers" [*ṛṣi*s] and the *gurukula* [teacher's family]—which was a germinal cenobitic setting comprising the *ṛṣi*s and their disciples), typically monastic features such as vows of chastity and an unequivocal rule of monastic comportment were not operative before the time of the Buddhist *saṅgha* ("monastic order") in the 6th century BC and its little-known early contemporary movements, such as the Ājīvikas, which are viewed as proto-Jaina, and other incipiently monastic institutions.

The most outstanding Hindu monastic founders and thinkers, comparable in many ways to the Christian St. Benedict of Nursia or the great theologian Thomas Aquinas (1225/26–1274) with regard to their importance in their respective areas, were Śaṅkarā (8th century AD) and Rāmānuja (11th century AD). Both of these teachers propounded the Vedānta theology (a religio-philosophical system concerned with the nature of ultimate reality), albeit in very different, mutually incompatible interpretations. Śaṅkarā's order of Dasanāmī *sannyāsin* (Holy Men of Ten Names) was and has remained the monastic order that set the monastic standards for the rest of Hindu India. Based on a scholastic, nondualistic reading of the four "great dicta" (*mahāvākya*) of the canonical *Upaniṣads* (one of the Hindu scriptures), the monk, following the example given by the founder, meditates constantly on the numerical identity of his individual soul (*Ātman*) with the cosmic soul (Brahman). All his other observances—such as group incantation of canonical liturgy, participation in the monastic assemblies with other orders (*kumbhamelā*) at various places and at astrologically computed time intervals, alms begging, teaching religious topics to the laity, and conducting scriptural discourse with lay and monastic scholars (*śāstrārtha*)—are ancillary to his main purpose; *i.e.*, meditation. He does no social work, and there is nothing to parallel the humanitarian deeds of some other orders in the Indian and most orders in the Judeo-Christian world. He cannot conduct ritual and he has no obligation whatever toward society; but society has its obligation toward him—it feeds and clothes him. For this, he instructs those who wish to be instructed in the methods of meditation leading to emancipation from rebirth, which is believed to be caused by one's good or evil thoughts, words, and deeds in one's previous existence. In a more formal manner, he may or may not initiate lay seekers and monastic postulants (potential monks) into meditation by imparting to them a *mantra*, a sacred secret formula aiding the emancipatory process. Since the monk's initiation entails the symbolic cremation of his body, he is not cremated at his death as is done in the case of lay Hindus, but is interred or immersed in the river.

Śaṅkarā and Rāmānuja

Most of the prestigious Hindu monastic orders follow this pattern, though their disciplinary codes are often radically different. Thus, the followers of Rāmānuja, referred to as Śrīvaiṣṇavas (worshipers of Viṣṇu [Vishnu] and his spouse), are largely lay, high-caste Hindus. The monastic order relating to this tradition emphasizes ritual and worship of the personally conceived deity; its rules of celibacy, compared with the strict and unexceptional

rules in the Daśanāmī *sannyāsin* order, are somewhat vague and flexible—so that, in theory at least, a person who claims the title of a monk could be a married man.

Celibacy in Hindu monasticism

Of the approximately 90 monastic orders in Hinduism, some 70 impose celibacy and a cenobitic rule on their ordained members. Others—such as the Dādū-panthīs (created by Dādū, an important Indian saint of the 16th century) and a number of other orders whose designation ends in *-panthīs* ("path-goers") of relatively recent origin (14th century and later)—follow specific theistic doctrines of medieval Hinduism. Unlike the Daśanāmī, who accept only Brahmins (highest caste Hindus) into their order, the *panthīs* do not discriminate on grounds of caste in their recruitment. In fact, most of these orders can be viewed as anti-Brahmanic revival, or even rebellion, movements.

Jainism. Jainism, founded in the 6th century BC by Mahāvīra in reaction to Brahmanic Hinduism, has fewer than 12,000,000 followers today; but due to its mercantile predilection, it is a wealthy community that has given traditional and substantial support to its monastic organizations. The two main lay divisions of the Jainas derive their primary designation from the monastic setting, which is unique, in India and the West. The Śvetāmbara ("white-clad") sect is so called because its monks wear a white robe and a white piece of cloth to cover the mouth, thereby preventing the inhaling and annihilation of microbes and insects, and carry a broom with which they sweep the ground in front of them as they walk so as to clear away insects and other living beings that would be hurt or killed by being stepped on. The Digambara ("heaven-clad"; *i.e.*, nude) sect is so called because its monks used to go naked, to signify their complete detachment from worldly things and social trappings. Under Muslim rule, this custom was interdicted and since then (15th century) the Digambara monks have been wearing the white robe, thus becoming sartorially indistinguishable from the Śvetāmbaras. The Jaina monk stresses mendicancy, extreme austerity, and detachment. Both Jainism and Buddhism were purely monastic religions—in fact the only ones in a cross-cultural perspective; the rules for the laity are derived from monastic rules. The founders of Jainism (Mahāvīra, the semi-legendary Pārśvanātha), as well as the Buddha, directed their instructions to monks and postulants exclusively—the vows of the monks are more numerous and more intensive, but the way of life enjoined on the laity is simply an abridged monastic rule with more dispensations and compromise.

The *saṅgha* of Buddhism

Buddhism. The generic term for the Buddhist monastic order is the *saṅgha* ("community"); the terms connoting the order in all Buddhist countries are literal translations of the Indian word. Far more than in other monastic traditions of the world—with the possible exception of Jainism—the Buddhist doctrine attaches central importance to the order. The recitation of the "threefold refuge" formula that makes a person a Buddhist, either lay or monastic, imparts a pledge of allegiance to the Buddha, the *dharma* ("teaching"), and the *saṅgha*; most commentaries imply that each of these three constituents is equally important. In northern Buddhism (*i.e.*, Mahāyāna, or the Greater Vehicle) the role of the historical Buddha has been and is negligible, and the order has acquired an even more exalted position. Since the Buddha began every single one of his sermons with the address *bhikkhave* ("O ye begging monks"), fundamentalistic Buddhism is tantamount to monastic Buddhism. The monastic discipline of the Buddhist clergy varies widely in the different parts of the Buddhist world. Ideally, the rules are laid down in the *vinaya* ("monastic rules") portion of the Buddha's sermons, but monastic traditions and regulations have been influenced by the ecological and cultural conditions of Buddhist areas. Rules of vicinity or distance from lay settlements had to be differently interpreted and implemented depending on whether tropical, moderate, or (as in the case of Tibet and Outer Mongolia) sub-Arctic climatic situations are involved. Although celibacy is postulated for the Buddhist clergy everywhere, there have been and are liberal exceptions. The married monks of pre-20th-century Ceylon and those of some of the Japanese Buddhist clergies were highly evident exceptions. Since the vows of the Buddhist monk in principle are not permanent, the theoretical emphasis on celibacy became academic in many parts of Asia. In the Theravāda (Way of the Elders) countries of South and Southeast Asia, the monks were and still are teachers to the people—not only of religious matters, but of basic education—particularly in Burma. There appears to be a correlation between the regional stress on educational and other forms of monastic involvement with the lay society and the provision of special amenities for monks who prefer a strictly contemplative life, as in Sri Lanka and Thailand; conversely, where the involvement is not spelled out, as in Japan, no such special division is apparent. The differences in living style between the northern (Mahāyāna, or the Greater Vehicle) and the southern (Theravāda, wrongly called "Hīnayāna," or the Lesser Vehicle) monastic institutions are quite radical. The fundamental feature, however, is meditation (Sanskrit *dhyāna*, Pāli *jhāna*, from which is derived the Chinese Ch'an and Japanese Zen). The path of meditation leads positively toward the intuitive understanding of momentariness—the condition of existence—or negatively toward the total rejection of all notions of permanence. The negation of existence is described by the term Nirvāṇa, "fading away"—Pāli Nibbāna; Tibetan Thar Pa, or "crossing over"—which is identical with the Zen term Satori, a translation of the Sanskrit *bodhi*, "enlightenment," which was the discovery of the Buddha. All these terms are operational synonyms.

Functions of Buddhist monks

Sikhism. Sikhism, founded by the Punjabi reformer Nānak, was and is a martial version of Hinduism. Of all the indigenous Indian religions, it was the least sympathetic to monastic inspirations. The aforementioned Sikh monastic Nirmal-akhāḍā and the quasi-monastic Nihaṅg are patently compromises with the overall Indian trend to establish monastic traditions to express full-time involvement with redemptive practice. During the last two centuries only, the monastic Udāsīn order (founded by Nānak's elder son Śri Chand) has achieved a most successful rapprochement with Hindu elements. Its disciplinary, sartorial, cenobitic settings are identical with those of the Hindu *sannyāsin* ("holy man"). They refer to the *Granth Sahib*, the sacred book of the Sikhs, as their basic text, in spite of the fact that their intramonastic and intermonastic discourse proceeds on Sanskritic, Vedāntic (religio-philosophical) lines similar to those of the orthodox Hindu orders. This accounts for the fact that the Udāsīns are now respected as equal to the most prestigious and ancient Hindu orders.

Taoism. Taoism, a Chinese religion based on folk religion with Buddhist influences, with some very weak emulation in Japan, holds a middling position with respect to monastic ventures, lying somewhere between the powerfully antimonastic Confucian schools that always represented the official culture and mainstream of sophisticated Chinese opinion and the radically monastic Buddhists. Some scholars believe that since Taoism originated in the southwestern parts of China, Indian influences are conceivable. The chief object of Taoism, however, is not the redemptive, salvational purpose evinced in other scripturally based religions, Eastern and Western. The ultimate aim of the Taoist sage was longevity or ultimate physical immortality rather than salvation from this world or the self. The Taoist quest after the elixir of life, and its expression in cryptic and enigmatic poetry that is well known to, and generally misunderstood by, European and American seekers of the mysterious East in the 20th century, is in no way comparable to the supererogatory search of the monastics thus far discussed. The Taoist settlements of sages, in forests and mountain glades as well as in the cities, are, at best, analogous to the eremitic type of proto-monasticism. When Taoist settlements were cenobitic or celibate, these features were indeed incidental to Taoism, which defies and rejects rules of any corporate kind.

Aim of the Taoist sage

Other Eastern varieties. Monastic orders are thus either of Mediterranean or of Indian origin. The connections between the monasticisms of the two regions are

immensely complicated, since the Indian subcontinent has absorbed or generated traditions that, though abhorrent to the orthodox Islām of the Arabic *'ulamā'* are nevertheless Muslim rather than Hindu in origin. North Indian Hindus have felt a strong attraction to Ṣūfī (Islāmic mystical) poetry, and dozens of Ṣūfī saints are worshipped by Muslims and Hindus alike. With India as one of the centres of monastic diffusion, new quasi-monastic types can be expected to spring up in the present or near future. In a strictly descriptive account, leaving aside historical development, the Ṣūfī orders must be seen as separate from both the Indian and the Mediterranean backgrounds.

South Asian quasi-monastic groups

Of the slightly less than 100 monastic and quasi-monastic orders in South Asia, well over half are local and regional growths. They usually lack even a body of rules and conventions that would be recognized or accepted by a wider Hindu-Buddhist-Jaina consensus. About a dozen orders are rejected and feared as heretical, outside the pale of the acceptable, and charged with using religious pretexts to indulge in antisocial behaviour. The Hindu and Buddhist Tantric sects (practicing occult, esoteric, meditative techniques) represent esoteric countermonasticism in India, though they have been accepted fully in certain Tibetan Buddhist hierarchies.

Of the not numerous but clearly evident monastic or quasimonastic organizations of recent origin in other parts of Asia, the Vietnamese Cao Dai achieved some recognition. They had a military organization and their own army "regulars" up to the mid-1950s; an eclectic theology with such heterogeneous patron saints as the 19th-century French novelist Victor Hugo, the World War II British prime minister Winston Churchill, and the Buddha; and a monastery-fortress in south central Vietnam. The members were bound by vows that did not include celibacy or poverty but stressed obedience to the hierarchy.

Religions of the West. *Judaism.* Judaism and Islām, the two Western religions that assume a father-oriented, social hierarchy as the supreme secular value, did not generate monastic traditions as part of their official institutions. The Essenes of the Qumrān community were, in their own vision, inimical to the ecclesiastic centre and marginal to the official Judaic complex. The weak eschatology (doctrine of the last times) of Jewish theology might account for the lack of a lasting monastic quest, which is typically inspired by individual salvational expectations.

Islām. In a somewhat milder fashion, this holds for Islāmic orders as well. Whereas some Jewish prophets might have had celibate leanings, the Prophet Muḥammad discouraged celibacy. Non-Arabic Islām, nevertheless, generated monastic orders, "Ṣūfīsm" being the generic name for a monastic attitude rather than for monastic institutions. The Bektaşi (see above) and the Sanūsīyah (conservative order founded in the 19th century) are typical for the marginal status of monastic settings in Islām; vestigial rules and formalized vows are discernible, but the main thrust of these monastics was of an interpersonal type, centring on the relation between the individual teacher of esoteric wisdom (*murshid*) and his disciple (*murīd*) and on the practices of chanting and meditation on the secret or known names of God (*dhikr*) and of other ecstasy-producing methods. The "way" (*ṭarīqah*) meant something that, by implication, was not accessible to the pious, orthodox Muslim alone. The Naqshbandīyah order, which originated in Turkic-speaking areas of southwestern Central Asia, became widespread in the Islāmic Middle Ages, rediffusing from India into the western reaches of the Ottoman Empire (14th–20th centuries). The actual or alleged ingestion of *Cannabis* drugs (such as hashish) and the nonconformist, antilegalistic doctrines of the order make it attractive on the popular level. The ritualization of the esoteric, contrasted with that of the social and the civil in the official Muslim prayer, seemed to provide an outlet and an alternative for a large number of devout but nonconformist Muslims, much as the modern cultic movements (such as spiritualist, "hippie," and similar groups) do for the religiously alienated in the West. Nonconformity to the official doctrine was often enhanced by unexpected or deviant behaviour; thus, the Sanūsīyah brethren prepared and used a variety of perfumes for their personal toilets. The element of rebellion, frequently manifested in eccentric behaviour, is typical for a setting where the official religion is antimonastic, as is the case of Islām.

Ṣūfī orders

Christianity. Since the terms monk and monastic are historically and etymologically Christian terms, the ideal-type monasticism is the Christian type by a semantic fiat —thus, the denotation of monastic terms has been extended to religions other than Christianity, the lexical donor for monastic terminology.

An overall view of Christian monastic history reveals a shift of emphasis from the contemplative to the socially active; this shift, if observed in a geographical dimension, moves roughly from East to West. The Eastern Orthodox and other Greek-liturgy-based sections of Christianity generated highly meditative orders, the Mt. Athos (Greece) complex being the most famous among them. The large variety of Roman Catholic orders displays eclectic emphases: the Benedictines, Franciscans, and all the orders designated as "minor" (in the Latin sense of humble or modest, rather than hierarchical or organizational) emphasize meditation; the Dominicans should be called "major"—though they are not—because the areas of preaching, scholastic continuity, and evangelizing rank far above contemplation in their order; and the Society of Jesus, or Jesuits (founded 1534), stands on the other end of the contemplative–social-centred continuum—teaching, social work, and the active life being regarded as the quintessence of supererogatory piety that is so important to this powerful organization.

Viewed from the social-centred end of the continuum, certain institutions in the Protestant tradition should be called monastic. The Taizé (France) communities of the Reformed Protestant tradition, founded in the 1940s, initiated an ecumenical monastic movement. Where emphasis is on faith rather than on works, monasticism would seem to be a viable expression of the Protestant tradition —yet, due to a set of historical accidents whose ideological summation was described in *The Protestant Ethic and the Spirit of Capitalism* by the German sociologist Max Weber (1864–1920), the operational emphasis has been on active engagement in the world rather than on seclusion. This explains why the various kinds of part-time Protestant retreats (linked to places especially designated for recuperation from the Christian involvement in work) are often placed in rural settings, and where they are urban, there is usually an apology at hand. Professional Western monasticism, by contrast, never made much of the rural–urban division, with important monastic centres randomly built in rural and urban settings; it is only the eremitic and some quasi-monastic types that insisted on rural or sylvan environments.

Ecumenical monastic groups and retreats

MONASTICISM TODAY

A discernible trend can be noticed in modern Western monasticism. It appears to be directed towards an enhanced individual liberalism, covertly or overtly condoned and granted to monks who wish to reinterpret fundamental traditions in the light of individualistic experience. If the work of Teilhard de Chardin, the Jesuit paleontologist, and of the Trappist monk Thomas Merton are in any way typical of this trend, it may be a clue for the future of Christian monasticism. Even though the disciplinary bases of the standard religious orders are likely to remain conservative for a very long time, there is much more ideological permissiveness. Individual monks and their disciples have been interpreting the monastic ideal in psychological, anthropological, phenomenological, and existentialist terms. It is perhaps the Christian monks of the future rather than nonmonastic priests and ministers who will seek a trans-Christian ecumenism through religious experimentation. Of significance is the fact that Thomas Merton was killed in an accident while in Bangkok to visit the Dalai Lama (leader of Tibetan Buddhists). Monks had been the religious entrepreneurs of early Christianity—they may well emerge as innova-

tors in an era of infinitely greater inter-religious information and diffusion.

Sponsored largely by college-campus religious organizations (both Protestant and Catholic), a retreat pattern based on ecumenical concerns has been emerging in North America and in Great Britain during the second half of the 20th century, though much less so in other parts of the Western world. There were roughly 1,000 such para-academic, campus-sponsored organizational retreats in North America alone in 1968. Most were Protestant, but with an increasing participation by Catholic and Jewish groups; almost all the participants were of college age. No vows of any sort are enforced or even recommended, but there is emphasis on meditation, discussion, and religious reflection. The projection of ideas about the future of monasticism is easier to discuss than that of the future of religion in general. Since the end of World War II (1945) there has been a highly visible resurgence of monastic ideals and an overall increase of monastic organizations in all parts of the world outside the Communist-dominated countries. The models were admittedly Christian, particularly Jesuit in the case of neo-Hindu orders—such as the Ramakrishna Mission (an eclectic Hindu order founded in the 19th century), which has over a dozen well-established centres in America and Europe. One Hindu monk, a swami—which correctly means an ordained Hindu monk—presides over each of these centres, often assisted by a younger monk; in theory, these centres train monks in the *sannyāsin* tradition, but they actually serve a European and American laity committed in various degrees to the Vedānta theology. On the model of the Ramakrishna Mission, there are some two dozen other India-originated organizations of this quasi-monastic or semimonastic type that have spread over the Western world, some of them with considerable wealth and popularity. Among such groups are the Self-Realization Fellowship, founded by the late Swami Yogananda Paramahamsa, with its headquarters in San Diego, California; and the Hare Ram Hare Krishna with chapters all over North America and western Europe. Founded by Swami A.C. Bhaktivedanta, it attracts and assembles young men and women who temporarily dress in the monastic ochre robe of Hindu monks and chant the names of the Lord as incarnated in Rāma and Kṛṣṇa (Krishna), accompanied by simple, arcane dance movements. Zen Buddhism has been established in North America and Europe, with Zen churches and centres in many major cities. Again, it is usually only the leader and the teacher who is a Japanese Buddhist monk, instructing a variously interested laity.

Infusion of Oriental groups in the West

Since the 1960s much attention has been given to the monastic potential of the communes, the "hippie" and kindred congregations in various parts of the Western world, some of whom have already established settings that have many genuine monastic traits—though with none of the foundational vows of traditional monasticism, Eastern or Western. The growing concern with mystical experience since the mid-20th century, with or without the use of psychedelic ("mind-expanding") drugs, and the general alienation from the official religions parallel many, if not most, of the monastic creations of the last 2,000 years or more.

In the countries of non-Christian monastic traditions (*i.e.*, India, Sri Lanka, Burma, Thailand, Laos, Vietnam, and Japan), monasticism is, in varying degrees, on the decline. If the monastic distribution of the future could be assessed, an intelligent appraisal might be that monastic commitments and organizations will increase in the West and decline in Asia, perhaps in direct proportion to the changing religious concerns of the societies involved —the West, often believed to have had a surfeit of material progress, might seek the spiritual; the East, tired of contemplation, might turn to the pursuit of material progress on the former Western model.

BIBLIOGRAPHY. AGEHANANDA BHARATI, *The Ochre Robe* (1962), and *The Tantric Tradition* (1966), partly autobiographical analyses of the official and the esoteric monastic traditions in Hinduism; OWEN CHADWICK, *Western Asceticism* (1958), a good survey, especially of the Roman Catholic orders; E.C. BUTLER, *Benedictine Monachism*, 2nd ed. (1962), a classic study on this subject; SUKUMAR DUTT, *Buddhist Monks and Monasteries of India: Their History and Their Contribution to Indian Culture* (1963), a classic work relating the monastic to the other cultural tradition of India; E.E. EVANS-PRITCHARD, *The Sanusi of Cyrenaica* (1949), an important anthropological study of a regional Ṣūfī tradition; J.N. FARQUHAR, *The Fighting Ascetics of India* (1925), a classic by default, as no other work has been written since dealing exclusively with the military orders of Hindu India; G.S. GHURYE, *Indian Sadhus* (1964), the only English language survey of the total contemporary monastic situation in India; P.S. JAINI, "Śramanas: Their Conflict with Brahmanical Society," in J.W. ELDER (ed.), *Chapters in Indian Civilization* (1970), an excellent short survey of the Jaina monastic tradition juxtaposed with the Hindu and Buddhist orders; DAVID KNOWLES, *Christian Monasticism* (1969), a good survey of the history of monasticism and religious orders; NORBERT MCMAHON, *The Story of the Hospitallers of St. John of God* (1959), a history of monastic knightdom in the Crusade and post-Crusade eras; R.J. MILLER, *Monasteries and Culture Change in Inner Mongolia* (1959), an anthropological account of Buddhist monachism in Mongolia (the only study in English to date); J.A.N. MULDER, *Monks, Merit and Motivation* (1969), a report on monastic behaviour in contemporary South and Southeast Asian Buddhism; WALTER NIGG, *Warriors of God: The Great Religious Orders and Their Founders* (1959), an excellent account of the inceptors of monastic traditions, with special reference to the paramilitary trends in early monastic attitudes; J.C. OMAN, *Mystics, Ascetics, and the Saints of India* (1903), a classic work that is a fair account of the monastic situation in both ancient and contemporary India; *Palladius: The Lausiac History*, trans. and annot. by R.T. MEYER (1965), a translation of one of the original classics on early Christian monasticism; J. PRIP-MØLLER, *Chinese Buddhist Monasteries*, 2nd ed. (1967), a good account of the monastic ecology and discipline of the Chinese orders; D.T. SUZUKI, *The Training of the Zen Buddhist Monk* (1934), a classic by a famous Zen scholar; L.A. WADDELL, *The Buddhism of Tibet*, 2nd ed. (1934), the classic by way of an overall view of Lamaism; J.B. WILLIAMSON, *The History of the Temple* (1924), an erudite historical account of the Templars; A.W. WISHART, *A Short History of Monks and Monasteries* (1902), a classical introduction to Christian monachism; NUR YALMAN, "Islamic Reform and the Mystic Tradition in Eastern Turkey," *Archives of European Sociology*, 10:41–60 (1969), an excellent sociological account of Ṣūfism in action; ERIK ZURCHER, *The Buddhist Conquest of China*, 2 vol. (1959), an excellent, learned study of Chinese monastic and lay Buddhism and its conflicts with the official Chinese culture.

(A.Bh.)

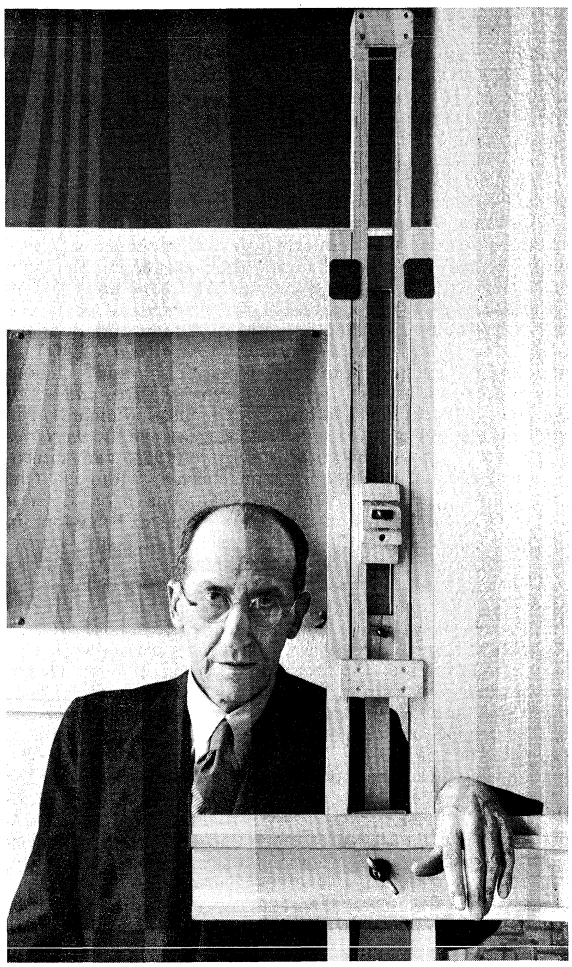
Mondrian, Piet

Piet Mondrian, Dutch painter and leading member of the abstract art movement de Stijl, developed a distinctive style of nonobjective painting based on the reduction of pictorial elements to vertical and horizontal lines, using the three primary colours and noncolours. His work has exerted a powerful influence on 20th-century art, including architecture, advertising art, and typography.

He was born Pieter Cornelis Mondriaan on March 7, 1872, in Amersfoort, The Netherlands, the second child of Pieter Cornelis Mondriaan, Sr., who was headmaster of a Calvinist primary school there. Mondrian (who dropped the second *a* in his name when he moved to France in 1912) grew up in a stable yet creative environment; his father was part of the Protestant orthodox circle that formed around the conservative Calvinist politician Abraham Kuyper, and his uncle Frits Mondriaan belonged to The Hague school of landscape painters. Both he and Mondrian's father, an amateur draftsman, gave him guidance at the age of 14, when he began to study drawing on his own.

Mondrian was determined to become a painter, but at the insistence of his family he first obtained a degree in education; by 1892 he was qualified to teach drawing in secondary schools. That same year, instead of looking for a teaching position, he took painting lessons from a painter in a small town not far from Winterswijk, where his family now resided, and moved to Amsterdam to register at the Rijksacademie. He became a member of the art society Kunstliefde in Utrecht, where his first paintings were exhibited in 1893, and in 1894 joined the two local

Early training



Mondrian, photograph by Arnold Newman, 1942.
© Arnold Newman

artist societies in Amsterdam, Arti et Amicitiae and St. Lucas; during this period he still attended the evening courses at the academy for drawing, impressing his professors with his self-discipline and effort. In 1897 he exhibited a second time, at Arti et Amicitiae.

Up to the turn of the century, his paintings followed the prevailing trend of art in The Netherlands: landscape and still lifes painted in subdued hues with picturesque lighting effects, and subjects chosen from the meadows and polders around Amsterdam. In 1899 he was commissioned to paint a ceiling in a large private house in Amsterdam, where, in homage to the youngest son, he painted an allegory of the seasons, somewhat reminiscent of the Dutch painter Matthijs Maris, the uncle of his close friend Simon Maris. In 1901, he competed for the Prix de Rome, was successful in the initial contest, but did not reach the final competition. During the same year, he made a short trip to Spain with Simon Maris, but the country left no lasting impression on him. Mondrian achieved artistic reputation and a certain notoriety in 1903, when he won the Willink van Collem Prize of Arti et Amicitiae for a still life. That same year he visited a friend in Brabant; the calm beauty of the landscape he discovered in the province proved to be an important influence on him. He stayed on in Brabant, and the following year became a period of self-discovery and intense concentration on his artistic gifts: and by the time he returned to Amsterdam in 1905, his art had visibly changed. The landscapes he painted of the surroundings of Amsterdam, mainly on the River Gein, show a pronounced rhythmic framework and lean more toward compositional structure than toward the usual picturesque values of light and shade; windmills, farmhouses, and series of trees were his preferred subjects at this stage of his art. But his painting still remained within the traditional boundaries of contemporary Dutch art.

Influence of Postimpressionists and Luminists. In 1907 the Quadrennial Exhibition was held in Amsterdam, featuring such painters as Kees van Dongen, Otto van Rees, and Jan Sluijters, Postimpressionists who used pure colours boldly. Their preference for colour must have impressed Mondrian considerably, for in the same year he painted "The Red Cloud," a rapidly executed sketch dominated by colour, which initiated a series of works in which Mondrian experimented with his newly discovered values of colourism. "The Red Cloud" was done at Oele, where Mondrian stayed at the invitation of a friend. There, as well, Mondrian's monumental work "Woods Near Oele" was conceived, the final version being executed in his studio in 1908. In this painting new values appear: a new linear movement, somewhat reminiscent of the Norwegian painter Edvard Munch, and a new colour scheme, based on hues of yellow, orange, blue, violet, and red, suggestive of Expressionist paintings. With this vigorous painting of considerable size (50⅜ × 62¼ inches), Mondrian broke away from the national tradition of Dutch painting and took his place in the ranks of European art.

In 1908 his new style was reinforced by his acquaintance with the Dutch artist Jan Toorop, whom he met during a summer trip to Domburg, a seaside resort where artists and art patrons spent their holidays. Toorop, at that time, led the Luminist movement in The Netherlands, which endeavoured to render light by a series of dots or short lines of primary colours (*e.g.*, blue and yellow), not completely mixed, so that the secondary colour (green) is formed in the viewer's eye as he stands away from the picture. This technique, known as Divisionism, produces brighter secondary colours than if they were mixed on the palette. In these years Mondrian concentrated on this use of colour and limited his palette to the primary hues: a painting such as "Windmill in Sunlight," executed mainly in yellow, red, and blue, proved his mastery in this evocation of strong, radiant sunshine. But in a painting such as "The Red Tree," also dated from the same year, 1908, Mondrian expressed his own vision of nature by creating a balance between the contrasting hues of red and blue, between the violent movement of the tree and the blue sky, thus producing a sense of equilibrium, which is his prevailing aim in his representation of nature. In 1909 Mondrian's Luminist works were exhibited in a large show at the Stedelijk Museum in Amsterdam, together with works of his more prominent contemporaries. Thus, Mondrian firmly established himself among Dutch avant-garde painters.

Luminist period

The year 1909 was important for Mondrian's career from another point of view: in May he joined the Theosophical Society, which espoused the mystical belief that all that exists in the universe emanates from one source, that spirit and matter are inseparably united. In the following years Mondrian's new philosophy was reflected in his painting: he concentrated on large forms in nature, such as the lighthouse in Westcapelle, which he had painted before in a Divisionist technique, but which he rendered now—in 1909–10—with a renewed interest in form. Subjects such as the church at Domburg and an impressive Zeeland windmill express a similar striving for monumental forms; "Evolution," his triptych of three standing human figures, painted in 1910–11, shows a surprising similarity between the human figure and architectural subjects, thus stressing the symbolic identity between man and the inanimate structures he creates. His change toward monumental forms is apparent in his paintings of dunes, first painted in 1909 in a Divisionist technique; but more versions of the same motif in 1910 show the concentration on large planes of colour, which were to be a characteristic of his later work. In 1910 Mondrian's Luminist works attracted considerable attention at the St. Lucas Exhibition in Amsterdam; and in 1911 he submitted one of the dune paintings, in which he expressed the infinitude of nature, to the Salon des Indépendants in Paris, his first bid for international recognition.

Cubist period in Paris. That same year, in Amsterdam, Mondrian saw for the first time the early Cubist works of Picasso and Braque. Mondrian was profoundly im-

pressed, so much so that early in 1912 he moved to Paris, where he took lodgings in the Montparnasse district; from this time, in order to make it easier to pronounce his name in French, he spelled his last name with only one *a*. Almost immediately he began to adapt the precepts of Cubism to his own use, and two versions of "Still Life with Gingerpot," done during the winter months of 1911–12, make this apparent. In the first version, the objects are rendered as recognizable forms from everyday life, a table, cloth, jars, etc., united into a rigid composition around the central object, a ginger pot; in the second version the same objects have lost their daily-life identity and have been transformed into compositional structures: what was a knife in the first version is now a diagonal accent in the pictorial score.

The second version of his "Still Life with Gingerpot" stands at the beginning of Mondrian's Cubist period, which lasted from 1912 to 1917. His Cubist development is characterized by an increasing interest in the painting's composition and by the constant subordination of the factual reality of objects to their role in the structural entity of the picture. His compositions of trees, of facades of Parisian houses, and of scaffoldings during this period are proof of this urge to reduce individual forms to a general formula; Mondrian went beyond the trend of analytical Cubism and its tendency to break forms up into their component parts by striving for a vision of reality that surpasses the individual object. His Cubist style, from 1913 onward, showed the first steps toward abstraction—abstraction being understood as the attempt to strip objects of their factual connotations. On the other hand, Mondrian still kept within the boundaries of Cubism by limiting his colours to ochre, brown, and gray so as not to distract from form and by observing the Cubist scheme of composition, in which the painting gravitates toward a central focus, thus stressing the central part of the composition and leaving the corners almost untouched; the result of this scheme is his oval compositions from 1913 to 1917. In an attempt to reduce the elements of his composition even further, Mondrian avoided curved lines and diagonal accents and increasingly used only vertical and horizontal lines. The course of his development can be followed in his sketchbooks (edited by R.P. Welsh and J.M. Joosten, Amsterdam, 1969) in the transformation of the theme of trees and of the Domburg church into Cubist compositions.

First steps toward abstraction

In the summer of 1914 Mondrian returned to The Netherlands to visit his father, who was seriously ill; much to his dismay, the outbreak of World War I prevented him from returning to Paris. In 1915 he settled at Laren, where he became acquainted with Dr. M.H.J. Schoenmaekers, a Theosophical philosopher whose works on the symbolical meaning of lines and on the mathematical construction of the universe had a decisive influence on his vision of reality. In Laren he made friends as well with S.B. Slijper, who supported him both as a patron and a friend, and with the painter Bart van der Leck, who had developed a monumental style derived from the principles of mural painting. Through van der Leck he met Theo van Doesburg, one of the avant-garde artists of this time. During these years he developed his Cubist style to its extreme limits, mainly in his painting of the church at Domburg and in a new theme, "Pier and Ocean." The ultimate version of this theme, completed in 1917 and shown at the Rijksmuseum Kröller-Müller, marks the final stage of his Cubist style: a composition of vertical and horizontal line fragments in black on a white background, concentrated into an oval shape.

De Stijl. After this painting a new and fundamental change occurred in Mondrian's work: the reduction of existing forms into a pictorial formula is now replaced by a complete rejection of external subject matter. This step was taken not only by Mondrian but by the four painters who together founded the art periodical and the movement of de Stijl ("the Style"): Theo van Doesburg, the initiator of the group, Mondrian, van der Leck, and Vilmos Huszar. The principles of the group were the complete rejection of visually perceived reality as subject matter, and the restriction of the pictorial language to its most basic elements—the straight line and the right angle (hence the use of the vertical and the horizontal line), the three primary colours—yellow, red, and blue—and the three primary noncolours—white, gray, and black. The scope of this new style, for which Mondrian coined the term neoplasticism, was to free the work of art from the momentary visual perception and from the personal temperament of the artist. The objectivity that Mondrian had sought for so long now seemed to be within reach: he could now render "a true vision of reality" in his painting.

Neoplasticism

Mondrian's first neoplastic paintings, which did not derive their composition from a fragment of reality but from an overall view of the harmony of the universe, date from the summer months of 1917: rectangles in soft hues of primary colours painted on a white background. His compositions, in opposition to his Cubist style, were now based on colour and appear to expand over the borders of the canvas into space beyond the picture. In 1918 he reintroduced lines into his painting, linking the colour planes to one another and to the background by a series of black vertical and horizontal strips, thus creating rectangles of colour or of noncolour; the "Composition: Colour Planes with Gray Contours" of 1918 initiates this phase. In 1918 and 1919 he executed a series of rhomboid compositions, subdivided into a pattern of regular squares differentiated by thick black lines and by soft hues of ochre, gray, and rose. In 1919 as well, he created two versions of a checkerboard composition, one in dark and one in light colours, in which the difference of the hues transforms this common pattern into a rhythmic sequence of squares. But only in 1920 and the following years did his neoplastic style reach its full maturity.

This definitive phase in his search for an objective vision of reality was reached in Paris; Mondrian returned there in 1919. In spite of his move, he retained his close collaboration with de Stijl. Mondrian had already published "The New Plastic Art in Painting," in the first issue of *De Stijl* in 1917, and was one of the signers in 1918 of the first manifesto of de Stijl, which stressed a universal, rather than an individualistic, view of the world. And in 1920, coinciding with his mature style, his important essay "Le Néo-plasticisme," published in France, further articulated his theories on abstract art previously published in *De Stijl*.

Mature style. In 1921 his mature style emerged clearly: Mondrian now limited his palette to the primary colours of yellow, red, and blue. In 1922 an exhibition of his works, organized at the Stedelijk Museum in Amsterdam by friends on the occasion of his 50th birthday, was a retrospective progression of his paintings; it traced the path from his beginnings in the Dutch traditional style to his abstract paintings depicting a universal harmony through the use of horizontal and vertical lines and primary colours and noncolours. He had now reached his goal: the creation of what he considered the true image of reality. But he did not stand still: from 1922 to 1938 he concentrated on achieving an ever increasing clarity and purity in his paintings. In 1925 he broke with de Stijl on account of Theo van Doesburg's use of diagonal elements in his work: he considered this an aberration from the strict universality of the neoplastic principles. As an answer to van Doesburg, he resumed his rhomboid compositions without the use of diagonals: the vertical-horizontal opposition was maintained, and only cut by the diagonals of the frame, which presented the rhomboid theme.

Stedelijk retrospective expedition

Mondrian did not exhibit frequently: from 1922 to his arrival in the United States, he never held a one-man show. In 1930 he exhibited with the Cercle et Carré group, and in 1931 with Abstraction-Création, the group that stood next to Cercle et Carré. In the years after 1933 his works became more austere and severe. During this period, in 1934, he met the young American artist Harry Holtzman and the English painter Ben Nicholson, both of whom were soon to become helpful to him. Nicholson urged him to publish his essay "Plastic Art and Pure Plastic Art" in the international publication *Circle*, of

which he was coeditor; and when Mondrian decided to leave Paris in 1938, under the shadow of the invasion of Czechoslovakia by Hitler, he was welcomed in London by members of the *Circle* group. For two years he worked and lived in a London suburb, but the bombardment of the city forced him to flee to New York in 1940, where he was welcomed by Harry Holtzman, the art collector Peggy Guggenheim, James Johnson Sweeney, the art critic and museum director, and other members of the American artistic vanguard.

American phase. There, the last change in Mondrian's work occurred. Inspired by his regained freedom, New York's pulsating life, and the new rhythms of American music, he broke away first from the austere patterns of black lines, replacing them with coloured bands, and then substituted for the continuous flow of these bands a series of small rectangles that coalesced into a rhythmic flow of colourful vertical and horizontal lines. His late masterpieces, "New York City" and "Broadway Boogie Woogie," exhibited in 1943 and 1944, in the first personal exhibition Mondrian had after 1922, express this new vivacity. Buoyed by his hope for a better future, Mondrian started his "Victory Boogie Woogie" in 1943; it remained, however, unfinished: Mondrian died from pneumonia on February 1, 1944, in New York, at the age of almost 72.

Objectives in art. Mondrian's work—the most important collection of which is found in The Hague Gemeentemuseum—should not be considered merely as an artistic innovation. Certainly, the consistent development of his neoplasticism was an outstanding feat in the history of modern art, and Mondrian's work foreshadowed what was later called "geometric abstraction." He must have been conscious of his revolutionary innovation and of its implications: in a self-portrait of 1918, he painted himself against a background of one of his new, and at that time strange, pictures, thus stressing the historical importance of the new de Stijl principles.

Ethical significance of his work

But Mondrian's work goes beyond merely aesthetical considerations: his search for harmony through his painting has an ethical significance. Rooted in a strict puritan tradition of the Dutch Calvinism of his paternal home, he continually strove for purity during his long career as an artist. This purity is most clearly explained by the double meaning of the Dutch word *schoon*, which means both "clean" and "beautiful." The beauty of Mondrian's work consists in its purity, a purity that points beyond a painting to show humanity the way toward a life of health and clarity. When, in 1920, Mondrian dedicated his booklet "Le Néo-plasticisme" to "future men," he tried to indicate that art can be a guide to humanity, that it can work toward eliminating casual facts of appearance and the arbitrary outlook of the individual, and thereby substitute a new harmonious view of life for the conventional tragic conception of existence. To this purpose Mondrian chose the strict and rigid language of geometry to produce first of all an extreme purity, and on another level, a Utopia of superb clarity and force. Mondrian is a pioneer in modern art, but he is champion for humanity as well: he described the goal he set out to meet through his paintings in his 1937 essay in *Circle*: "One serves mankind by enlightening it."

MAJOR WORKS

"The Red Cloud" (1907; Haags Gemeentemuseum, The Hague); "Woods near Oele" (1908; Haags Gemeentemuseum, The Hague); "Windmill in Sunlight" (1908; Haags Gemeentemuseum, The Hague); "The Red Tree" (1908; Haags Gemeentemuseum, The Hague); "Still Life with Gingerpot" (1911–12; Haags Gemeentemuseum, The Hague); "The Gray Tree" (1912; Haags Gemeentemuseum, The Hague); "Oval Composition (Trees)" (1913; Stedelijk Museum, Amsterdam); "Blue Facade" (1914; Museum of Modern Art, New York); "Composition" (1916; Solomon R. Guggenheim Museum, New York); "Composition in Black and White" (1917; Rijksmuseum Kröller-Müller, Otterlo, the Netherlands); "Composition in Colour A" (1917; Rijksmuseum Kröller-Müller, Otterlo); "Composition III with Colour Planes" (1917; Haags Gemeentemuseum, The Hague); "Composition: Colour Planes with Gray Contours" (1918; Max Bill Collection, Zürich); "Composition in Diamond Shape" (1919; Rijksmuseum Kröller-Müller, Otterlo); "Composition Checkerboard" (1919; Haags Gemeentemuseum, The Hague); "Composition with Red, Blue and Yellow Green" (1920; Wilhelm Hack Collection, Cologne); "Composition with Red, Yellow, and Blue" (1921; Haags Gemeentemuseum, The Hague); "Composition 2" (1922; Solomon R. Guggenheim Museum, New York); "Composition 1 with Blue and Yellow" (1925; Kunsthaus, Zürich); "Composition with Yellow and Blue" (1929; Boymans-van Beuningen Museum, Rotterdam); "Fox Trot A" (1927; Yale University Art Gallery, New Haven); "Composition with Red, Blue and Yellow" (1930; Mr. and Mrs. Armand P. Bartos Collection, New York); "Composition with Yellow Lines" (1933; Haags Gemeentemuseum, The Hague); "Vertical Composition with Blue and White" (1936; Kunstsammlung Nordrhein-Westfalen, Düsseldorf, West Germany); "Composition with Blue" (1937; Haags Gemeentemuseum, The Hague); "Composition London" (1940–42; Albright-Knox Art Gallery, Buffalo); "New York City I" (1941–42; Sidney Janis Gallery, New York); "Broadway Boogie Woogie" (1942–43; Museum of Modern Art, New York); "Victory Boogie Woogie" (1943–44; Burton Tremaine Collection, Meriden, Connecticut, on loan to Museum of Modern Art, New York).

BIBLIOGRAPHY. PIET MONDRIAN has written a series of essays on art and on the task of art in daily life. The first, "De Nieuwe Beelding in de Schilderkunst" (on the new plastic art in painting) was published in *De Stijl*, vol. 1 (1917), a supplement to it in vol. 2 (1918). The "Dialog over de Nieuwe Beelding" (dialogue about the new plastic art) also appears in vol. 2. Volume 3 contains a trialogue Naturlijke en Abstracte Realiteit, published in English translation ("Natural Reality") in M. SEUPHOR, *Piet Mondrian: Life and Work* (1956). In 1920 Mondrian published "Le Néo-plasticisme: principe général de l'équivalence plastique," reprinted in The Hague *Catalogue of Mondrian Exhibition* (1955). A German adaptation of this essay, with four other texts, was published as a Bauhausbuch, *Die neue Gestaltung* (1925). In 1931 Mondrian wrote "Kunst en Leven" (art and life), published in English translation in H.L.C. JAFFE, *De Stijl, 1917–1931* (1956). His later writings are collected in *Plastic Art and Pure Plastic Art, and Other Essays*, ed. by R. MOTHERWELL in 1945, after Mondrian's death.

The first comprehensive monograph on Mondrian was M. SEUPHOR, *Piet Mondrian: Life and Work* (1956), with a catalogue raisonné of all Mondrian's paintings and an extensive bibliography. This important publication was followed by a series of other monographs: O. MORISANI, *L'Astrattismo di Piet Mondrian* (1956), dealing with the theoretical background of Mondrian's art; D. LEWIS, *Mondrian* (1957); and S. HUNTER, *Mondrian* (1958), two works that give a selection of his most characteristic phases; and C.L. RAGGHIANTI, *Mondrian e l'arte del XX secolo* (1962), a brilliant survey, but proposing some extravagant hypotheses. Further monographs on Mondrian are: F. MENNA, *Mondrian, cultura e poesia* (1962); F. ELGAR, *Mondrian* (Eng. trans. 1968); A. BUSIGNANI, *Mondrian* (Eng. trans. 1968); I. TOMASSONI, *Piet Mondrian* (1969), in Italian; L.J.F. WIJSENBEEK, *Piet Mondrian* (Eng. trans. 1969), written by the director of the Hague Municipal Museum, where the largest collection of Mondrian's work can be seen; and H.L.C. JAFFE, *Mondrian* (1970).

Various exhibition catalogs give important information on Mondrian's work and contain valuable introductions: *Zürich*, Kunsthaus (1955), with introduction by M. BILL; *Basel*, Galerie Beyeler (1964–65); with introduction by G. SCHMIDT ("Mondrian Today"); *Toronto, Philadelphia, The Hague*, ed. by R.P. WELSH (1966); *The Hague* (1968), catalog of the Mondrian collection in the Municipal Museum, ed. by C. BLOK, with an introductory essay; *Paris*, Orangerie des Tuileries (1969), with introduction by M. SEUPHOR; and *New York*, Solomon R. Guggenheim Museum (1971).

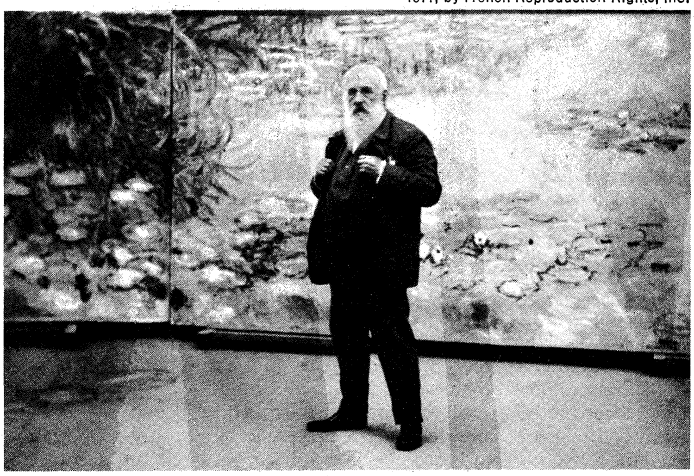
A range of articles deal with various aspects of Mondrian's art and with its implications: the series starts with Mondrian's pupil, C. VON WIEGAND, "The Meaning of Mondrian," *Journal of Aesthetics*, 2:62–70 (1943); and continues soon after Mondrian's death with two articles by J.J. SWEENEY: "Mondrian," in *Eleven Europeans in America*, MMA *Bulletin*, no. 13, pp. 35–36 (1946), and *Piet Mondrian* (MMA, 1948); G.C. ARGAN published a fundamental article on Mondrian in his *Studi e Note*, pp. 155–183 (1955). See also MARTIN S. JAMES, "The Realism Behind Mondrian's Geometry," *Art News*, 56: 34–37 (1957), and "Mondrian and the Dutch Symbolists," *Art Journal*, 23:103–111 (1963–64), dealing mainly with Mondrian's early period. M. SEUPHOR, "L'humanisme de Mondrian," *Aujourd'hui*, no. 2, pp. 5–9 (1955), deals with the sense of Mondrian's art, as experienced by a close friend. R.P. WELSH and J.J. JOOSTEN published *Two Mondrian Sketchbooks, 1912–1914* (1969), thereby casting light on the working process of Mondrian in his transitory phase.

(H.L.C.J.)

Monet, Claude

The art movement known as Impressionism began to take form about 1864, but it was not so identified until ten years later, when Claude Monet exhibited an oil sketch of the harbour at Le Havre under the title "Impression: Sunrise." This small, unpretentious study was sarcastically criticized by a journalist, who headed his article "Exhibition of the Impressionists," thus christening the movement that was to succeed Realism, led by the painters Gustave Courbet and Édouard Manet. Other great artists, among them Pissarro, Renoir, and Sisley, contributed much to the Impressionist style, but it was Monet who was its initiator, leader, and most undeviating advocate. As the critic Félix Fénéon wrote in 1888, "the word 'impressionism' was created for him, and it fits him better than it does anyone else."

H. Roger-Viollet, permission S.P.A.D.E.M. 1971, by French Reproduction Rights, Inc.



Monet standing in front of one of his water-lily paintings in his studio, 1920.

Childhood and early works. Claude Monet, the eldest son of Adolphe Monet, a grocer, was born in Paris on November 14, 1840. In 1845 the family moved to Sainte-Adresse, near Le Havre, where Monet's father took over the management of his family's thriving ship-chandlering and grocery business. This event has more than biographical significance, for it was Monet's childhood, spent along the beaches, and his intimate knowledge of the sea and the rapidly shifting Norman weather from which his fresh vision of nature arose. Monet's first success as an artist came when he was 15, with the sale of caricatures that were carefully observed and well drawn. Also preserved from these early years is a page of pencil sketches of sailing ships, almost technical in their clear descriptiveness. His aunt, Mme Lecadre, was an amateur painter, and, perhaps at her suggestion, Claude was sent to study drawing with a local artist. But his life as a painter did not begin until he was befriended by Eugène Boudin, who introduced the somewhat arrogant student to the practice—then uncommon—of painting in the open air. A single day working with Boudin focussed his familiarity with outdoor light and his talent as an artist. The experience was a revelation that set the direction for more than 60 years of concentration on visible phenomena and on the innovation of effective methods to transform perception into pigment.

Influence of Boudin

Although oil landscapes have been painted at least since the 16th century, they were produced in the studio—ideas of nature rather than direct perceptual records. The English painters John Constable and J.W.M. Turner made small oil sketches outdoors before 1810, but it is unlikely that Monet knew these studies. He first visited Paris in 1859–60, where he was impressed by the work of the Barbizon school painters Charles Daubigny and Constant Troyon. To his family's annoyance, he refused to enroll in the École des Beaux-Arts. Instead, he frequented the haunts of advanced artists and worked at the Académie Suisse, where he met Camille Pissarro. This informal training was interrupted by a call to military service, which he began in Algeria, where he was excited by the African light and colour. In 1862 he returned to Sainte-Adresse, perhaps because of illness, again painted the sea with Boudin, and met the Dutch marine painter Johan Barthold Jongkind. Later that year he continued to study in Paris, this time with the academician Charles Gleyre, in whose atelier he met Frédéric Bazille, Alfred Sisley, and Pierre-Auguste Renoir. After disagreements with their master, the group departed for the village of Chailly-en-Bière, near Barbizon in the forest of Fontainebleau. It was also during this period—or at least before 1872—that Monet discovered Japanese prints, the decorativeness and flatness of which were to have a strong influence on the development of modern painting in France. Perhaps also, Monet's choice of Algeria for service was a result of his admiration for the Romantic painter Eugène Delacroix, whose colouristic work he had seen in Paris and who had visited Morocco in 1832.

The exceptional achievements of Monet's prolific youthful period can be measured in works completed between 1865 and 1870, before he had begun to fragment his brushstrokes into the characteristic broken touches that were to become the hallmark of Impressionist style. One of the most ambitious of these early works (which was never finished, supposedly because of negative comments by Courbet) was "The Picnic" (1865–66), named after Édouard Manet's notorious painting of the same name ("Déjeuner sur l'herbe") shown in the Salon des Refusés in 1863. But, quite differently from Manet's masterpiece, which was a parody of a Renaissance painting in a contemporary setting, Monet's painting, about 15 feet high by 20 feet wide, was a direct, unprovocative representation of a group of fashionably dressed picnickers in the forest of Fontainebleau. For a young painter attempting plein-air style, the size of his composition was absurd—far too large to complete in the open. Only a fragment of the original canvas, now in the Louvre, has been preserved, but the entire composition can be seen in a superb smaller version in the Pushkin Museum, Moscow, and its early Impressionist style in a large study of Bazille and Camille Doncieux, Monet's mistress, in the National Gallery of Art, Washington, D.C. Monet shared with Manet, however, the new concern for actual scenes of modern life rather than contrived historical, romantic, or fanciful subjects. Thus, Monet's "Picnic" was an extension, by virtue of a more immediate empiricism, of the Realism of Courbet. Camille wore the same fashionable clothes for the famous "Women in the Garden" (1866–67), also in the Louvre. In the garden of his rented house at Ville-d'Avray, near Versailles, Monet fashioned a device to lower the eight-foot-high canvas into a trench, so he could paint a life-size composition entirely in the open. Another major figure composition of this period, "The Terrace at Sainte-Adresse," was purchased by the Metropolitan Museum of Art in 1967 for $1,411,200. Unique among Monet's early works, it represents his family regarding the sea from their flower-covered terrace. Monet's letters and documents from this period speak plaintively of the critical state of his finances, the difficulties of his life, trouble with his eyes, and the problems attending the birth of his first child, Jean, in 1867. Indeed, until the 1890s, when he was a revered and prosperous master, Monet was continually in dire financial straits. These complaints are not verified, however, by his paintings of the same period, which reflect a comfortable and middle class home life: "The Luncheon" (1868; Stadelsches Kunstinstitut, Frankfurt, West Germany) represents Camille, fastidiously dressed, seated with young Jean at the table. An equally well dressed friend leans, watching, against the window sill, and a uniformed maid appears at an open door. The food, dishes, napery, and glassware are superbly and almost impressionistically rendered.

"The Picnic"

To paint *la vie moderne* ("modern life") was not to be the primary aim of Monet's art, although Impressionism, broadly viewed, was a celebration of the pleasures of middle class life. Of more significance in Monet's case was his ceaseless search for painterly means to implement his radical view of nature. "The River" (1868) or "The

Beach at Sainte-Adresse" (1867), both in the Art Institute of Chicago, give a better accounting than do the ambitious figure paintings of his advance toward the Impressionist style. Of all types and periods of Monet's painting, moreover, the beach and sea pictures of 1865–67—more powerfully painted than those of Boudin—seem closest to "normal" (*i.e.*, casual) vision. Monet was plainly trying not to copy the scene before him as examined in detail but to record on the spot the impression that relaxed, momentary vision might receive—what is seen rather than what is known, with all its vitality and movement. Boats, buildings, incidental figures, and the pebble beach are swiftly brushed in as flat colour patterns, with little attention to their weight or solidity.

Advance toward Impressionism

First Impressionist paintings. Monet's life during the '60s was precarious and itinerant, and he sold almost nothing; but several works were accepted for exhibition in the yearly salons, most notably, and with great success, a fine but not yet Impressionist portrait of Camille. In 1868 he won a medal in Le Havre. Despite difficulties, moreover, he was able to paint in Paris, Le Havre, Chailly, Honfleur, Trouville, Fécamp, and at other stations between Paris and the sea. The '60s came to an auspicious close at the Seine River resort known as La Grenouillère, at Bougival, where he and Renoir worked together for the first time. In canvases almost identical in style, they made rapid notations of pleasure-seekers and bathers, rowboats bobbing in the foreground, and the scintillating reflections in the lapping water. Regarded by Monet as "bad sketches," they were already Impressionist. Both artists' Bougival studies interpret the light and movement of outdoor life in strong, abbreviating strokes, improvised at the moment of perception, that together serve as equivalents for visual experiences never before so truthfully committed to canvas. In 1870 at Trouville, in broad, assured gestures, Monet painted a study of Camille on the beach (Tate Gallery, London). It is as animated an example of visual realism as had ever been painted: grains of sand remain embedded in the pigment.

As the '70s began, Monet continued his pursuit of natural phenomena. In order to avoid the Franco-Prussian War, he left his son and Camille, whom he had just married, and travelled to London. Here, with Pissarro, he must have seen work by English masters—perhaps even Constable and Turner—and was introduced by Daubigny to Paul Durand-Ruel, who was to become his dealer. In 1871 and 1872 he painted canals, boats, and windmills in Holland and worked again at Le Havre; "Impression: Sunrise" was a product of this stop. On his return, Monet rented a house at Argenteuil, on the Seine near Paris. The years he lived there mark the height of the group Impressionist movement. Sometimes joined by Renoir or another friend, he painted his home and family, the town, the river, and the pleasures of life on it and along its banks. In 1874 even Manet visited Monet and painted him working in the "floating studio" he had constructed the year before. Indeed, it was the younger artist who introduced the elder to the practice of painting in the open—not the other way around, as the historical sequence of the Realist and the Impressionist movements suggest. The radiant, luminous paintings of this period (1872–75) by Monet, Renoir, Sisley, Pissarro, and others remain today the most coveted and characteristic works of Impressionism as a collaborative venture.

Work at Argenteuil

Later Impressionism. Monet's celebrated method of producing works in series, each representing the same motif under different light and weather conditions, was not fully implemented until the '90s, but what is usually regarded as the first series was executed in or around the Gare Saint-Lazare in Paris during the winter of 1876–77. A total break with the customary Impressionist subjects, the engines belching smoke and steam in the great shed recall Turner's "Rain, Steam, and Speed—The Great Western Railway" of 1844 and prefigure the mechanical subjects painted by Italian Futurists after 1909. The summer of 1877, moreover, was the end of the idyllic art of Argenteuil. Monet's life was less happy after he moved to Vétheuil, farther from Paris. In 1876 a liaison had begun between Monet and Alice Hoschedé, the wife of a department-store owner and collector. Monet had incurred a burden of debts in Argenteuil, and Camille was pregnant and ill. At Vétheuil they were joined by Mme Hoschedé, who had left her husband, and at least five of her children. From her dowry she assumed Monet's debts and cared for Camille, who died in September 1879. Many of the paintings done in the new locale radiate the same joy in nature as those of Argenteuil, but one group, done after Camille's death, is surely the most melancholy of Monet's career. It represents the Seine in flood during the winter, blocks of ice heaped above its surface, and trees crushed by the flow. Rendered in broad jabs of grayed greens, blues, and browns, these are scenes of profound sadness and desolation. The "Break-up of the Ice" series also demonstrates, as do paintings from every period, that Impressionism was not a theoretical art devoted merely to simulating sunlight by juxtaposing bright hues, as it once was assumed. From the beginning, Monet's palette always responded to the unique appearances before him; if the day was overcast, his tones were muted and gray. For Monet, winter was as valued a subject as summer, despite the discomforts it entailed. And, though Impressionist colour was surely of a higher key than previous landscape painting had been and more roughly brushed, it was an art not of formula but of poetic response and innovation. Once a motif was selected, Monet composed as a photographer does on his ground glass, accepting the pattern of the elements before him but framing them with great care. Later he emphasized those elements—vertical masts, tall trees, the horizontals of riverbanks, boats, or rooftops—that provide pictorial structure.

By 1881 the original Impressionist group began to disintegrate, although it was still to hold two more exhibitions—the eighth and last (in which Monet did not show) in 1886, after the advent of Neoimpressionism. Only Monet continued with the same fervour to carry on the scrutiny of nature. If one regards science as empirical rather than deductive, his intense application can be seen as "scientific." Yet Monet's landscapes were always immediate, poetic, and continually changing in treatment from motif to motif. Among the sites he chose during the '80s were Pourville, Étretat, Fécamp, and Varangéville in Normandy, the rugged and isolated Breton island of Belle-Île, the wild Creuse River Valley, Menton and Antibes in the Midi, and Bordighera in Italy. In 1886, he made a second visit to Holland to paint the tulip fields, before important sojourns at Étretat and Belle-Île.

Continuing scrutiny of nature

In 1883 Monet, Mme Hoschedé, her children, and Monet's sons Jean and Michel settled at Giverny, a hamlet near Vernon, 52 miles from Paris, on the tiny Epte River but still close to the Seine. Here Monet purchased a farmhouse surrounded by an orchard, which was to be his home until his death and is now a French national monument. After the travels of the '80s, the '90s were spent at or near Giverny in concentrated application to one series after another, but were interrupted from January to April in 1895 by a bold painting trip to Norway to paint the Arctic winter and a rest cure that June in the Pyrennees. The series were usually dated and often exhibited in groups: "Poplars" (1890–91), "Haystacks" (1889–93), "Rouen Cathedrals" (1892–94), "Mornings on the Seine" (1896–97), etc. The most radical, perhaps, was the "Haystacks," of which 15 were shown at the Galeries Durand-Ruel in 1891. By comparison with earlier Impressionist subjects, they provided a neutral motif, devoid of the lyricism of the '70s. The mute heaps of hay served simply as modulators of light—for light, season, moment of the day, weather, and the psychological and technical process of envisioning and transmitting impressions to canvas and pigment were all that seemed to interest Monet at that time. Surely these unprecedented canvases were as radical for their time as the abstract paintings of Wassily Kandinsky in 1912 or the poured and dripped paintings of Jackson Pollock in 1947. The "Cathedrals," 20 of which were shown together in 1895, were the most systematic and theoretical of all Monet's works. Begun in Rouen, they were completed at Giverny but represented the facade of Rouen Cathedral at dawn, at various times during

The "Cathedrals"

the day, at sunset, in bright and gray weather, in sun and mist. They were no more religious in content than the "Haystacks" were rustic. The architectural screen of the worn facade served, like the spongy fabric of the stacks, as an armature, pulverized by sun or drowned in mist, on which to hang an iridescent curtain of light.

Last years. After 1900, two ambitious projects, both far from Giverny, concluded Monet's search for new motifs. The first (for which he made at least three trips to London between 1899 and 1904) was the extensive, multiple series representing the Thames River, the Waterloo and Charing Cross bridges, and the House of Parliament. Recalling, in their exotic coloration and mysterious, romantic mood, the Thames paintings of Turner and Whistler, it is atmosphere, more than structures, that is the subject; buildings and bridges are less tangible than the pulsating brush strokes that give volume to the light-filled fog and mist. The second and last of the architectural motifs was the canals and palaces of Venice. They were begun in 1908 and continued in 1909, although Monet worked on the series at Giverny until 1912. Venice was a perfect Impressionist subject, but the light, water, movement, architecture, and reflections in the water are more generalized than the specific weather effects of the systemic "Haystacks" and "Cathedrals."

In 1890 Monet had bought a strip of marshland across the road from his house and flower garden, through which flowed a tributary of the Epte. By diverting this stream, he began to construct his now world-renowned water-lily garden. Soon weeping willows, iris, and bamboo grew around a free-form pool, clusters of lily pads and blossoms floated on the quiet water, and a Japanese bridge closed the composition at one end. By 1900 this unique product of Monet's imagination (for his Impressionism had become more subjective) was in itself a major work of environmental art—an exotic lotus land within which he was to meditate and paint for over 20 years. The first canvases of lilies, water, and the Japanese bridge were only about one yard square, but their unprecedented open composition, with the large blossoms and pads suspended as if in space, and the azure water in which clouds were reflected, implied an encompassing environment beyond the frame. This concept of embracing spatiality, new to the history of painting and only implicit in the first water-lily paintings, was expanded by 1925 into a cycle of huge murals to be installed in Paris in two 80-foot oval rooms in the Orangerie of the Tuileries. These were described in 1952 by the painter André Masson as "the Sistine Chapel of Impressionism." This crowning achievement of Monet's long, probing study of nature—his striving to render his impressions, as he said, "in the face of the most fugitive effects"—was not dedicated until 1927, more than a year after his death. The many large studies for the Orangerie murals, as well as other unprecedented and unique works painted in the water garden between 1916 and 1925, were almost unknown until the 1950s but are now distributed throughout the major private collections and museums of the world. Among these is a bequest made to the Musée Marmottan, Paris, by Monet's son Michel and opened to the public in 1971. Despite failing eyesight, Monet continued to paint almost until his death in Giverny, on December 5, 1926.

The Orangerie murals

MAJOR WORKS

"Still Life" (*c.* 1859; Leigh B. Block collection, Chicago); "Farm in Normandy" (*c.* 1864; Louvre, Paris); "View of the Coast at Le Havre" (1865; Minneapolis Institute of Arts, Minnesota); "Camille (The Green Dress)" (1866; Kunsthalle, Bremen, West Germany); "Garden of the Princess, Paris" (1866; Oberlin College, Ohio); "The Terrace at Sainte-Adresse" (1866; Metropolitan Museum of Art, New York); "Women in the Garden" (1866–67; Louvre, Paris); "The River" (1868; Art Institute of Chicago); "La Grenouillere" (1869; Metropolitan Museum of Art, New York); "The Beach at Trouville" (1870; Tate Gallery, London); "Hyde Park, London" (1871; Philadelphia Museum of Art); "The Thames and the Houses of Parliament" (1871; Lord Astor of Hever collection, London); "Impression: Fog" (1872; Musée Marmottan, Paris); "Impression: Sunrise" (1872; Musée Marmottan, Paris); "Wild Poppies" (1873; Louvre, Paris); "The Railroad Bridge at Argenteuil" (1875; Philadelphia Museum of Art); "La Japonaise" (1876; Museum of Fine Arts, Boston); "Gare Saint-Lazare, Paris" (1877; Fogg Museum of Art, Harvard University); "Rue Montorquiel Bedecked with Flags" (1878; Musée des Beaux-Arts, Rouen); "Vétheuil in Summer" (1880; Metropolitan Museum of Art, New York); "Haystacks at Giverny" (1884; Josef Rosensaft collection, New York); "Antibes" (1888; Toledo Museum of Art, Ohio); "Two Haystacks" (1891; Art Institute of Chicago); "Rouen Cathedral, Sunset" (1894; Museum of Fine Arts, Boston); "Water-Lily Pool" (1899–1906; Louvre, Paris); "Les Nymphéas" series (1906–26; Jeu de Paume, Paris).

BIBLIOGRAPHY. No *oeuvre* catalog or definitive biography of Monet exists. A volume purporting to fill the latter need, by CHARLES MERRILL MOUNT, *Monet: A Biography*, appeared in 1967. Its account of Monet's life is distorted and prejudiced. The author, however, presents some new and important documentation of Monet's early life and especially of his relationship with his mistress and first wife, Camille Doncieux. Indispensible for the study of Monet, as for other Impressionist painters, is JOHN REWALD, *History of Impressionism*, rev. ed. (1962). It includes copious notes, with references to publications, problems, and documents, and an exhaustive annotated bibliography. WILLIAM C. SEITZ, *Claude Monet* (1960), is a most useful recent book, including colour reproductions and an extensive outline biography. Among the books published during Monet's lifetime, perhaps that by Monet's friend GUSTAVE GEFFROY, *Claude Monet, sa vie, son temps, son oeuvre* (1922), is the most rewarding. See also GEORGE HEARD HAMILTON, *Claude Monet's Paintings of Rouen Cathedral* (1960), and for an account of the water garden, including photographs, STEPHEN GWYNN, *Claude Monet and His Garden* (1934).

(W.C.Se.)

Money

The subject of money has fascinated wise men from the time of Aristotle to the present day because it is so full of mystery and paradox. The piece of paper labelled one dollar or 100 francs or 10 kroner or 1,000 yen is little different, as paper, from a piece of the same size torn from a newspaper or magazine, yet it will enable its bearer to command some measure of food, drink, clothing, and the remaining goods of life while the other is fit only to light the fire. Whence the difference? Why should these particular pieces of paper be accepted in transactions for goods and services?

The basic function of money

The easy answer, and the right one, is that people accept money as such because they know that others will. The pieces of paper are valuable because everyone thinks they are, and everyone thinks they are because in his experience they always have been. At bottom money is, then, a social convention, but a convention of uncommon strength that people will abide by even under extreme provocation. The strength of the convention is, of course, what enables governments to profit by inflating the currency. But it is not indestructible. When great variations occur in the quantity of these pieces of paper —as they have during and after wars—they may be seen to be, after all, no more than pieces of paper. People will then seek substitutes—like the cigarettes and cognac, which for a time became the medium of exchange in Germany after World War II. As John Stuart Mill wrote:

> There cannot, in short, be intrinsically a more insignificant thing, in the economy of society, than money; except in the character of a contrivance for sparing time and labour. It is a machinery for doing quickly and commodiously, what would be done, though less quickly and commodiously, without it: and like many other kinds of machinery, it only exerts a distinct and independent influence of its own when it gets out of order. (*Principles of Political Economy*, W.J. Ashley [ed.], 1909, p. 488.)

Mill was perfectly correct although one must add that there is hardly a contrivance man possesses that can do more damage to a society when it goes wrong.

The basic function that money performs is to enable buying to be separated from selling, thus permitting trade to take place without the so-called double coincidence of barter. If a person has something to sell and wants something else in return, he does not have to search for someone able and willing to make the desired exchange of items. He can sell what he has for general purchasing

power—that is, "money"—to anyone who wants to buy it; he can then use the proceeds to buy what he wants from anyone who wants to sell it.

The importance of this function of money is dramatically illustrated by the experience of Germany just after World War II, when paper money was rendered largely useless because, despite inflationary conditions, price controls were effectively enforced by the American, French, and British armies of occupation. People had to resort to barter or to inefficient money substitutes. The result was to cut total output of the economy in half. The German "economic miracle" just after 1948 reflected partly a currency reform by the occupation authorities, but some economists hold that it stemmed primarily from the German government's elimination of all price controls, thereby permitting a money economy to replace a barter economy.

Separation of the act of sale from the act of purchase requires the existence of something that will be generally accepted in payment—this is the "medium of exchange" function of money. But also, there must be something that can serve as a temporary abode of purchasing power, in which the seller holds the proceeds in the interim between his sale and his subsequent purchase, or from which the buyer can extract the general purchasing power with which he pays for what he buys. This is the "asset" function of money.

VARIETIES OF MONEY

Anything can serve as money that habit or social convention and successful experience endow with the quality of general acceptability, and a very wide range of items have so served—from the wampum (beads made from shells) of American Indians to cowries (brightly coloured shells) in India, to whales' teeth among the Fijians, to tobacco among early colonists in North America, to large stone disks on the Pacific island of Yap, and to cigarettes and liquor in post-World War II Germany. The wide use of cattle as money in primitive times survives in the word pecuniary, which comes from the Latin word *pecus*, meaning cattle.

Metallic money. The use of metals as money has occurred throughout history. As Aristotle observed,

> The various necessities of life are not easily carried about, and hence man agreed to employ in their dealings with each other something which was intrinsically useful and easily applicable to the purposes of life, for example, iron, silver, and the like. Of this the value was at first measured by size and weight, but in process of time they put a stamp upon it, to save the trouble of weighing and to mark the value.

The use of metal for money can be traced back to over 2,000 years before the birth of Christ. But standardization and certification in the form of coinage, as referred to by Aristotle, did not occur except perhaps in isolated instances until the beginning of the 7th century BC. Historians generally assign to Lydia, a Greek state in Asia Minor, priority in using coined money. The first coins were made of electrum, a natural mixture of gold and silver, and were crude, bean-shaped ingots bearing a primitive punchmark certifying to either weight or fineness, or perhaps both.

Debasing the currency

The use of coins enabled payment to be by "tale" or count rather than weight, greatly facilitating commerce. But this in turn encouraged clipping (shaving off tiny slivers from the sides or edges of coins) and sweating (shaking a bunch of coins together in a leather bag and collecting the dust that was thereby knocked off) in the hope of passing on the lighter coin at its face value. Gresham's law (that "bad money drives out good" when there is a fixed rate of exchange between them) came into operation, and heavy, good coins were held for their metallic value, while light coins were passed on. The coins became lighter and lighter, and prices higher and higher. Then payment by weight would be resumed for large transactions, and there would be pressure for recoinage. These particular defects were largely ended by the "milling" of coins (making serrations around the circumference of a coin), which began in the late 17th century.

A more serious matter was the attempt by the sovereign to benefit from his monopoly of coinage. In this respect, Greek and Roman experience offers an interesting contrast. Though Solon, on taking office in Athens in 594 BC, did institute a partial debasement of the currency, for the next four centuries, until the absorption of Greece into the Roman Empire, the Athenian *drachma* had an almost constant silver content (67 grains of fine silver until Alexander, 65 grains thereafter) and became the standard coin of trade in Greece and in much of Asia and Europe as well. Even after the Roman conquest, the *drachma* continued to be minted and widely used.

The Roman experience was very different. Not long after the silver *denarius* patterned after the Greek *drachma* was introduced (in 269 BC), the prior copper coinage (*aes* or *libra*) began to be debased until by the time the Empire began its weight had been reduced from one pound to half an ounce. The silver *denarius* and the gold *aureus* (introduced about 87 BC) suffered only minor debasement until the time of Nero (54 AD), when almost continuous tampering with the coinage began. The metal content of the gold and silver coins was reduced, and the proportion of alloy was increased to three-fourths or more of its weight. Debasement in Rome (as ever since) was a reflection of the state's inability or unwillingness to finance its expenditures "through explicit taxes." But the debasement in turn worsened Rome's economic situation and undoubtedly contributed to the collapse of the empire.

Paper money. In the late 18th and early 19th centuries, paper money and bank notes spread widely. The bulk of the money in use came to consist not of actual gold or silver but of fiduciary money—promises to pay specified amounts of gold and silver. These promises were initially issued by private individuals or companies as bank notes or transferrable book entries that have come to be called deposits. But gradually the state assumed an increasing role.

From fiduciary paper money promising to pay gold or silver, it is a short step to fiat paper money—that is, notes that are issued on the "fiat" of the sovereign, are specified to be so many dollars or francs or yen and are legal tender but are not promises to pay something else. The first large-scale issue in Western countries occurred in France in the early 18th century (though there are reports of paper money in China many centuries earlier). Later, the French revolutionary government issued paper money in the form of *assignats* from 1789 to 1796. The American colonies and later the Continental Congress issued bills of credit that could be used in making payments. These early experiments gave fiat money a deservedly bad name. The money was overissued, and prices rose drastically until the money became worthless or was redeemed in metallic money (or promises to pay metallic money) at a small fraction of its initial value.

Subsequent issues of fiat money in the major countries during the 19th century were temporary departures from a metallic standard. In Britain, for example, payment of gold for the outstanding bank notes was "suspended" during the Napoleonic Wars (1797–1816). As a result, gold coin and bullion became more expensive in terms of paper. Similarly, in the U.S. during the Civil War, convertibility of Union currency (greenbacks) into specie was suspended and resumption did not occur until 1879. At its peak, in 1864, the greenback price of gold, nominally equivalent to $100, reached over $250.

In the Middle Ages, when money consisted primarily of coins, silver and gold coins circulated simultaneously. As governments came increasingly to take over the coinage, and especially as fiduciary money was introduced, they tended to specify their nominal monetary units in terms of fixed weights of both silver and gold—to adopt a national bimetallic standard. Gresham's law, however, usually assured that the bimetallic standard degenerated into a monometallic standard: if the quantity of silver designated as the monetary equivalent of one ounce of gold was less than the quantity that could be purchased in the market for one ounce of gold; *i.e.*, if the mint overvalued silver, no one would bring gold to be coined. If he had gold, it was better to buy silver and bring it to be coined.

Silver, the cheaper metal, "drove out" gold and became the standard. This happened in most of the countries of Europe, so that by the early 19th century all were effectively on a silver standard. In Britain, on the other hand, the ratio established in the 18th century at the advice of Sir Isaac Newton, then serving as master of the mint, overvalued gold and therefore led to an effective gold standard. In the United States, a ratio of 15 ounces of silver to 1 ounce of gold was set in 1792. This ratio overvalued silver, so silver became the standard. In 1834, the ratio was altered to 16 to 1, which overvalued gold, so gold became the standard.

Gold. The great gold discoveries in California and Australia in the 1840s and 1850s produced a temporary decline in the value of gold in terms of silver. This price change plus the dominance of Britain in international finance led to a widespread shift from a silver standard to a gold standard. Germany adopted gold in 1871–73; the Latin Monetary Union (France, Italy, Belgium, Switzerland) in 1873–74; the Scandinavian Union (Denmark, Norway, and Sweden) and the Netherlands in 1875–76. By the final decades of the century, silver remained dominant only in the Far East (China, in particular). Elsewhere the gold standard reigned.

The international gold standard

The early 20th century was the great era of the international gold standard. Gold coins circulated in most of the world; paper money, whether issued by private banks or by government, was convertible on demand into gold coins or gold bullion at an official price (with perhaps the addition of a small fee); and bank deposits were convertible into either gold coin or paper currency that was itself convertible into gold. In a few countries, a minor variant prevailed—the so-called gold-exchange standard, under which the currency was converted at a fixed price into the currency of another country (usually the British pound sterling) that was itself convertible into gold.

There was, in effect, a single world money called by different names in different countries. A U.S. dollar, for example, was defined as 23.22 grains of pure gold (25.8 grains of gold 0.9000 fineness). A British pound sterling was defined as 113.00 grains of pure gold 123.274 grains of gold 11/12th fine). Accordingly, one British pound equalled 4.8665 U.S. dollars (113.00/23.22) at the official parity. The actual exchange rate could deviate from this only by an amount that corresponded to the cost of shipping gold. If the price of the pound sterling in terms of dollars rose to a considerably higher value than this in the foreign exchange market, someone in New York who had a debt to pay in London might find that, rather than buy the pounds he needed on the market, it was cheaper to get gold for his dollars at his bank or at the U.S. subtreasury, ship the gold to London, and get pounds for his gold from the Bank of England. This set an upper limit to the exchange rate. Similarly, the cost of shipping gold from Britain to the U.S. set a lower limit. These limits were known as the gold points.

Under such an international gold standard, the quantity of money in each country is determined by the specie-flow adjustment mechanism analyzed by 19th-century economists. If for whatever reason, the quantity of money in a country rises unduly, this will tend to raise prices in that country relative to prices in other countries, discouraging exports and encouraging imports. The decreased supply of foreign currency from the sale of exports plus the increased demand for foreign currency to pay for imports will tend to raise the price of foreign currency in terms of domestic currency. As soon as this price hits the upper gold point, gold will be shipped out of the country to other countries. The decline in the amount of gold will produce in turn a reduction in the total amount of money—because banks and government institutions, seeing their gold reserves decline, will want to protect themselves against further demands by reducing the claims against gold that are outstanding. This will tend to lower prices at home. The influx of gold abroad will have the opposite effect, increasing the quantity of money there and raising prices. These adjustments will continue until the gold flow ceases or is reversed.

This is precisely the mechanism that operates within a unified currency area; the mechanism that determines how much money there is in Illinois compared to how much there is in other states or how much there is in Wales compared to how much there is in other parts of the United Kingdom. In the early 20th century, most of the world was a unified currency area so the gold standard functioned throughout the world. Its great advantage, which accounts for the widespread support that still exists for a real gold standard, is that—if permitted to operate—it greatly limits the power of any national government and prevents it from engaging in irresponsible monetary expansion. This is also its great disadvantage. In an era of big government and of full employment policies, a real gold standard would tie the hands of national governments in one of the most important areas of policy.

The decline of gold. World War I ended the real international gold standard. Most belligerents suspended the free convertibility of gold. The U.S., even after its entry into the war, maintained convertibility but embargoed gold exports. For a few years after the end of World War I, most nations had inconvertible national paper standards—inconvertible in the sense that paper money was not convertible into gold or silver, rather than in the more recent sense of its not being convertible into other currencies. The exchange rate between any two currencies was a market rate that fluctuated from time to time. During this period this was regarded as a temporary phenomenon, like the British suspension of gold payments during the Napoleonic era, and the U.S. suspension during the Civil War greenback period. The great aim was a restoration of the prewar gold standard.

Managed gold-exchange standard

This aim dominated monetary developments during the 1920s. Britain, still a major financial power, returned to gold in 1925. Winston Churchill, then Chancellor of the Exchequer, decided to follow prevailing financial opinion and adopt the prewar parity (*i.e.*, to define a pound sterling once again as equal to 123.274 grains of gold 11/12th fine). This produced exchange rates that, at the existing prices in sterling, overvalued the pound and so tended to produce gold outflows, especially after France returned to gold in 1928 at a parity that undervalued the franc. By 1929 the important currencies of the world, and most of the less important ones, were again linked to gold.

The gold standard that was restored, however, was a far cry from the prewar gold standard. The establishment of the Federal Reserve System in the United States introduced an additional link in the international specie flow mechanism. That mechanism no longer operated automatically. It operated only if the Federal Reserve chose to let it do so, and the Federal Reserve did not so choose; it took certain measures that prevented gold inflows from producing a corresponding expansion in the money supply. The United Kingdom had recurring difficulties in retaining its gold. It followed a restrictive monetary policy, but, because of rigidities in prices and particularly in wages, the result was unemployment rather than a lowering of prices. The unemployment reduced imports and stimulated exports, thus preventing gold losses. But this was not a tenable long-term position, as a lower level of prices might have been. Most other countries adopted a gold-exchange standard. Everywhere national central banks took scattered measures to loosen the connection between changes in the money supply and inflows or outflows of gold.

If the Great Depression had not occurred, this system might have grown and matured and improved. But the Great Depression brought the managed gold standards to a quick end. It originated in the United States, spread to all the countries linked to the U.S. by the gold exchange standard, and then reverberated back to the U.S. Britain was the first major country to cut the link, leaving the gold standard in 1931. The United States followed in March 1933.

The best simple description of the world's monetary system today is perhaps that the world is on a dollar standard. The dollar is widely used in international trade, even in trade between countries other than the U.S. It is the unit in terms of which the exchange rates of other currencies are expressed. Other countries maintain their

"official" exchange rates by buying and selling U.S. dollars and hold dollars as their primary reserve currency for this purpose. The existence of a dollar standard does not mean that other countries may not change their exchange rates from time to time—as Germany and France did in 1969—just as the gold standard did not mean that countries could not "devalue" or "appreciate" in terms of gold. What it does mean is that the U.S. cannot determine its own exchange rate or its balance of payments position. It also means that U.S. monetary policy has a major effect on the world economy.

MODERN MONETARY SYSTEMS

Monetary systems are today very much alike in all the major countries of the world. They consist of three levels: (1) the holders of money (the "public")—individuals, businesses, governmental units; (2) commercial banks (privately or governmentally owned), which borrow from the public and make loans to individuals, firms, or governments; and (3) central banks, which have a monopoly on the issue of certain types of money, serve as the bankers for the central government and the commercial banks, and have the power to determine the quantity of money.

The money holdings of the public

The public holds its money in two forms: (1) currency (including coin) and (2) bank deposits.

In most countries the bulk of the currency consists of notes issued by the central bank. In the United Kingdom these are Bank of England Notes, in the United States Federal Reserve Notes, and so on. It is hard to say precisely what "issued by the central bank" means. In the U.S., for example, the currency bears the words "Federal Reserve Note," but these notes are not obligations of the Federal Reserve Banks in any meaningful sense. The holder who presents them to a Federal Reserve Bank has no right to anything except other pieces of paper adding up to the same face value. The situation is much the same in most other countries.

The other major item of currency held by the public is coin. In almost all countries this is token coin, worth as metal much less than its face value.

Bank deposits. Bank deposits are also counted as part of the money holdings of the public. In the 19th century most economists regarded only currency and coin as "money," treating deposits only as claims to money. As deposits became more and more widely held, and as a larger fraction of transactions came to be effected by check, economists started to include, not the checks but the deposits they transferred, as money on a par with currency and coin.

The definition of money has been the subject of much dispute. The chief point at issue is which categories of bank deposits to call money and which to regard as near money. Many economists include as money only deposits transferrable by check (demand deposits). Others include nonchecking deposits such as "time deposits" or "current deposits" in commercial banks. Still others include deposits in other financial institutions such as savings banks, saving and loan associations, and so on.

The term deposits is highly misleading. It connotes something deposited for safekeeping, like currency in a safe-deposit box. Bank deposits are not like that. When one brings currency to a bank for "deposit," the bank does not put the currency in a vault and keep it there. It may put a small fraction of the currency in the vaults as "reserves," but it will lend most of it to someone else or buy an investment—that is, a bond or some other security. As part of the inducement to depositors to lend it money, it provides facilities for transferring demand deposits from one person to another by check.

The deposits of commercial banks are assets of their holders but liabilities of the banks. The assets of the banks consist of "reserves"—currency plus deposits at other banks—and "earning assets"—loans plus investments in the form of bonds and other securities. The reserves are only a small fraction of the aggregate deposits. Initially, in the history of banking, the amount of reserves held was determined by each bank separately in terms of its judgment of the likely demands on it. The growth of deposits enabled the total quantity of money (including deposits) to be larger than the total sum available to be held as reserves. A bank that received, say, $100 in gold might add $25 to its reserves and lend out $75. But the recipient of the $75 loan would himself spend it. Some of those who received gold this way would hold it as gold but others would deposit it in this bank or in other banks. If, for example, two-thirds was redeposited, some bank or banks would find $50 added to deposits and to reserves and would repeat the process, adding $12½ to its reserves and lending out $37½. When this multiple expansion process worked itself out fully, total deposits would have increased by $200, bank reserves by $50, and $50 of the initial $100 deposited would have been retained as "currency outside banks." There would be $150 more money in total than before (deposits up by $200, currency outside banks down by $50). Although no individual bank "created" money, the banking system as a whole did. This is the well-known multiple expansion process that lies at the heart of the modern monetary system.

Central banking. An important part of the monetary system is the central bank. The Bank of England was the first modern central bank, serving as the model for many others. It was established as a private bank in 1694 but quickly came to be largely an agency of the government. The Bank of France was established as a governmental institution by Napoleon in 1800. In the U.S., the 12 Federal Reserve Banks, which together with the Board of Governors in Washington constitute the Federal Reserve System, are technically owned by their member commercial banks, but this is a pure formality. Member banks get only a fixed annual percentage dividend on their stock and have essentially no formal power. To all intents and purposes, the system is a governmental agency and is best treated as such.

High-powered money

The notes issued by a central bank (or another governmental agency) plus deposits at the central bank are often called high-powered money, because when held as bank reserves each dollar or pound or franc may correspond to several dollars or pounds or francs of commercial bank deposits. Generally speaking, there is now no formal limit to the amount of notes and deposits that a central bank may have as liabilities.

The way in which a central bank increases or decreases the total amount of high-powered money is, typically, by making loans (discounting) or by buying and selling government securities (open-market operations). If, for example, the Federal Reserve System purchases $1,000,000 of government securities, it will pay for these securities by a check on itself, adding $1,000,000 to its assets and $1,000,000 to its liabilities. The seller of the security can take the check to a Federal Reserve Bank, which will give him $1,000,000 in Federal Reserve Notes. Or the seller may deposit the check at his bank, and the bank will in turn present it for payment to a Federal Reserve Bank which will "pay" the check by making a book entry increasing that bank's deposits with it by $1,000,000. The bank may, in turn, transfer this sum to a borrower who again will either convert it into Federal Reserve Notes or deposit it.

The important point is that these bookkeeeping operations simply record a process whereby the central bank has created, out of thin air as it were, additional high-powered money—the direct counterpart of printing Federal Reserve Notes. Similarly, if the central bank sells government securities, it destroys high-powered money.

In addition to the high-powered money of the central bank, the total quantity of money at any given time depends on the preferences of the public as to the relative amounts of money it wishes to hold as currency and as deposits and on the preferences of the banks as to the ratio they wish to maintain between their reserves and their deposits. (The reserve ratio is of course dominated by legal reserve requirements, where they exist, but may vary somewhat as banks think it prudent to keep a larger or smaller cushion in excess of required reserves.)

It follows that, by controlling the quantity of high-powered money and by other less important means, a central

bank can vary the total quantity of money as it wishes within broad limits. The major problem of modern monetary policy is how the central bank should use this power.

THE QUANTITY THEORY OF MONEY

The relation between money and what it will buy has always been a central issue of monetary theory. Economists have generally held that the level of prices is determined by the quantity of money. But precisely how the quantity of money affects the level of prices, and what the effects are of changes in the quantity of money, have been conceptualized in different ways at different times. For an interval of two or three decades from the mid-1930s to the mid-1960s there was a widespread rejection of the quantity theory by professional economists. More recently, there has been something of a counterrevolution involving a partial return to and modification of earlier views. The following discussion of monetary theory will therefore deal first with the quantity theory of money, as it developed before the 1930s, then with the Keynesian revolution, and finally with the recent counterrevolution.

The basic principles of the quantity theory

In whatever way it has been formulated, the quantity theory of money has involved a distinction between the *nominal* quantity of money and the *real* quantity of money. The nominal quantity of money is that expressed in whatever units are used to designate money—talents, shekels, pounds, francs, lire, drachmas, dollars, yen, and so on. The real quantity of money is that expressed in terms of the volume of goods and services that the money will purchase. The quantity theory assumes that what ultimately matters to holders of money is the real quantity rather than the nominal quantity. If this is so, then whatever factors may determine the nominal quantity of money, it is the holders of money who determine the real quantity and, in the process, the price level.

The following may be considered a hypothetical example. In a certain community the quantity of money in existence is $1,000,000, and the total income of the community is $10,000,000 a year. On the average, each member of the community holds an amount of money equal in value to one-tenth of a years' income, or to 5.2 weeks' income. Put differently, the income velocity of circulation is equal to ten per year; that is, each $1 on the average is paid out ten times a year. For the sake of simplicity there are no business enterprises; the members of the community buy and sell services from and to one another.

The quantity of money is somehow then doubled, but in such a way that no one expects the quantity to change again. Each member of the community will regard himself as better off. He now has 10.4 weeks' income in the form of cash, where as before he had only 5.2 weeks'. If everyone were to hold onto the extra cash, nothing further would happen. But people will try to spend it. One man's spending, however, is another man's receipts. All the people together cannot spend more than all the people receive. The attempt of each to do so is bound to be frustrated. In the attempt to spend more than they receive, people will simultaneously try to buy more of various services from each other and to sell less. To induce others to sell, they will offer higher prices; to induce others not to buy, they will ask higher prices. Whether the quantity sold goes up or down depends on whether the attempt to buy more is stronger or weaker than the attempt to sell less. But in either case total spending is sure to go up and so are total income and prices paid. When income has doubled, to $20,000,000, the amount of money in existence will again be equal in value to 5.2 weeks' income. The community will have succeeded in reducing its real cash balances to their former level, not by reducing nominal balances but by raising incomes and prices. The process of adjustment may not be smooth; spending may go too far and leave people with real balances that are too small, requiring a subsequent fall in the price level; but the final position will tend toward a doubling of prices, and the previous real flows of services will be resumed with no one any better off than before the new money was distributed.

This simple example embodies most of the basic principle of monetary theory: (1) the central distinction between the *nominal* and the *real* quantity of money; (2) the equally crucial contrast between the alternatives open to the individual and to the community as a whole. To each individual separately (in the hypothetical example and in the real world) it looks as if his income is outside his control, but he can determine how much cash he holds. To the community as a whole, the total amount of cash is fixed but it is able to determine the size of its income in dollars; (3) the importance of attempts. The *attempt* of people as a whole to spend more than they receive, even though doomed to frustration, has the effect of raising total nominal expenditures and receipts.

The quantity equations. A useful device for clarifying the variables stressed in the quantity theory is the so-called quantity equation. The quantity equation has taken different forms according as quantity theorists have stressed different variables. The most famous is the so-called transactions version:

$$MV = PT \qquad (1)$$

This version starts with the identity that any transaction has two sides: an amount of money paid and a corresponding quantity transferred at a certain price—for example, $10 paid for two books at $5 each. The left side of the equation sums up the amount of money paid out; the right side, the value of the transaction. Generalized to the whole economy, the left side of the equation becomes the amount of money in existence (M) multiplied by the average number of times per year (or other unit of time) that each dollar is used in effecting a transaction—the transaction velocity of circulation (V). The right side of the equation then becomes the product of an average price (P) times an *aggregate* quantity of goods (T).

In this version, all transactions of whatever kind are included—purchases of securities along with purchases of restaurant meals, purchases of the food ingredients by the restaurant as well as of the final restaurant meals. The inclusion of capital as well as current transactions, and of intermediate as well as final goods, makes the concepts of total transactions (T) and the "general price level" (P) highly ambiguous and difficult to handle statistically. The emergence of national income accounting in recent decades has led to the substitution of an income version of the quantity equation:

$$MV = Py \qquad (2)$$

In this equation y represents national income in constant prices (the total value at fixed prices of all final goods and services including additions to the stock of capital), P is a price index, and V is the average number of times that the money stock is used for making *income* transactions (that is, payments for final goods and services, or alternately, for final productive services—services of labour, land, and so forth).

In both these equations the two sides are, by definition, precisely equal to one another—that is, each equation is an identity (like the equality of total assets and total liabilities on a double-entry balance sheet). In practice, the data on transactions or on income—the right sides of the equations—come from different statistical sources than do the data on the quantity of money, and only indirect data exist for velocity. As a consequence, when these equations are used to describe the real world, V is usually computed in such a way as to make the two sides equal, thus embodying all the statistical errors of the other terms.

These versions of the quantity equation emphasize the function of money as a medium of exchange—its use in making payments. But money also serves as a temporary store of purchasing power in the interim between sale and purchase. This aspect of money is emphasized in the cash balance approach.

It has generally been supposed that the amount of money people will want to hold for this purpose bears some relation to their income. This leads to the equation

$$M = kPy \qquad (3)$$

in which M, P, and y are defined as in equation (2), and k is the ratio of money stock to income. Although equation (3) is simply a mathematical transformation of equation (2)—k being numerically equal to the reciprocal of V—it brings out the difference between the aspects of money stressed by the transactions approach and those stressed by the cash-balances approach.

In particular, the cash-balances approach fits in better with general economic analysis than does the transactions approach. Equation (3) can be regarded as describing the demand for money, with P and y being two of the variables on which the demand for money depends and with k symbolizing all the other variables. For completeness, the analysis requires another equation showing how the supply of money is determined. This approach provides a better bridge to the later Keynesian and more recent monetarist development than does the transactions approach.

Their uses. The purpose of the quantity equations is mainly to provide an analytical filing system to help sort out the forces at work. Their usefulness derives from the empirical hypothesis that each of the four main labels (M, V, P, and y for the income version) refers to a fairly distinct set of forces subject to influences that are different from those classified under another label.

Interpretation of the quantity equations

In its simplest form, the quantity theory regards output (y) as being determined by the productive opportunities of the community and its tastes and preferences; it regards the quantity of money (M) as being determined by the conditions of production of the monetary metal plus the characteristics of the financial and banking structure (for a commodity standard) or the decisions of the monetary authorities and the financial structure (for a fiat standard); it regards velocity (V) as being determined by payments practices, the financial and economic arrangements for effecting transactions, and the costs and returns from holding money instead of other assets; and, finally, it regards the price level (P) as the variable that adjusts to reconcile the conditions of production determining y with the financial conditions determining M and V.

In the most rigid form of the quantity theory, velocity is seen as very stable. This view regards the level of prices as determined directly by changes in the quantity of money, and as moving in strict proportion to the quantity of money. But the more sophisticated quantity theorists never took this simplistic view. They all recognized that prices could not adjust instantaneously and fully and that, during so-called transition periods, changes in the quantity of money would affect output as well as prices. They also recognized that V was not a constant but itself a function of other variables such as interest rates and that it was subject to substantial random perturbations. As a result, they all recognized that changes in output and prices could be produced in principle by independent changes in either V (that is, the demand for money) or M (that is, the supply of money).

On an analytical level the quantity theory of money is not the tautology of the equation; it is, rather, an analysis of the factors determining the quantity of money that the community wishes to hold. On an empirical level, it does not assume that V is constant; it holds that changes in the demand for money tend to proceed slowly and gradually or to be the result of events set in train by prior changes in supply, whereas, in contrast, substantial changes in the supply of nominal balances can and frequently do occur independently of any changes in demand. The conclusion is that substantial changes in prices or nominal income are almost invariably the result of changes in the nominal supply of money rather than of V.

An implication of the foregoing is that monetary policy —policy as to changes in the quantity of money—is the appropriate means for preventing inflation or deflation and for avoiding wide cyclical fluctuations.

The Keynesian critique. The groundwork for the Keynesian attack on the quantity theory was laid by two historical experiences. One was the British experience after the return of gold in 1925. In order to maintain the parity set for the pound in terms of gold, the Bank of England followed a persistently tight monetary policy. But prices did not move downward sufficiently, and there was persistent unemployment. This raised doubts about the allegedly classical postulate that money was only a veil, that total output was determined by real conditions, and that full employment was the "equilibrium" or normal position. The second, and even more dramatic experience, was the Great Depression. Later study has shown that the severity of the depression, certainly in the U.S., can be attributed to bad monetary policy—a policy that permitted or forced the quantity of money to decline by one-third from 1929 to 1933. This was not known or understood at the time; on the contrary, the Federal Reserve authorities insisted that they were doing everything they could to offset the forces making for depression, and that they were unable to do more.

This stimulated the British economist John Maynard Keynes—who had been a highly sophisticated quantity theorist on strictly classical lines—to develop an alternative interpretation of the debacle after 1929 and an alternative prescription for policy. It also produced a sympathetic climate for his new theory, so that the Keynesian revolution swept the economics profession in a remarkably short time.

Keynes did not, of course, deny the validity of the quantity equation. What he maintained was very different: that the right-hand side—Py or prices times output—was determined by forces largely independent of the left-hand side. P, he argued, was largely an institutional datum, linked to customary nominal rates of wages that were very resistant to change, or at least to decline. Real income or output, y, he argued, was determined by aggregate demand, which depended primarily on the amount of "autonomous spending"—that is, spending not linked directly to current income—and consisted mostly of investment expenditures by enterprises and spending by government. The rest of aggregate demand consisted mostly of consumption, which was linked closely to current income and so determined by autonomous spending. On this view, the key factor determining the level of real income and output, and thereby the level of nominal income, was the strength of investment and of government spending.

How then could the quantity equation hold? Keynes's answer was that V was so pliable that it would adjust to keep the two sides equal. Let autonomous spending increase without a change in the quantity of money and V would simply increase in response; let the quantity of money increase without an increase in autonomous spending, and V would decline to offset the increase in M.

In its extreme form—as much a caricature of Keynes's thinking as the attribution of a constant V is of his predecessors'—the quantity of money simply did not matter at all for the course of prices or real output but only for the movement of V.

How did Keynes explain the plasticity of V? He did so by means of a very special explanation of the demand for money, to which he applied the phrase "liquidity preference." He regarded the quantity of money demanded as divided into two parts, one part M_1, "held to satisfy the transactions and precautionary-motives," the other, M_2, "held to satisfy the speculative-motive." He regarded M_1 as a roughly constant fraction of income, and with respect to this part he accepted a rather rigid form of the quantity theory. He regarded M_2, however, as largely independent of income and as highly sensitive to current interest rates and to the expectations of future interest rates. This explained the plasticity of V.

According to Keynes, if current interest rates are above the level many holders of money expect to prevail in the future, people will prefer to hold their speculative balances in bonds instead of money—because they expect a fall in interest rates, which means a rise in bond prices. Conversely, if current interest rates are below the level expected to prevail in the future, they will prefer to hold money. Over short periods, he argued, these expectations are widely and strongly held at about the same level. In the extreme with everyone holding the same expectations and everyone sensitive, liquidity preference will become, as he termed it, absolute. Under such circumstances, let the monetary authorities try to in-

crease the amount of money by buying bonds. This will tend to raise bond prices and lower interest rates, but even the slightest lowering will lead speculators to absorb the additional money and sell bonds. The result will be a lower V or a higher k for the community as a whole. The converse will hold if the authorities try to reduce the amount of money.

The decline and rise of the quantity theory

Alternatively, if nominal income rises, for whatever reason, without a change in the quantity of money, that will require an increase in M_1, but it can and will come out of M_2 without any further effects. The conclusion is that *under circumstances of absolute liquidity preference* income can change without a change in M or in interest rates, and M can change without a change in income or in interest rates. The holders of money are in metastable equilibrium, like a tumbler on its side on a flat surface; they will be satisfied with whatever the amount of money happens to be.

For the long run, Keynes regarded the size of M_2 as depending on the level of interest rates for a different reason. The interest rate measures the cost of holding money instead of bonds—it is the price, as Keynes put it, that the holder of money pays for parting with liquidity. If the interest rate is low, the cost of liquidity is low; hence people will hold a large fraction of their wealth in the form of money, and the fraction will be sensitive to small changes in the interest rate—liquidity preference being absolute in the long run as well as in the short run. If interest rates were high, and M_2 low—a condition that Keynes expected to hold only at times of full employment—then the quantity theory would "come into its own," in the double sense that velocity would be constant and that prices rather than output would respond to changes in the quantity of money.

Keynes himself regarded absolute liquidity preference as a strictly "limiting case"; while it "might become practically important in future," he knew "of no example . . . hitherto." In practice, however, he treated velocity as if its behaviour frequently approximated this limiting case, and his disciples went much further than he did.

If liquidity preference was not absolute, then in the Keynesian view a change in the quantity of money would first affect interest rates in order to induce the community to hold the changed quantity of money. This change in interest rates might in turn affect autonomous spending—a decline encouraging investment and a rise discouraging investment—and in this way aggregate demand, income and employment. The more sensitive is the quantity of money demanded to interest rates, and the less sensitive is investment to interest rates, the smaller will be the final effect of changes in the quantity of money, and conversely. (For a fuller discussion of the Keynesian view see INCOME AND EMPLOYMENT THEORY.)

The monetarist view. The Keynesian revolution swept the economics profession in the late 1930s and early 1940s; by the 1960s Keynesian ideas had become deeply embedded in the thinking of informed intellectuals. The course of events, however, did not follow the pattern that many had expected. Their analysis had predicted a great postwar depression if war spending was not replaced by other government spending. But government spending fell precipitously in the U.S. and elsewhere without ill effect. The analysis of the Keynesians had called for "cheap money" policies—that is, low interest rates—with a view to stimulating investment and thereby avoiding mass unemployment. Those policies proved in the main to be unnecessary; where they were adopted they were consistently followed by inflation that could be restrained only by abandoning the cheap-money policies. A re-examination of the Great Depression of the 1930s demonstrated that, contrary to general belief, it had been a tragic testament to the great power of monetary policy, not to its impotence. The U.S. Federal Reserve System could have prevented the decline of one-third in the quantity of money that occurred from 1929 to 1933. Had it done so, the evidence indicates, the depression would have been far milder and briefer. Most important of all, extensive empirical research demonstrated that the velocity of circulation of money, far from moving to offset changes in the quantity of money, has generally moved in the same direction and reinforced the effect of these changes.

The result has been a major change in the views of today's followers of Keynes. They attribute far more importance to the quantity of money than they earlier did, regarding the quantity of money demanded as less elastic with respect to interest rates, and investment as more elastic, than they did in the first flush of enthusiasm. Some economists have been waging a counter-revolution in the name of monetary policy, to which the name monetarism has become attached. This doctrine is in an early stage of development and was still, at the beginning of the 1970s, a minority view among professional economists. It had made significant gains, however, and had influenced considerably the views even of those who still regarded themselves as Keynesians.

No counterrevolution ever produces a return to the status quo. The emerging monetarist doctrine has much in common with the earlier quantity theory, but it has not expunged the Keynesian revolution. As one monetarist has put it, Keynes's liquidity preference analysis reinforced the shift of emphasis that had already been going on "from the transactions version of the quantity equation to the cash-balances version—a shift of emphasis from mechanical aspects of the payments process to the qualities of money as an asset." His analysis, though strictly in the cash-balances tradition, was much more explicit in stressing the role of money as one among many assets, and of interest rates as the relevant cost of holding money. The recent work of others "has gone still further in this direction treating the demand for money as . . . concerned with the composition of the balance sheet or portfolio of assets." (Milton Friedman, "A Theoretical Framework for Monetary Analysis," *Journal of Political Economy*, March/April 1970.) The Keynesian emphasis on the plasticity of velocity has led to much more extensive work on the stability of the demand for money, and in particular on the exploration of time lags between monetary changes and economic changes and on the separate effects of monetary changes on prices, output, and interest rates.

Key propositions. For the long run, however, monetarism accepts almost wholly the earlier quantity theory. Its main contribution to the earlier theory is a much more detailed and sophisticated analysis of short-run effects, and a more detailed set of empirical generalizations about these effects. The key propositions of monetarism are as follows:

1. The rate of growth of the quantity of money is consistently, though not precisely, related to the rate of growth of nominal income. That is, if the quantity of money grows rapidly, so will nominal income, and conversely. The velocity of circulation, though not constant, is fairly predictable.

2. This relation is not obvious to the naked eye, mainly because it takes time for changes in monetary growth to affect income.

The time factor in monetary policy

3. On the average, a change in the rate of monetary growth produces a change in the rate of growth of nominal income six to nine months later. But this is an average.

4. If the rate of monetary growth is reduced, then about six to nine months later the rate of growth of nominal income and also of physical output will decline, but the rate of price rise will be affected very little. There will be downward pressure on prices only as a gap emerges between actual and potential output.

5. The effect on prices comes on the average, six to nine months after the effect on nominal income and output, so that the total delay between a change in monetary growth and a change in the rate of inflation averages about 12 to 18 months.

6. The above relationships are not invariable. There is many a slip between the monetary change and the income change.

7. Monetary changes affect output only in the short run—though "short run" may mean five to ten years. Over the longer run the earlier quantity theory seems to hold, and the rate of monetary growth affects only prices. What

happens to output in the long run depends on such "real" factors as the enterprise, ingenuity, and industry of the people; the extent of thrift; the structure of industry and government; the relations among nations; and so on.

8. It follows from these propositions that *inflation is always and everywhere a monetary phenomenon,* in the sense that it cannot occur without a more rapid increase in the quantity of money than in output. There are, of course, many possible reasons for monetary growth—gold discoveries, the manner in which government spending is financed, and even the manner in which private spending is financed.

9. Government spending may or may not be inflationary. It will clearly be inflationary if it is financed by creating money—that is, by printing currency or creating bank deposits—and if the resultant rate of monetary growth exceeds the rate of growth of output. If it is financed by taxes or by borrowing from the public, the main effect is that the government spends the funds instead of someone else. Government fiscal policy is extremely important in determining the fraction of total income that is spent by the government and who bears the burden of that expenditure. But it is not necessarily inflationary.

10. One of the most difficult things to explain in simple fashion is the way in which a change in the quantity of money affects income. Generally, the initial effect is not on income at all but on the prices of existing assets (bonds, equities, houses, and other physical capital). This is the liquidity effect stressed by Keynes; it occurs on the balance sheet, not on the income account. An increased rate of monetary growth raises the amount of cash people (or businesses) have relative to other assets. The holders of the excess cash will try to correct this imbalance by buying other assets. But one man's spending is another man's receipts. All the people together cannot change the amount of cash all hold—only the monetary authorities can do that. However, their attempts will tend to raise the prices of assets and to reduce interest rates. These changes will in turn encourage spending (by businessmen and others) to produce new assets. That is how the initial effect on balance sheets gets translated into an effect on income and spending. In this connection the monetarists stress a far wider range of assets and interest rates than do the Keynesians; they give importance to such assets as durable consumer goods and other real property; they regard market interest rates as only a small part of the whole complex of rates that are relevant.

11. One important feature of this mechanism is that a change in monetary growth affects interest rates in one direction at the outset, and in the opposite direction later on. More rapid monetary growth at first tends to lower interest rates. But later on, as it raises spending and stimulates price inflation, it also produces a rise in the demand for loans that will tend to raise interest rates. That is why, around the world, interest rates are highest in the countries that have had the most rapid rise in the quantity of money and also in prices—countries like Brazil, Chile, and Korea in the 1960s. Taking the opposite case, a slower rate of monetary growth at first raises interest rates but later on, as it reduces spending and price inflation, it lowers interest rates. That is why interest rates are lowest in countries that have had the slowest increase in the quantity of money and in prices—countries like Switzerland and Germany in the 1960s. This inconsistent relation between the quantity of money and interest rates explains why monetarists insist that interest rates are not a good guide to monetary policy.

12. With respect to policy, these propositions clearly imply that monetary policy is important and that what is important about monetary policy is its effect on the quantity of money, not on bank credit or total credit or interest rates. Wide swings in the rate of change of the quantity of money are evidently destabilizing and should be avoided. Beyond this, different monetarists draw different conclusions. Some conclude that the monetary authorities should make deliberate changes in the rate of monetary growth in order to offset other forces making for instability; these changes should be gradual and small and make allowance for the lags involved. Others maintain that not enough is known about the relations between changes in the quantity of money and in prices and output to assure that a discretionary monetary policy will do good rather than harm. They believe that a wiser policy would be simply to have the quantity of money grow at a steady rate over time. They contend that this would provide a stable monetary framework for economic growth without itself being a source of instability and disturbance.

Conclusion. No other subject in economics has been studied longer or more intensively than the subject of money. The result is a vast amount of documented experience and a well-developed body of theoretical analysis. The extent to which the students of monetary problems agree in their basic conclusions is concealed by the tendency of laymen to exaggerate their differences. But even among professional economists there remain important disagreements, centring mainly on empirical judgments about the stability and form of some of the relations between money and other economic magnitudes.

BIBLIOGRAPHY. Works on various aspects of monetary history include: PHILLIP CAGAN, "The Monetary Dynamics of Hyperinflation," in MILTON FRIEDMAN (ed.), *Studies in the Quantity Theory of Money* (1956); RUPERT J. EDERER, *The Evolution of Money* (1965); PAUL EINZIG, *Primitive Money in Its Ethnological, Historical, and Economic Aspects,* 2nd ed. rev. (1966); ALBERT E. FEAVERYEAR, *The Pound Sterling: A History of English Money,* 2nd ed. (1963); MILTON FRIEDMAN and ANNA J. SCHWARTZ, *A Monetary History of the United States, 1867–1960* (1963); ELGIN E. GROSECLOSE, *Money: The Human Conflict* (1934); and EARL J. HAMILTON, *American Treasure and the Price Revolution in Spain, 1501–1650* (1934, reprinted 1965).

Useful readings in monetary theory, of varying levels of difficulty, include: WALTER EUCKEN, *This Unsuccessful Age* (1951); MILTON FRIEDMAN, *A Theoretical Framework for Monetary Analysis* (1971); JOHN G. GURLEY and EDWARD S. SHAW, *Money in a Theory of Finance* (1960); HARRY G. JOHNSON, *Essays in Monetary Economics,* 2nd ed. (1969); JOHN MAYNARD KEYNES, *A Treatise on Money,* 2 vol. (1930), and *The General Theory of Employment, Interest and Money* (1936); DON PATINKIN, *Money, Interest, and Prices: An Integration of Monetary and Value Theory,* 2nd ed. (1965); and D.H. ROBERTSON, *Money,* 4th ed. (1956).

(M.Fr.)

Money Market

The functions of a money market

Every country with a monetary system of its own has to have some kind of market in which dealers in bills, notes, and other forms of short-term credit can buy and sell. The "money market" is a set of institutions or arrangements for handling what might be called wholesale transactions in money and short-term credit. The need for such facilities arises in much the same way that a similar need does in connection with the distribution of any of the products of a diversified economy to their final users at the retail level. If the retailer is to provide reasonably adequate service to his customers, he must have active contacts with others who specialize in making or handling bulk quantities of whatever is his stock-in-trade. The money market is made up of specialized facilities of exactly this kind. It exists for the purpose of improving the ability of the retailers of financial services—commercial banks, savings institutions, investment houses, lending agencies, and even governments—to do their job. It has little if any contact with the individuals or firms who maintain accounts with these various retailers or purchase their securities or borrow from them.

The elemental functions of a money market must be performed in any kind of modern economy, even one that is largely planned or socialist, but the arrangements in socialist countries do not ordinarily take the form of a market. Money markets exist in countries that use market processes rather than planned allocations to distribute most of their primary resources among alternative uses. The general distinguishing feature of a money market is that it relies upon open competition among those who are bulk suppliers of funds at any particular time and among those seeking bulk funds, to work out the best practicable distribution of the existing total volume of such funds.

In their market transactions, those with bulk supplies of

funds or demands for them, rely on groups of intermediaries who act as brokers or dealers. The characteristics of these middlemen, the services they perform, and their relationship to other parts of the financial mechanism vary widely from country to country. In many countries there is no single meeting place where the middlemen get together, yet in most countries the contacts among all participants are sufficiently open and free to assure each supplier or user of funds that he will get or pay a price that fairly reflects all of the influences (including his own) that are currently affecting the whole supply and the whole demand. In nearly all cases, moreover, the unifying force of competition is reflected at any given moment in a common price (that is, rate of interest) for similar transactions. Continuous fluctuations in the money market rates of interest result from changes in the pressure of available supplies of funds upon the market and in the pull of current demands upon the market.

BANKS AND THE MONEY MARKET

Commercial banks. Commercial banks are at the centre of most money markets, as both suppliers and users of funds, and in many markets a few large commercial banks serve also as middlemen. These banks have a unique place because it is their role to furnish an important part of the money supply. In some countries they do this by issuing their own notes, which circulate as part of the hand-to-hand currency. More often, however, it is checking accounts at commercial banks that constitute the major part of the country's money supply. In either case, the outstanding supply of bank money is in continual circulation, and any given bank may at any time have more funds coming in than going out, while at another time the outflow may be the larger. It is through the facilities of the money market that these net excesses and shortages are redistributed, so that the banking system as a whole can at all times provide the means of payment required for carrying on each country's business.

In the course of issuing money the commercial banks also actually create it by expanding their deposits, but they are not at liberty to create all that they may wish whenever they wish, for the total is limited by the volume of bank reserves and by the prevailing ratio between these reserves and bank deposits—a ratio that is set by law, regulation, or custom. The volume of reserves is controlled and varied by the central bank (such as the Bank of England, the Bank of France, or the Federal Reserve System in the U.S.), which is usually a governmental institution, is always charged with governmental duties, and almost invariably carries out a major part of its operations in the money market.

Volume of reserves

Central banks. The reserves of the commerical banks, which are continually being redistributed through the facilities of the money market, are in fact mainly deposit balances that these commercial banks have on the books of the central bank or notes issued by the central bank, which the commercial banks keep in their own vaults. As the central bank acquires additional assets, it pays for them by crediting depositors' accounts or by issuing its own notes; thus the potential volume of commercial bank reserves is enlarged. With more reserves, the commercial banks can make additional loans or investments, paying for them by entering credits to depositors' accounts on their books. And in that way the money supply is increased. It may be reduced by reversing the sequence. The central bank can sell some of its marketable assets in the money market or in markets closely interrelated with the money market; payment will be made by drawing down some of the commercial bank reserve balances on its books; and with smaller reserves remaining, the commercial banks will have to sell or reduce some of their investments or their loans. That, in turn, results in a shrinkage of the outstanding money supply. Central bank operations of this kind are called open-market operations.

The central bank may also increase bank reserves by making loans to the banks or to such intermediaries as bill dealers or dealers in government securities. Reduction of these loans correspondingly reduces bank reserves. Although the mechanics of these lending procedures vary widely among countries, all have one feature in common: the central bank establishes an interest rate for such borrowing—the bank rate or discount rate—pivotally significant in the structure of money market rates.

Money market assets may range from those with the highest form of liquidity—deposits at the central bank—through bank deposits to various forms of short-term paper such as treasury bills, dealers' loans, bankers' acceptances, and commercial paper, and including government securities of longer maturity and other kinds of credit instruments eligible for advances or rediscount at the central bank. Although details vary among countries, the touchstone of any money market asset other than money itself is its closeness—*i.e.*, the degree of its substitutability for money. So long as the institutions making use of a money market regard a particular type of credit instrument as a reasonably close substitute—that is, treat it as "liquid"—and so long as the central bank acquiesces in or approves of this approach, the instrument is in practice a money market asset. Thus no single definition or list can apply to the money markets of all countries nor will the list remain the same through the years in the money market of any given country.

THE INTERNATIONAL MONEY MARKET

Each central bank usually holds some form of reserve that is acceptable in settling international transactions. International monetary reserves are mainly gold, or "money market assets" in some country whose currency is widely used, such as the United States dollar. The monetary laws of all countries provide for the establishment of some kind of parity between their currencies and those of other countries. This parity may be defined either in terms of gold or in relation to a key currency such as the British pound sterling or the United States dollar, which in turn has a fixed parity with gold. A country maintains the "convertibility" of its currency by standing ready to buy and sell gold or other currencies in exchange for its own at prices within a fixed and rather narrow "spread" above or below the "exchange rate" for its own currency that is implied by the declared parity.

The buying and selling of currencies

Because world trade continually gives rise to various needs for payment in various currencies, an international money market must exist so that traders with an excess of one currency can use it to buy another currency for which they have a need. Within the scope of convertibility arrangements, this trading in currencies is carried out by skilled intermediaries, usually banks or specialized foreign exchange brokers and dealers. Trading in currencies is extensive both for immediate use ("spot") and for future ("forward") delivery. Quotations vary according to changes in supply and demand, over the range between the upper and lower buying and selling prices set by official parity. If no parity has been set quotations may fluctuate widely. If a currency is subject to exchange controls, there may be two or more quotations for different uses of the same nominal currency.

Changes in a country's balance of payments may affect the usefulness or prestige of its currency. A sustained and substantial balance of payments deficit (outpayments larger than inpayments), for example, will result in continuous large increases in the world supply of its currency, possibly leading to some decline in its acceptability abroad and to a loss of international monetary reserves. At the same time, an outward drain may reduce the reserves of the commercial banks (the base for the domestic money supply), unless the central bank takes off-setting action.

Since 1944 most of the countries that have domestic money markets or that play a role in the international money market have been joined together in the International Monetary Fund, which represents a pooling of part of the foreign exchange reserves (including gold) of more than 100 member countries. Drawings on the pool may be made by member countries to meet some of the reserve drains arising from balance of payments deficits and in amounts related to the quota that each has subscribed.

The internal money markets of a surprisingly high proportion of the countries of the world are quite rudimentary. The work of the money market in these countries is done largely by transfers of deposit balances, government securities, or foreign exchange among a few banks and between them and the central bank. But in nearly all such cases there is genuine discontent with the rigidity of these limited facilities and a desire to develop a structure, as well as instruments and procedures, which would provide the open-market attributes of the arrangements that have evolved in the leading countries. Several of the more fully developed money markets are described below (see also BANKS AND BANKING; EXCHANGE AND PAYMENTS, INTERNATIONAL).

THE U.S. MONEY MARKET

The Federal Reserve System

The domestic money market in the United States carries out the largest volume of transactions of any such market in the world; its participants include the most heterogeneous group of financial and nonfinancial concerns to be found in any money market; it permits trading in an unusually wide variety of money substitutes; and it is less centralized geographically than the money market of any other country. Although there has always been a clustering of money market activities in New York City and much of the country's participation in the international money market centres there, a process of continuous change during the 20th century has produced a genuinely national money market.

By 1935 the financial crises of the Great Depression had resulted in a basic revision of the banking laws. All gold had been withdrawn from internal circulation in 1933 and was henceforth held by the U.S. Treasury for use only in settling net flows of international payments among governments or central banks; its price was raised to $35 per ounce, and the U.S. dollar became the key currency in an international gold bullion standard. Domestically, the changes included legislative recognition of the primary importance of unified open-market operations by the Federal Reserve System and delegation to the board of governors of the Federal Reserve System of authority to raise or lower the ratios required between reserves and commercial bank deposits. Although about half of the 30,000 separate banks existing in the early 1920s had disappeared by the mid-1930s, the essential character of commercial banking in the U.S. remained that of a "unit" (or single-outlet) banking system in contrast to those of most other countries, which had a small number of large branch-banking organizations.

The unit banking system. This system has led inevitably to striking differences between money market arrangements in the United States and those of other countries. At times, some smaller banks almost inevitably find that the wholesale facilities of the money market cannot provide promptly the funds needed to meet unexpected reserve drains, as deposits move about the country from one bank to another. To provide temporary relief, pending a return flow of funds or more gradual disposal of other liquid assets in the money market, such banks have the privilege, if they are members of the Federal Reserve System, of borrowing for reasonable periods at their own Federal Reserve bank. At times some large banks, which serve as depositories for part of the liquid balances of many of the smaller ones (including those that are not members of the Federal Reserve System) also find that demands converging on them are much greater than expected. These large banks, too, can borrow temporarily at a Federal Reserve bank if other money market facilities are not adequate to their needs. Because these borrowing needs are unavoidably frequent in a vast unit banking system and, as a rule, do not indicate poor management, the discount rate charged by the Federal Reserve banks on such borrowing is not ordinarily put at punitive or severe penalty levels—thus, contrary to practice in many other countries, the central bank does not always maintain its interest rate well above those prevailing on marketable money market instruments. To avoid abuse, there is continuous surveillance of the borrowing banks by the Federal Reserve banks.

Along with this practice of borrowing at a Federal Reserve bank has developed the market for "federal funds." This specialized part of the money market provides for the direct transfer to a member bank of balances on the books of a Federal Reserve bank in return for payment of a variable rate of interest called the "federal funds rate." These funds are immediately available. There are transactions, too, in funds that are on deposit at commercial banks—by means of loans between banks, or through loans by one large depositor to another. Because these must be collected through a clearing process, they are usually called "clearinghouse funds."

Money market instruments. Transactions in federal funds and clearinghouse funds are further supplemented by transactions in which either kind of money is exchanged for some other liquid, money market instrument, most frequently government securities. The magnitude of the market for government securities became so great after World War II that it overshadowed all other elements of the money market. Trading in outstanding "governments" is virtually all done through dealers who buy and sell for their own account at prices which they quote on request (standing ready to "bid" for or to "offer" any outstanding issue). Most of these dealers have head offices located in New York City, but all are engaged in nationwide operations. Their transactions and the lending arrangements through which they finance their own inventories of government securities have evolved into a particularly sensitive indicator of the pressures of supply and demand on the money market from day to day. The most common form of dealer financing is the repurchase agreement, through which dealers sell parts of their inventory temporarily, subject to repurchase.

Closely interrelated, often through trading operations conducted by the same dealers, are the much smaller markets for bank drafts, bills of exchange, and commercial paper. Alongside these other markets and actually somewhat larger in outstanding volume are the markets for securities issued by various "agencies" created by federal statute, such as the Federal Home Loan banks and Federal Land banks. Another money market instrument is the negotiable time certificate of deposit (CD), issued in large volume by commercial banks, which first became significant in 1962. While the owner of a time CD cannot withdraw his deposit before the maturity date initially agreed upon, he can sell it at any time in a secondary market that is conducted by government securities dealers.

Certificates of deposit

The Federal Reserve System conducts day-to-day operations in the money market on its own initiative in order to assist the smooth working of the nation's financial machinery and to exert a general influence aimed at fostering economic growth and limiting economic instability. Its transactions include substantial outright purchases or sales of government securities, relatively small purchases and run-offs of bankers' acceptances, and a considerable volume of loans made for a few days at a time to dealers in government securities or acceptances in the form of repurchase agreements. While it is still the commercial banks as a group that have the greatest continuing need for the combined facilities of the nationwide money market, there is frequent and continuous participation by a great variety of institutional investors who channel the public's savings into various uses and who must always also make some provision for their own liquidity.

Perhaps the most unusual feature in the composition of the U.S. money market is the great importance attained by nonfinancial business concerns and local units of government since World War II. Corporate treasurers and the treasurers of many states and local political subdivisions and authorities have become so keenly sensitive to the profitable possibilities of managing their own liquid holdings instead of relying on the commercial banks as most had done formerly that this group at times provides nearly as large a part of the volatile financing needs of government securities dealers, for example, as comes from the banks. Moreover, banks outside New York City sometimes supply more of the financing needed by these

dealers than do the traditional "money market banks" in New York City. The nationwide character of the money market is also shown by the participation of nearly 200 banks in the federal-funds market—banks that are widely scattered among all Federal Reserve districts, although the bulk of all transactions is executed through facilities located in New York.

While the U.S. money market has become truly national, it still needs a final clearing centre upon which the net impact of changes in overall supply or demand can ultimately converge and where the final balancing adjustments of the market as a whole can be accomplished. In filling that need, New York City continues to be the centre of the national money market.

(R.V.R.)

THE BRITISH MONEY MARKET

The discount houses. In Great Britain the money market consists of a number of linked markets, all of them concentrated in London. The 12 specialist banks known as discount houses have the longest history as money market institutions; they have their origin in the London bill broker who in the early 19th century made the market in inland commercial bills. By selling bills through this market, the growing industrial and urban areas were able to draw upon the surplus savings of the agricultural areas. Quite early many bill brokers began to borrow money from banks in order to buy and hold these bills, instead of simply acting as brokers, and thus became the first discount houses. Since then the major assets held by the discount houses have at different times been commercial bills (first inland bills as described above and later bills financing international trade), treasury bills, and short-dated government bonds. During the 1960s there was a resurgence of the commercial bill, which finally became the discount houses' largest single class of asset, only to be overtaken later by the certificate of deposit.

Important changes were introduced into the British monetary system in 1971, but money at call with the discount houses retained its role as a reserve asset. Such is the safety and liquidity of call money that, despite the fractionally lower rate on it compared with other reserve assets, the banks hold about half of their required reserves in this form. This in turn provides the discount houses with a large pool of funds, which they invest in relatively short-dated assets, of which the most important is sterling certificates of deposit, followed by commercial bills, local authority securities, and treasury bills. This pattern of assets is greatly influenced by the fact that all call loans to the discount houses are secured loans, parcels of assets being deposited *pro rata* with the lending banks as security, and the assets held by the discount houses must therefore be suitable for use as such security.

They also need to hold a substantial proportion of assets that are rediscountable at the Bank of England in case of need, and the Bank of England limits their holding of assets other than public sector debt to a maximum of 20 times their capital resources.

Operations in money

On the liabilities side of the discount house's balance sheet, operating in call money is part of its day-to-day work. A bank that expects to make net payments to other banks during the day (for example, in settlement of checks paid by its customers to their customers) will probably call in some of its call loans, and by convention this is done before noon. Since the banks that have called in money then pay it to other banks, these other banks will have an equal amount to re-lend to the discount houses in the afternoon. Thus the discount houses can "balance their books"—borrow enough money in the afternoon to replace the loans called from them in the morning.

It is not uncommon for perhaps £100,000,000 to be called from, and re-lent to, the discount houses on an active day.

There is one main reason why the money position may not balance in this way. The British government accounts are kept with the Bank of England, which does not lend at call as other banks do. Thus net payments into or out of these government accounts will cause a net shortage or surplus of money for the discount houses in the afternoon and will tend to cause money rates to rise or fall. The Bank of England can allow such shortages or surpluses to affect interest rates, or it can offset them by buying or selling bills or by lending overnight to the discount houses at market rates. Even if the Bank of England does not act in this way to meet a shortage of funds, the discount houses are always finally able to secure the funds they need by their right to borrow from the Bank of England (the lender of last resort) against approved security at the "minimum lending rate" (the penalty rate).

Dealing in assets

On the assets side of their balance sheets, the discount houses are active dealers in a number of the assets they hold. They make the market in sterling certificates of deposit and in commercial bills, quoting buying and selling rates for different maturities. They also quote selling rates for treasury bills that they acquire at the weekly tender in competition with each other and with any other banks that may tender, including the Bank of England. Most of these other banks tender for treasury bills in order to hold them to maturity, but the discount houses sell theirs on the average when only a few weeks of the bills' 91-day life has passed. A large proportion of these bills is sold to the clearing banks, which do not tender on their own account.

The Bank of England minimum lending rate is normally determined for each week 0.5–0.75 percent above the average treasury bill rate at the previous Friday's tender. The bank, however, has the power to fix it at a different level if it so wishes, and this has been done.

Other markets. Important changes have also occurred outside the discount market described above; after the mid-1950s there was steady growth in public borrowing by local authorities. This led to an active local authority loan market conducted through a number of brokers, where money can be lent on deposit for a range of maturities from two days up to a year (and indeed for longer periods). Much more rapid was the growth after the mid-1960s of the interbank market, in which banks borrow and lend unsecured for a range of maturities from overnight upward. This market also is conducted through brokers, often firms that also operate in the local authority and other markets; a number of these firms of brokers are subsidiaries of discount houses.

In addition to the markets mentioned, there is the gilt-edged (government bond) market on the stock exchange; short-dated bonds are held by the discount houses and by banks and other money market participants, as are short-dated local authority stocks and local authority "yearling" (very short-dated) bonds. With flexibility of bank deposit rates (at least for deposits of large denomination), both banks and nonbank transactors are faced with a wide and competitive range of sterling money market facilities in London.

Finally, mention should be made of the Euro-dollar market, because London is its centre; this is an entrepôt market with a very large volume of business in U.S. dollar balances, conducted through brokers (often the same firms that operate in the sterling markets), and U.K. banks are active participants. However, owing to exchange control there has been little significant interaction between the Euro-dollar market and the U.K. domestic money market.

(R.F.G.A.)

THE MONEY MARKETS OF OTHER COUNTRIES

The Canadian money market. The Canadian money market was substantially broadened in 1954 with the introduction of day-to-day bank loans against Government of Canada treasury bills and other short-term government and government-guaranteed securities. Treasury bills of 91 days' and 182 days' maturity are issued weekly with the occasional offering of a longer maturity of up to one year. Government of Canada bonds and Government of Canada guaranteed bonds are issued at less regular intervals.

Groups involved in the money market are the following: the government, as the issuer of the securities; the Bank of Canada, acting as issuing agent for the government and as a large holder of market material; the chartered banks, as large holders and as distributors and potential buyers and sellers of bills and bonds at all times; the security dealers, as carriers of inventories and traders in such securities; and the public (mainly provincial and municipal governments and larger corporations), as short-term investors.

Treasury bills are sold by competitive tender in which the Bank of Canada, the chartered banks, and a small number of investment dealers participate. Bonds are normally issued at a price at which the yield is in line with outstanding comparable issues.

The central bank, through its tender at the weekly treasury-bill sale, active manipulation of its own bill and bond portfolio, and regulation of the money supply, has workable instruments for active monetary control. For both banks and qualified dealers, the Bank of Canada acts as lender of last resort. The rate is set slightly above the average rate of the last treasury bill auction to discourage regular borrowing.

The West German money market. In West Germany, where the money market developed strongly after World War II, transactions are to a large extent confined to interbank loans. In addition, insurance companies and other nonbank investors are also important lenders of short-term funds. Treasury bills and other short-term bills and notes from government agencies (railways and post) were gaining in importance by the 1960s, whereas in 1955 certain nonmarketable securities (the so-called equalization claims, created during the 1948 currency reform) held by the Bundesbank were transformed into short-term marketable securities in order to obtain suitable market material for the open-market operations of the Bundesbank. Banks are not used to dealing in short-term government securities between each other. They generally either hold these securities to maturity or resell them to the central bank at its buying rates, so that a true money market has not developed.

The market for commercial paper is of some significance, and dealing in it takes place from time to time between banks, especially in times of tight market conditions. Comprehensive regulations have been given through the Bundesbank about the rediscountability of the several kinds of commercial paper.

The influence of the Bundesbank on the monetary situation through open-market operations by the 1960s was greatly hampered by the vast liquidity of the banking system as a consequence of the persistence of Germany's favourable balance of payments situation.

The French money market. The French money market is fairly well established, but its size is restricted by the fact that in France currency still plays an important role in the money supply, whereas by regulations the nonfinancial private sector of the economy is excluded from dealing in the market. Banks as well as a few public or semipublic agencies working in the financial sphere and intermediaries—brokers and discount houses—constitute the market. Transactions take place in commercial paper and in treasury bills. The monetary authorities maintain a special bookkeeping system for all the treasury bills held by banks and other financial institutions, under which such bills are not represented by actual certificates but by entries in special accounts administered by the Banque de France for the treasury.

The central bank's open-market operations, which were normally limited to smoothing out disturbances in the local money market, have gained importance in recent years. Open-market transactions are effected to keep domestic money-market rates in line with international rates, in an effort to prevent unwanted capital flows. The possibilities of the central bank's influencing the monetary situation through the money market are limited to the large government needs for short-term funds, no market for long-term government borrowing being established.

(C.G.)

The Japanese money market. In Japan's rapidly growing economy the demand for funds, both short-term and long-term, has been persistently strong. Commercial banks and other financial institutions have therefore had an important role. The monetary authorities (the Ministry of Finance and the Bank of Japan) have been unwilling to allow market forces to equilibrate demand and supply in many financial markets for fear that interest rates would become excessively high. Most interest rates have been set administratively at levels high by international comparison (until the late 1960s) but lower than market forces would have dictated. Monetary policy is implemented by controls on both the availability of credit and its cost.

Under these circumstances Japan has had a very restricted money market. The market for short-term government securities is negligible; the low, pegged interest rate means that the Bank of Japan is the main buyer and that open-market operations are impossible. Transactions in commercial paper are minimal, being discouraged because they would tend to undermine the structure of interest rates and financial institutions.

Call money market in Japan

Only the call money market is well developed. It is restricted to transactions among financial institutions. The interest rate on call money has been relatively free, and persistently above most other short-term and long-term rates. Although small amounts are lent overnight, most are "unconditional loans" (repayment after one day's notice, with a minimum of two days) or "over-month-end-loans" (repayment on a fixed day the following month). The pattern of flows is rather stable, despite seasonal and cyclical fluctuations. City banks are the major borrowers; they have a strong demand for loans by large enterprises and use call funds as a major source of liquidity. Major lenders are local banks, trust banks, credit associations, and agricultural cooperatives, which collect individual urban and rural savings and are attracted by the high yields, liquidity, and low risk of call loans relative to other uses. Call brokers help make a market, though most funds flow directly from one financial institution to another. About three-quarters of the funds flow through the Tokyo market, and there are also call markets in Ōsaka and Nagoya.

Money markets in developing countries. Well-developed money markets exist in only a few high-income countries. In other countries money markets are narrow, poorly integrated, and in many cases virtually nonexistent. Despite the many differences among countries, one can say in general that the degree of development of a country's financial system, including its money markets, is directly related to the level of its economy. Most very-low-income countries have limited financial systems in which money markets play no role. In many former colonies, notably in Africa, expatriate commercial banks had substituted for a local money market; the banks met fluctuations in loan demand by changing their balances at head offices in London or elsewhere. More recently, government policies have encouraged these banks to develop domestic channels for temporary surpluses and deficits. Persistent inflation has been another factor inhibiting the growth of money markets in developing countries, notably in Latin America.

Most developing countries, except those having socialist systems, have the encouragement of money markets as a policy objective, if only to provide outlets for short-term government securities. At the same time many of these governments pursue low-interest-rate policies in order to reduce the cost of government debt and to encourage investment. Such policies discourage saving and make money market instruments unattractive. Nevertheless, a demand for short-term funds and a supply of them exist in all market-oriented economies. In many developing countries these pressures have led to "unorganized money markets," which are often highly developed in urban areas. Such markets are unorganized because they are outside "normal" financial institutions; they manage to escape government controls over interest rates; but at the same time they do not function very effectively because interest rates are high and contacts between localities and among borrowers and lenders are limited. In all developing countries traditional forms of moneylending continue, particularly for agriculture and small enterprise.

(H.T.P.)

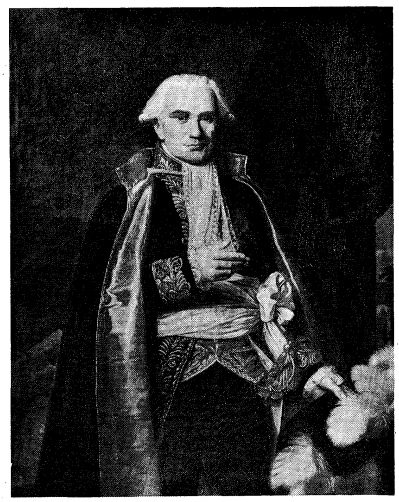
BIBLIOGRAPHY. For the United States money market, see ROBERT V. ROOSA, *Federal Reserve Operations in the Money and Government Securities Markets* (1956); and MILTON FRIEDMAN and ANNA J. SCHWARTZ, *A Monetary History of the United States, 1867–1960* (1963). For Great Britain, W.M. SCAMMELL, *The London Discount Market* (1968), is a good general treatment, while R.S. SAYERS, *Gilletts in the London Money Market 1867–1967* (1968), is a useful case study. See also the bibliography in D.R. CROOME and H.G. JOHNSON (eds.), *Money in Britain, 1959–1969* (1970). German works include HELMUT LIPFERT, *Der Geldmarkt* (1962); and JOACHIM VON SPINDLER, *Geldmarkt, Kapitalmarkt, internationale Kreditmärkte* (1960). The best single source for Japan is HANNAN EZEKIEL, "The Call Money Market in Japan," *International Monetary Fund Staff Papers*, 13:26–49 (1966). HUGH T. PATRICK, *Monetary Policy and Central Banking in Contemporary Japan* (1962), includes and explanation of the nondevelopment of money markets in Japan, and an analysis of the instruments of monetary policy in that environment. THE BANK OF JAPAN, *The Japanese Financial System* (1969), is a brief description of financial institutions, financial markets, and characteristics of the financial structure. For undeveloped countries, SATYENDRANATH SEN, *Central Banking in Undeveloped Money Markets*, 3rd ed. (1961), is a sensible, general study, but rather dated. A good discussion of policy is EDWARD NEVIN, *Capital Funds in Underdeveloped Countries* (1961), although in some respects it is out of date. JOSEPH O. ADEKUNLE, "The Demand for Money: Evidence from Developed and Less Developed Countries," *International Monetary Fund Staff Papers*, vol. 15 (1968), is a good theoretical and empirical paper relating money both to other financial assets and to real assets.

(H.T.P./R.V.R./R.F.G.A./C.G.)

Monge, Gaspard

The French mathematician Gaspard Monge founded the techniques of descriptive geometry and pioneered the development of analytical geometry, both of which have since become part of projective geometry (*q.v.*). As a prominent figure during the French Revolution, he helped to found the École Polytechnique, where his lectures and textbooks substantially advanced mathematics at the beginning of the 19th century.

By courtesy of the Musee des Beaux-Arts de Beaune, France



Monge, detail of an oil painting by Jean Naigeon, 1811. In the Musée des Beaux-Arts de Beaune, France.

Born on May 10, 1746, at Beaune, Monge was educated at the college of the Oratorians at Beaune and at their college at Lyon, where for a time at age 16 he was a physics teacher. During a vacation visit to Beaune in 1762, he made a large-scale plan of the town, devising methods of observation and constructing the necessary surveying instruments. Impressed with the plan, a military officer recommended him to the commandant of the aristocratic École Militaire de Mézières, where he was accepted but as a draftsman only; because his father was an itinerant peddler, he could not be admitted to the commission program.

A further opportunity to display his skill as a draftsman occurred when he was required to work out from data the proper locations of gun emplacements for a proposed fortress. At that time such an operation could be performed only by a long arithmetic process, but Monge, substituting for this a geometrical method, obtained the result so quickly that the commandant at first refused to receive it. On careful examination, Monge's plan was considered of sufficient value that it was guarded as a military secret until 1795, when he was finally allowed to describe it in his lectures. Continuing his researches at Mézières, Monge developed his general method of applying geometry to problems of construction; this subject later became known as descriptive geometry.

Development of descriptive geometry

Between 1768 and 1783 Monge taught physics and mathematics at Mézières; in 1777 he married the widow Madame Horbou, and in 1780 he was appointed to a chair of hydraulics, held concurrently, at the Louvre in Paris. The same year he was elected an associate of the Académie des Sciences. In a paper discussing the problem of earthworks in 1781, "Sur la théorie des déblais et des remblais" ("On the Theory of Excavation and Emplacements"), he used calculus to determine the curvature of a surface. His memoir stated the ordinary differential equation of curvature and established the general theory.

Leaving Mézières permanently about 1783, Monge became increasingly active in public affairs in Paris. Between 1783 and about 1789 he was an examiner of naval cadets; he served on the committee of weights and measures that established the metric system in 1791; from 1792 to 1793 he was Minister for the Navy and Colonies and had occasion to welcome the young artillery officer who became the emperor Napoleon; and in 1795 he participated in the founding of the Institut National de France. Although at times during the Revolution his position was precarious, Monge continued to be influential. When an appeal was made to scientists to assist in producing materials for national defense, he supervised foundry operation and wrote factory handbooks, *Description de l'art de fabriquer les canons* (*The Art of Manufacturing Cannon*) and *Avis aux ouvriers en fer sur la fabrication de l'acier* (*Advice to Iron Workers in Steel Production*). In 1794–95 he taught at the short-lived École Normale, where he was given permission for the first time to lecture on the principles of descriptive geometry he had developed at Mézières.

Particularly important for mathematics was his substantial role in the founding in 1795 of the École Polytechnique, originally for training engineers; he was an administrator and an esteemed teacher of descriptive, analytical, and differential geometry. Since no texts were available, his lectures were edited and published for student use. Among them were two important texts in 1799: in his *Géométrie descriptive*, based on his lectures at the École Normale, he developed his descriptive method for representing a solid in three-dimensional space on a two-dimensional plane, by drawing the projections—known as plans, elevations, and traces—of the solid on a sheet of paper; his text *Feuilles d'analyse appliquée à la géométrie* established the algebraic methods of three-dimensional geometry. Engineering design was revolutionized by his new procedures. Moreover, mathematics education was significantly advanced by his successful texts and popular lectures.

Teaching and writing

In 1796 Monge officially met Napoleon and henceforth became his confidant and one of his closest friends. Napoleon sent him on a commission to Italy to choose the paintings and statues that were taken to help finance his military campaigns; many of these works of art went to the Louvre. From 1798 to 1801 he accompanied Napoleon to Egypt and in Cairo helped to establish in 1798 the Institut d'Égypt, the cultural organization that was patterned after the Institut de France.

With the fall from power of Napoleon in 1814, the Bourbons deprived Monge, as a Bonapartist, of all his honours and excluded him in 1816 from the list of members of the reconstituted Institut.

He died at Paris on July 28, 1818, following a stroke.

BIBLIOGRAPHY. Biographies of Monge include LOUIS DE LAUNAY, *Un Grand Français: Monge, fondateur de l'École*

polytechnique (1933); RENE TATON, *Gaspard Monge* (1950) and *L'Oeuvre scientifique de Monge* (1951); and PAUL V. AUBRY, *Monge, le savant ami de Napoléon Bonaparte, 1746–1818* (1954)—all in French. Shorter studies of Monge's life and work are also found in ERIC T. BELL, *Men of Mathematics* (1937, reprinted 1961); JULIAN LOWELL COOLIDGE, *A History of Geometrical Methods* (1940); GERHARD KOWALEWSKI, *Grosse Mathematiker* (1938); and NIELS NIELSEN, *Géomètres français sous la Révolution* (1929).

Mongolian People's Republic

The Mongolian People's Republic is a vast but sparsely populated highland country lying at the heart of eastern Asia, between the Soviet Union on the north and the People's Republic of China on the south. Its area is 604,000 square miles (1,565,000 square kilometres)—more than twice the size of Turkey, four times that of Japan. Its shape is that of an elongated oval, measuring almost 1,500 miles from west to east—a distance as far as that from Paris to Moscow—and, at its maximum, over 780 miles from north to south. The Soviet frontier accounts for just over 39 percent of its total boundary, the Chinese frontier for the remainder.

Mongolia has an average altitude of almost 5,200 feet above sea level but has a remarkable variety of scenery. The highest point is a towering 14,311 feet (4,362 metres), at the summit of Mönh Hayrhan Uul in the Mongol Altayn Nuruu (Mongolian Altai), one of the three great mountain ranges that thrust into the republic from the north and west. Some of these mountain areas are capped with eternal snow and ice, and their forest-clad slopes alternate with lake-filled basins and valleys cut out by tempestuous rivers. The east of Mongolia is a vast, hilly plain; and to the south lies the plateau of the Gobi, a desert covering a third of the country. Desert conditions proper are found in only about a third of the Gobi, the rest being oasis-dotted semidesert grassland; indeed, a typical Mongolian landscape is one with an undulating steppeland providing fine pastures.

The extent and character of the contemporary country

The Mongolians, who lent their name to one of the three great traditional racial groupings of the globe (the other two refer to peoples of Caucasoid and Negroid stock), have a long prehistory and a most remarkable history. Their ancestors were the Huns, who lived in Central Asia from the 3rd to the 1st century BC. A single Mongolian feudal state was eventually formed from nomadic tribal groupings in 1206, and its leader, Genghis Khan, and his successors plundered and conquered much of Eurasia. By 1691—weakened by internecine strife among the khans and the effects of foreign plundering—the Mongolian state fell prey to Manchurian conquerors. An independent feudal monarchy did not emerge until 1911. Stimulated by the influence of the October Revolution of 1917, Damdiny Sühbaatar, the modern national hero, organized militant antifeudal underground forces that began to drive out both White Russians and Chinese. Sühbaatar's forces achieved power on July 11, 1921, traditionally the founding date of the present state; and the Mongolian People's Republic was officially proclaimed on November 26, 1924, with its capital at Ulaanbaatar (Red Hero).

In 1918 Mongolia had a population of just under 650,000. Although endowed with a rich folk culture, it was unquestionably one of the most backward countries of Asia. The economy was based on centuries of nomadic cattle breeding; modern industry was nonexistent; secular feudal lords and the 100,000 or so religious leaders known as lamas owned 80 percent of the cattle; Chinese usurers and other foreign financial interests had a stranglehold on trade and credit; only 1 or 2 percent of the populace was literate; and mortality rates and health conditions generally were so bad that the population was in danger of extinction. By the 50th anniversary celebrations in 1971, the population had almost doubled, to 1,250,000; an intensive development program had brought industry's share of the national income up to nearly a quarter; agriculture had been transformed by the introduction of Socialist agricultural cooperatives and state farms; and a rapid development of cultivation enabled the country to become a net exporter, rather than importer, of grain. Urbanization was increasing, and Ulaanbaatar alone had a population of over 280,000, and half of its people were under age 20. Social services and cultural life, it was generally acknowledged, had also shown remarkable improvements. Mongolian authorities believe that their country's difficult and historically short transformation, aided by the Soviet Union and other Socialist countries, offers—notably to other nations of the developing world—an early model for development directly from a feudal to a Socialist society, bypassing capitalism. Mongolian plans for the middle and late 1970s called for a further development of industry until it became the predominant factor in the economy. Contemporary Mongolia has thus developed a marked and ever-changing personality of its own, with ancient traditions and an increasingly modern way of life often existing side by side. (The following article surveys life in contemporary Mongolia; for a historical survey, see INNER ASIA, HISTORY OF and MONGOLS; for related information on physical features, see GOBI [DESERT] and ALTAI MOUNTAINS; for geological evolution, see ASIA; see also ASIAN PEOPLES AND CULTURES; CENTRAL ASIAN PEOPLES, ARTS OF; ALTAIC LANGUAGES; and BUDDHISM, HISTORY OF).

(Sh.Nj./D.Go./D.Ds.)

THE LAND

Topography and drainage. Mongolia can be divided into three major topographic zones: the fingerlike mountain ridges that thrust into the northern areas; the basin areas that lie between and around them; and the enormous plateau belt that sweeps across the southern sector.

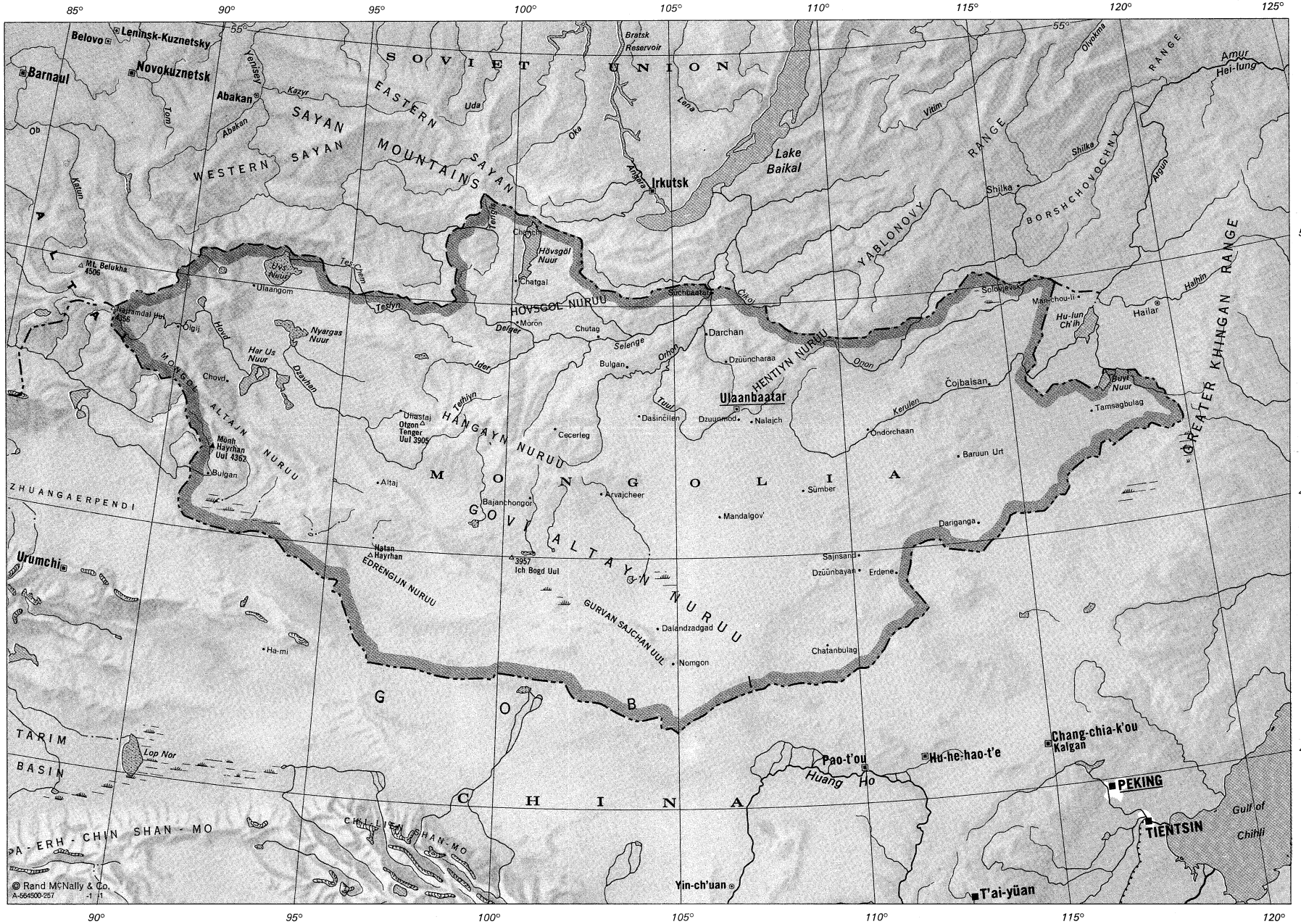
MAP INDEX

Cities and towns

Altaj, see Altay
Altay.....46·25n 96·20e
Arvajcheer, see Arvayheer
Arvayheer.....46·15n 102·45e
Bajanchongor..46·10n 100·45e
Baruun-Urt.....46·40n 113·12e
Bulgan.....44·05n 103·39e
Bulgan.....46·53n 91·05e
Cecerleg, see Tsetserleg
Chanch.....51·30n 100·40e
Chatanbulag....43·11n 109·10e
Chatgal, see Hatgal
Chovd, see Hovd
Choybalsan.....48·34n 114·50e
Chutag, see Hutag
Cojbalsan, see Choybalsan
Dalandzadgad..43·35n 104·30e
Darchan, see Darhan
Darhan.....49·28n 105·56e
Dariganga.....45·21n 113·38e
Dašinčilen.....47·51n 104·03e
Dzüünbayan....44·26n 110·03e
Dzüüncharaa...48·52n 106·28e
Dzuunmod.....47·45n 106·58e
Erdene.....44·40n 111·05e
Hatgal.....50·26n 100·07e
Hovd.....44·38n 102·30e
Hutag.....49·23n 102·43e
Mandalgovi.....45·45n 106·20e
Mörön.....48·15n 100·23e
Nalajch, see Nalayn
Nalayh.....47·45n 107·13e
Nomgon.....45·27n 105·05e
Ölgij, see Ölgiy
Ölgiy.....48·59n 92·01e
Önclörhaan.....47·19n 110·39e
Öndörchaan, see Önclörhaan
Sajnšand.....44·55n 110·11e
Süchbaatar, see Sühbaatar
Sühbaatar.....50·17n 106·10e
Sümber.....46·21n 108·25e
Tamsagbulag...47·16n 117·17e
Tsetserleg.....47·30n 101·27e
Ulaanbaatar....47·55n 106·53e
Ulaangom.....49·58n 92·02e
Uliasta, see Uliastay
Uliastay.....47·45n 96·49e

Physical features and points of interest

Altai, *mountains*.....49·00n 87·30e
Buyr Nuur, *lake*.....47·48n 117·42e
Delger, *river*....49·17n 100·40e
Dzavhan, *river*..48·54n 93·23e
Edrengiyn Nuruu, *mountain range*.....44·15n 97·45e
Gobi, *desert*.....43·00n 106·00e
Govĭ Altayn Nuruu, *mountains*.....44·00n 102·00e
Gurvan Sayhan Uul, *mountains*.....43·50n 103·30e
Hangayn Nuruu, *mountains*.....47·30n 100·00e
Har Us Nuur, *lake*.....48·00n 92·10e
Hatan Hayrhan, *mountain*.....44·55n 96·12e
Hentiyn Nuruu, *mountains*.....48·30n 108·30e
Hovd, *river*.....48·06n 92·11e
Hovsgol Nuruu, *mountains*.....49·14n 101·40e
Hövsgöl Nuur, *lake*.....51·00n 100·30e
Hyargas Nuur, lake.....49·12n 93·24e
Ider, *river*.....49·16n 100·41e
Ih Bogd Uul, *mountain*.....45·00n 100·15e
Kerulen, *river*..48·10n 115·30e
Mongol Altayn Nuruu, *mountains*.....47·00n 92·00e
Mönh Hayrhan Uul, *mountain*..46·50n 91·40e
Mönh Sarĭclag, Gora, *mountain*.....51·45n 100·32e
Najramdal Uul, *mountain*.....49·10n 87·52e
Onon, *river*.....49·35n 112·30e
Orhon, *river*....50·21n 106·05e
Otgon Tenger Uul, *mountain*..47·36n 97·36e
Sayan, *mountains*.....52·45n 96·00e
Selenge, *river*...50·25n 106·01e
Tengis, *river*....51·27n 99·02e
Terhiyn, *river*...48·08n 99·36e
Tesiyn, *river*....50·28n 93·04e
Tuul, *river*.....48·57n 104·48e
Uvs Nuur, *lake*..50·20n 92·45e



MONGOLIAN PEOPLE'S REPUBLIC

Size of symbol indicates relative size of town • ⊙ ▣ ▣ ■

Elevations in metres

The mountains. The present relief is due to the geologically recent upheavals of the Alpine mountain-building period. There are three major mountain belts. The highest and the longest spine of mountains is the westernmost, the Altai, which sweeps in from the northwestern tip of the country and thrusts toward the southeast for 1,000 miles. The main range—the only one where contemporary glaciation has developed—is known as the Mongol Altayn Nuruu, and a lesser range splitting off to the southwest is known as the Govĭ Altayn Nuruu (Gobi Altai). The southeastern extremities of the main range also split into a number of smaller hills, all following the same general trend, losing themselves in the expanses of the Gobi. The highest peaks of the Altai are in the west and north. The terrain further south and east is less rugged.

The three major mountain belts

The Hangayn Nuruu (Khangai Mountains), also trending northwest to southeast, forms a solid mountain mass near the centre of the republic, with peaks towering to over 12,000 feet, culminating in Otgon Tenger Uul, 12,811 feet (3,905 metres). A characteristic feature is the gentle slopes and crests, often covered with fine pastures. The higher central portions are nevertheless rugged and precipitous. To the north, the Hövsgöl Nuruu (Khubsugal Mountains), adjoining the vast Sayan Mountains that curve away beyond the Soviet border, have a similar topography.

The third mountain block, the Hentiyn Nuruu (Khentei), is smaller and lower, exhibiting in smaller compass some of the isolated grandeur of the Hangayn Nuruu. The Hentiyn Nuruu lies northeast of Ulaanbaatar, trending southwest to northeast, and reaches heights of between 6,500 and 8,000 feet. The enormous Great Khingan Mountains rise along and beyond the eastern frontier with China.

The northern intermontane basins. Around and between the main ranges mentioned lie an important series of basins, often choked with accumulations of material swept down from nearby mountains. In the northwest of the country, tucked between the Altai, the Hangayn, and the mountains of the Soviet frontier, lies a scenic basin complex known as the Great Lakes region, in which are strewn more than 300 lakes. Another basin complex lies between the eastern slopes of the Hangayn Nuruu and the western foothills of the Hentiyn Nuruu. The southern portion (Tuul Gol–Orhon Gol Basin) is a fertile region important in Mongolian history as the cradle of settled ways of life, and its landscapes are strewn with the ruins of ancient communities. Further north, on the northern flanks of the Hangayn Nuruu, lies the remarkable Khorgo region, in which as many as half a dozen extinct volcanoes and numerous volcanic lakes are found in quite a small area. Swift and turbulent rivers have cut jagged gorges. At the source of the Orhon Gol there is another volcanic region, with a cluster of lakes, deep volcanic vents, hot springs, and the constant roar of the Tsagan-Azraga Mörön (White Stallion River), almost invisible in its deep gorge. Near the northern border, Hövsgöl Nuur (Lake Khubsugal) is the focus of another rugged, lake-strewn region, noted for its huge subterranean caves.

The plateau and desert belt. The eastern part of Mongolia has a rolling topography of hilly steppe plains, supplanted in the extreme east by clusters of small, flat plains lying at altitudes of 2,000 to 2,300 feet. Here and there, small, stubby massifs contain the clearly discernible cones of extinct volcanoes. The Dariganga area of Mongolia's eastern tip contains some 220 such extinct volcanoes. Most of the southern part of the country is a vast rolling oasis-dotted plain, forming the northern fringe of the Gobi, which is predominantly stony, with sands covering only about 3 percent of the terrain. The flat relief is occasionally broken by low, heavily eroded ranges. Several spectacular natural features are found in the Gobi region. Huge basalt columns, arranged in clusters resembling bundles of six-sided pencils, are found in the eastern and central regions. The south Gobi contains three mountain ranges known as Gov Gurvan Saikhan (Three Gobi Beauties) and the famous Yelyn Am (Valley of Condors), now a national park, with deep gorges surrounded by the towering rocky cliffs where the condors have made their nests. This well-watered region offers a startling contrast to the surrounding steppes, which seemingly extend endlessly, covered with stone fragments and brown patches of withered grass.

Features of the Gobi

Rivers and lakes. Throughout the whole country, many streams and rivers, some of which are seasonal, run into salt lakes or the stones and sand of the desert areas. In the northern regions, the turbulent mountain streams and small rivers merge into quite deep, well-developed rivers. In the southern areas, where there are only a few constantly flowing rivers, numerous lakes, saltwater and freshwater springs, and wells draw on subterranean water supplies. The great divide, separating waters that flow into the Arctic and Pacific oceans and into the interior basins of Central Asia, runs along the crests of the Hentiyn, Hangayn, and, in part, the Altai mountains. Mongolia's greatest but third longest river, the Selenge Mörön (Selenga), with its tributaries, the Delger Mörön, Ider Gol, Orhon Gol, and others, drains northward toward Lake Baikal, crossing the Soviet border alongside the Trans-Mongolian Railway. The Mongolian portion of the Selenge Mörön Valley, measuring 109,000 square miles, is in the central northern portion of the country. The second major river, the Onon, and the Kerulen and the Halhïn Gol run latitudinally across the eastern part of the country, finally emptying into the Pacific. The largest river draining to the Central Asian interior is the Hovd (Kobdo), which rises from the glaciers of the Altai and drains part of the northwest corner of the country. A series of rivers runs down from the southern flanks of the Hangayn, ending in salt lakes or disappearing in the Gobi expanses. Generally, the Mongolian rivers are swift and with a steep gradient, offering considerable hydroelectric potential, as well as occasional flood dangers.

There are over 3,000 lakes in Mongolia with an area of half a square mile or more. Hövsgöl Nuur, already mentioned, the deepest lake (1,781 feet [238 metres]) in Central Asia, occupies a structural depression. Many of the lakes are of volcanic origin, as is Terhiyn Tsagaan Nuur on the northern Hangayn slopes. A thick layer of lava once blocked the waters of the Terhiyn Gol to form this 24-square-mile lake, lying at an altitude of 6,760 feet (2,060 metres). In the lake a small island is a famous bird sanctuary. Bayan Nuur, on the western Hangayn slopes, one of the country's loveliest lakes, fringed by golden sand dunes and rich in wild life, is associated with a romantic legend of a former city reputed to have vanished into the depths, a tale given some support by the archaeological remains dredged from the lake. Mongolia also has a number of medicinal springs, one of the best known being associated with Hatan Hayrhan, a heart-shaped massif rising from sand dunes in the western Gobi. The medicinal spring flows from a crevice in its side and has, over the centuries, worn the seven deep basins known locally as the Kettles of Hatan Hayrhan.

Volcanic lakes

Earthquakes. Mongolia is prone to seismic movements, with some earthquakes reaching extreme limits of severity, although their effects are fortunately diminished by the low population density.

Climate. Situated far from the moderating influences of the world's oceans, Mongolia experiences a markedly continental climate, with cold winters and hot summers. The dry mountain air and the tempering of the heat by rainfall, which falls mainly in the summer, nevertheless ameliorate the more extreme conditions. Weather is often unpredictable, and the sudden downpours of the summer months can occasion severe flooding. Spring meltwater, too, can swell the rivers: the spring ice run on the Tengis Gol, which runs through the Hövsgöl area, is an awesome spectacle, with its thunderous cannonading of breaking ice, swept down through gorges by the pressure of meltwater. Heavy snow occurs only in the mountain regions, but fierce blizzards of almost indescribable severity can sweep across the steppes, and even a thin coating of snow can prevent animals from getting at their pasture. One famous feature of the climate is the number of clear sunny days, averaging between 220 and 260 each year. Climatic statistics reveal a contrast between the northern and southern regions. In the Hövsgöl and Hentiyn moun-

tains, the mean annual temperature is about 28.6° to 25.7° F (−1.9 to −3.5° C), and the annual precipitation is from 16 to 20 inches; in the Gobi region, the annual precipitation is not more than 4 to 5 inches and only half this, or less, in some areas. The mean annual temperature is 39° F (4° C). There is great variation between the coldest and hottest months. Mean January temperatures in the north fall to −29° F (−34° C), and in the Gobi to −2.2° F (−19° C), while the corresponding July figures are 59° F (15° C) and 73° F (23° C), respectively. Temperature variations of as much as 55° F (30° C) may occur in a single day, and sand or hail storms may spring up almost without warning. (Sh.Ts.)

Plant and animal life. *Vegetation zones.* Climatic conditions are reflected in soil and vegetation patterns. Soils are predominantly of the chestnut or brown type, but latitudinal and altitudinal belts of vegetation, often changing rapidly over a few miles, are probably the most obvious feature of the local Mongolian landscape. There are four basic divisions running latitudinally from north to south: the forest–steppe zone of the mountains; the mountain–steppe zone; and, in the south and extreme south, respectively, the semidesert and desert zones. In addition, the mountain areas are characterized by two successive altitudinal belts: the mountain–taiga and the alpine zones. The steppes predominate, covering more than three-quarters of the national territory.

The importance of latitudinal and altitudinal belts

The mountain forest–steppe zone, with the two superimposed altitudinal belts mentioned, covers about 30 percent of the country and exhibits the richest diversity of plant and animal life. Forests grow thickest on the northern shady slopes, the most widespread tree being the Siberian larch, followed by the cedar, with a varying admixture of spruce, pine, and fir. Deciduous trees include the birch, aspen, and poplar. Steppe vegetation is found in the intermontane basins, the wide river valleys, and the sunny southern flanks of the mountains. These huge expanses of pastureland are covered with feather grass, couch grass, wormwood, and many fodder plant species. In summer, the steppes are carpeted with a magnificent covering of bright violet, blue, red, and yellow flowers. On the highest mountain slopes, the damp, dark, coniferous forest of the taiga gives way to the thin grasses and occasional flowers of the alpine zone, merging into the bare rocks and rugged glaciers of the summit zone.

Semideserts are found in the Great Lakes intermontane depression and, in general, over the Gobi, giving way to true desert conditions near the southern border. Vegetation is scarce there but heavy enough to feed camels, goats, and sheep. Tracts of saksaul (xerophytic [adapted to very dry conditions] vegetation) and groves of elm and poplars cluster around springs or other underground water sources. Only in the far south is there a narrow belt of the Gobi. This is a typical hammada, or rock-floored desert with gravel cover: only the extreme east has small areas of sandy desert.

Animal life. The varied natural conditions, the interior location, and the sparse human population of Mongolia all contribute to a very rich and diverse wildlife that has attracted international attention and has great commercial importance. Lying on the borders of several distinct zoogeographical regions (the Tibetan, the Afghano-Turkistani, the Siberian, and the North-Chinese-Manchurian), the country has a fauna combining species from each of them. The northern forests harbour lynx, maral (Asiatic red deer), elk, roe deer, musk deer, brown bears, snow leopards, wolverines, boars, squirrels, and sables. The steppes are the home of, among others, the marmot—whose pelts are very important economically—and the lithe Mongolian gazelle. Clustering around water holes in the semidesert and desert region may be found the wild sheep known as *argal*, *kulans* (wild asses), Przewalski's wild horses (*takhs*), wild camels (*khavtgays*), and the Gobi bear (*mazalai*), some of which are extremely rare and found nowhere else in Asia. Domesticated animals include sheep, camels, cattle, the hairy highland yak, goats, dogs, and the famous Mongolian horses. Bird life includes larks, partridges, cranes, bustards, and falcons in the steppes; geese, ducks, gulls, pelicans, swans, and cormorants in the rivers and lakes; the snowy owl, the golden eagle, and the condor, which frequent some areas. The freshwater lakes and rivers harbour some 60 fish species, including salmon, trout, grayling, perch, and pike. Hunting and fishing, for sport and for commercial purposes, are important; but the government has introduced stringent hunting regulations and other conservationist measures, including the establishment of national parks and nature reserves.

(D.Ba./O.Sh.)

The human imprint. *Rural patterns.* The hand of man has long been in evidence in this part of what is sometimes called High Asia; and the ruins of long-abandoned cities, as well as archaeological remains dating back to the earliest days of prehistory, have attracted the attention of Mongolian and international scholars. The huge rural areas of the country have their own charm. The steppelands and other pasture regions cover, in all, 83 percent of all Mongolian territory, and almost two-thirds of the population are livestock breeders and farmers, although a more settled way of life has gradually replaced the old nomadic traditions. A distinctive feature of the countryside is the *gher*, or *yurt*, the traditional Mongolian dwelling. It is a cone-shaped, latticed structure, light, strong, and easy to assemble, transport, and reerect, which offers warmth in winter and coolness in summer. It is still used by herdsmen moving from pasture to pasture, and the clusters of white felt and canvas cones against a green background still dominate the landscape in many areas. As the agricultural population becomes more settled, however, the *yurt* clusters are becoming associated with some 300 enormous cooperatives and state farms, often at centres with more permanent dwellings. The sinking of mines, extension of communications, the development of virgin lands, and extensive irrigation projects were all altering the face of the countryside in the 1970s.

The *gher*

Urban patterns. It is in the cities, however, that Mongolia presents its modern aspect. By the 1969 census, the population defined as urban had reached almost 530,000 (compared with a rural total of 670,000). Ulaanbaatar, with over 280,000 people, is most important by far among the urban centres. It lies on the banks of the Tuul Gol (Tula River) in the north central portion of Mongolia. Formerly known as Urga or Ikh Khuree, it was originally a nomadic princely residence and became settled on the present site in 1639. The old city—which numbered some 60,000 people in 1921—consisted mainly of monasteries, a few adobe structures, and clusters of *yurts*. A number of sections divided according to class and nationality showed marked contrasts between rich and poor areas. By the 1970s, however, the "city of felt" had become transformed into a modern industrial city, with broad avenues and streets, apartment complexes, and massive governmental, cultural, and educational buildings. The *yurt* areas and the fenced-off small houses were a fast diminishing feature of the peripheral regions. A vast urban planning scheme, initiated in 1961, is scheduled to run to 1980, by which time it is planned to divide the capital into 19 districts, each with a modern central complex of service and cultural centres. The contemporary city thus gives the impression of a giant construction site. It grew by 20 percent in the 1965–70 period alone.

Growth of Darhan

The city of Darhan, lying on the Trans-Mongolian Railway link to the north of Ulaanbaatar on the way to the Soviet frontier, is a spectacular example of planned urbanization. Its foundation stone was laid in 1961 in an existing settlement of 1,500 people. By 1970 its youthful population (three out of four people were under 40) had passed the 23,000 mark and was growing rapidly as the new city became the centre of a major industrial complex, second only to Ulaanbaatar itself. Much of the construction was achieved with the help of specialists and workers from other countries of the Socialist bloc, some from as far away as Hungary. Sühbaatar, founded in 1937, is the third largest industrial centre. Choybalsan, in the east, is another new industrial complex. Dzüün-bayan, in the eastern Gobi, is an example of a new town that has sprung up as a result of mining and oil-drilling activi-

ties. A steady growth is also taking place in other, smaller, urban centres, one or more of which are to be found in each of the 18 *aymags*, or provinces. By the 1969 census, the urban percentage in each *aymag* ranged from a low of 15 to a high of 58.9, and further incursions of modernity were expected to continue.

THE PEOPLE

Ethnic, religious, and social background. Anthropologically, the Mongolians are quite homogeneous, with nine out of 10 belonging to the classic physical type to which they lent their name. In appearance, they are of short, stocky build, with yellow to dark-brown skin colour and dark hair, and with eyes characteristically slanted. Ethnically, Kahlka-speaking Mongolians predominate, with just over 75 percent of the total population. Mongolians also form a minority in neighbouring countries, notably in the Inner Mongolian region of China and also in parts of the Soviet Union. The modern republic is also the home of over 60,000 Kazakhs (about 5 percent of the total population) and smaller numbers of Durbet, Buryat, Dariganga, Uryankhai, Dzakhchin, Torgut, Ould, Russian, and other peoples.

The theocratic background

At the time of the founding of the modern state, the social composition was strongly influenced by the then prevailing religious traditions of the lamas, who followed tenets derived from Tibetan Buddhism, with a strong admixture of more primitive elements. About 100,000 monks of all classes (about 40 percent of the entire male population), 20,000 concentrated in the capital alone, inhabited some 700 monasteries and were bound by vows of celibacy. The secular and clerical feudal lords owned half the national wealth (although comprising only about 8 percent of the population); and control lay in the hands of the head of the Mongolian Tibetan Buddhist Church (who was proclaimed the khan of all Mongolia), together with various local khans, hundreds of princes and noblemen, and the higher clergy. The whole theocratic system, it is generally acknowledged, was in a state of extreme backwardness—made worse by various internecine struggles—and served to preclude national development on any modern basis, as well as exposing the country to outside influence. The ensuing dramatic changes in social composition that accompanied the modernization of Mongolia may be seen by comparing 1925, 1956, and 1969 census data. In the first year, workers, employees, and members of their families, taken as a social category, were nonexistent; by 1956 they had formed 25.9 percent, and by 1969, 56.4 percent of the total population. Members of agricultural cooperatives, related artisans, and members of their families again were a nonexistent category in 1925 but formed 11.1 and 43.5 percent, respectively, in 1956 and 1969. Individual (that is, noncollectivized) livestock breeders, related artisans, and their families, however, formed 86.6 percent of the total in 1925, but their share had fallen to 62.8 in 1956 and to less than 1 percent in 1969. Other social groups—chiefly the landowners, secular and clerical—formed 13.4 percent of the total in 1925 but less than 1 percent in 1956 and 1969. The social development indicated by these changes is the key factor underlying contemporary Mongolian life. Although occasionally a lama may be seen striding in his bright robes through the capital, and many aspects of the national cultural traditions are preserved in museums, it is the young, educated Mongolian—often an urban worker —who is increasingly setting the pace of national life.

Mongolian People's Republic, Area and Population

	area		population*	
	sq mi	sq km	1963 census	1969 census
Autonomous municipalities				
Darhan	100	200	8,000	23,000†
Ulaanbaatar	800	2,000	224,000	263,000†
Provinces (*aymags*)				
Arhangay	21,000	55,000	63,000	72,000
Bayanhongor	45,000	116,000	46,000	52,000
Bayan-Ölgiy	18,000	46,000	48,000	58,000
Bulgan	19,000	49,000	33,000	37,000
Dornod	47,000	122,000	36,000	43,000
Dornogovi	43,000	111,000	28,000	31,000
Dundgovi	30,000	78,000	29,000	31,000
Dzavhan	32,000	82,000	60,000	71,000
Govi-Altay	55,000	142,000	43,000	47,000
Hentiy	32,000	82,000	37,000	40,000
Hovd	29,000	76,000	47,000	54,000
Hövsgöl	39,000	101,000	64,000	75,000
Övörhangay	24,000	63,000	56,000	67,000
Ömnögovi	64,000	165,000	23,000	26,000
Selenge	17,000	42,800	37,000	43,000
Sühbaatar	32,000	82,000	32,000	35,000
Töv	31,000	81,000	53,000	64,000
Uvs	27,000	69,000	52,000	60,000
Total Mongolia	604,000‡	1,565,000	1,017,000§	1,198,000

*De jure. †1969 estimate. ‡Converted area figures do not add to total given because of rounding. §Figures do not add to total given because of rounding.
Source: Official government figures.

Demographic trends. At the time of the foundation of the present Mongolian state, a high rate of disease and, particularly, a high infant mortality rate had produced a situation that, if it had been allowed to continue, would, many now believe, have endangered the survival of the entire population. In 1925, for example, the census recorded only two physicians, five medical assistants, and three nurses in the entire country. Since then, a concerted effort at affording the populace a reasonable level of medical care has reduced mortality rates considerably: as a result, contemporary Mongolia has a natural growth rate (excess of births over deaths) of about 3 percent per annum, one of the highest in Asia. The total population doubled in the first half century of the republic's history; and life expectancy also doubled, reaching approximately 65 years. Increasing urbanization and industrialization have already been noted: in absolute terms, the urban population rose from 183,000 in 1956 to 403,000 in 1963 and to 527,000 in 1969; the corresponding rural figures for the same years were 663,000; 614,000; and 670,000, respectively. The population density rose slowly over the first 50 years. In 1918 it was 1.09 per square mile (0.42 per square kilometre); by 1970 it had risen to 2.05 per square mile (0.79 per square kilometre)—a figure still remarkably low by global standards. In the individual *aymags* (first-order administrative divisions), density ranges from a low of 0.41 per square mile (0.16 per square kilometre) in Ömnögovĭ to 3.39 per square mile (1.31 per square kilometre) in Arhangay. Urbanization and an increasingly youthful, fast-growing population seem likely to characterize the demographic structure of the country for decades to come. (B.Gu.)

Population growth

THE ECONOMY

The resource policy background. Mongolia is rich in mineral resources. Geological surveys have confirmed the existence of large deposits of coal and iron, tin, copper, gold, and silver ore and a number of lesser known minerals. Mongolia's biological resources consist largely of the great herds of livestock in the country: in 1918, the total number of livestock was 9,645,600: 228,700 camels; 1,150,500 horses; 1,078,400 cattle; 5,700,100 sheep; and 1,487,900 goats. By 1970 livestock totalled 22,574,900: 633,500 camels; 2,317,900 horses; 2,107,800 cattle; 13,-311,700 sheep; and 4,204,000 goats. Overall per capita livestock figures thus rose from 14.9 in 1918 to 18.1 in 1970, providing a rich agricultural resource base which even enabled some exports of meat to be made. In addition, the northern rivers of Mongolia offer great potential for hydroelectric development, whereas the wildlife of the country offers considerable potential for commercial exploitation, a circumstance attested by the fact that the Mongolian Huntsmen's Association has some 30,000 members who kill, under conservationist restrictions, some 2,000,000 to 3,000,000 animals yearly.

Livestock

The present economic status of the country, however, cannot be understood without noting the development programs that have molded the economy of the country since its inception, because they underlie many of the changes now visible in Mongolian economy and society. In the Mongolian viewpoint, the basic task facing the new state at its inception was to abolish the backwardness of the economy (particularly its exclusive role as a supplier

of raw materials to other countries) by setting it on what was officially characterized as an anti-imperialist, anti-feudal, and noncapitalist path, first by building an agrarian–industrial economy and then by proceeding to the building of an industrial–agrarian state within a Socialist framework. The more than half century of intensive effort along these lines is of considerable interest because Mongolia was the first of the eastern nations of Eurasia to take this difficult course. With not a single industrial unit in existence in the country at its commencement, the foundations of industrialization were laid in the 1930s; and gross industrial output was growing steadily in the 1970s; in 1971 industry was yielding 20 percent of the national income, 36 percent of the gross national product, and 15 percent of the country's exports and was developing at the high annual rate of about 10 percent. Technical and economic assistance from the Soviet Union and other Socialist countries (more recently through the Council for Mutual Economic Assistance, or Comecon) played an important role in bringing this situation about. Agriculture still remained the main branch of the Mongolian economy, and 64 percent of the population were livestock breeders. Agriculture saw many technical advances after the introduction of a cooperative system in the 1950s. This continuing economic transformation has been the foundation of the social and cultural advances in the country.

The role of agriculture. Mongolian agriculture underwent radical changes in the decades after 1921. Until the late 1930s, the main emphasis was on dispossessing the large landowners and state functionaries; and some cooperative trading enterprises were introduced. After the post-World War II economic expansion, the then existing small individual *arat* (herdsmen and farmers) households were officially regarded as increasingly anachronistic, and as a result of an intensive campaign beginning in the late 1950s, agriculture changed over from small, extensive production to large-scale commodity production; and the 200,000 or so *arat* households were replaced by over 270 cooperative agricultural associations and over 30 state farms. On the average, each of these associations has 60,000 head of stock, 1,000 acres (400 hectares) of sown land, seven tractors, and seven motor vehicles. A state farm has, on the average, 27,000 head of stock, more than 32,000 acres (13,000 hectares) of arable land, over 60 tractors, and nearly 30 vehicles. The number of livestock per capita (over 20) is the highest in the world. *Arat* members, in addition to sharing the income from the sale of produce, have the right to pursue personal husbandry, with up to 50 head of stock in the north and 75 in the south. The other major change was the plowing up of virgin lands and a substantial development of crop farming, which contributed about a fifth of the gross agricultural output by the early 1970s. This development has enabled the country to cover its needs in flour, and to grow vegetables and fodder crops on an increasing scale. The introduction of large-scale activities has been accompanied by the mechanization of grain growing, vegetable growing, and hay mowing; and the spread of electricity has aided sheep shearing and milking. Fodder storage and livestock premises—hitherto virtually unknown—help protect against adverse weather conditions. Livestock breeding is now carried out on scientific principles.

The introduction of cooperatives

The role of industry. The introduction and development of industry have been the key to government planning since the foundation of Mongolia. Enormous initial difficulties were encountered, and the lack of technical knowledge and the need for huge capital investment were only overcome by outside assistance from the Soviet Union. The first factories, built in the 1920s and 1930s, processed timber, animal raw materials, and building materials; and mining and power-generation activities were also started. Through a system of producer cooperatives, industry was extended over the 1930s; but the main thrust toward industrialization took place as a result of the 1948–52 and the 1953–57 five-year plans, and the 1958–60 three-year plan. In the late 1950s, the national flour-milling industry, a result of the virgin-lands program, joined earlier endeavours in the light, food, and mining industries. Fuel and power industries were also extended. The heavy program of capital investment in industry was accelerated in the 1960s, when it totalled more than three times the entire investment in the previous 40 years and produced a 250 percent increase in industrial output over the decade. By 1971 industry was contributing almost a quarter of the national income. Emphasis was then placed on diversification, technical efficiency, and overall productivity.

The need for investment capital

The economic outlook. The economic development of Mongolia is by no means complete. Agriculture formerly stimulated industrial development by enabling the processing of animal products to raise overall economic efficiency; but industry in the 1970s was sufficiently important to enable development of the country's great natural mineral wealth. This, in turn, helped to overcome the heavy dependence on animal husbandry in the nation's exports. Animal husbandry, while capable of yielding high quality and cheap produce, is not, unfortunately, free from natural calamities. Thus efforts were directed, in agriculture, at increasing productivity, improving quality of products, and achieving stable crop yields, with the aim of satisfying the ever increasing demands of industry and foreign trade and the needs of the urbanizing population for vegetables and milk. These tasks require the implementation of programs reaching forward to the mid-1980s. Heavy investment in equipment and great emphasis on the rational use of resources in the still relatively new large-scale agriculture units were planned. Full cost accounting and the training of experts were also being emphasized. In industry, plans for the 1970s called for efficient use of scientific and technological knowledge, with special attention to geological prospecting, fuel and power development, and the growth of export industries. The ultimate aim was to transform Mongolia into an industrial–agrarian country with a rational and balanced mining and manufacturing structure and with the various elements of the national economy operating in optimal and efficient combination.

Transportation. Transportation, too, has played a major role in integrating economic development. In 1921, the sole means of transport was the cartage handled by individual *arat* households on the basis of feudal duty, with the purpose of servicing the theocratic state structure. Only camel caravans were used. When the theocratic state structure was overthrown, a new transport system had to be created. The first state-run enterprise was set up in 1925, with meagre funds and seven motor vehicles. In 1929 a Mongolian–Soviet joint stock company known as Mongoltrans was formed to handle the freight and passenger traffic flow along roads and waterways. Bridges and highways were constructed, and in 1936 the entire system was turned over to the Mongolians. The new organization, Mongoltransport, had some 300 vehicles and employed 1,000 persons. Other branches of the system were also developing: as early as 1926 steamships plied the Hövsgöl Nuur and Selenge Mörön routes, handling imports and exports; and in 1938 Mongolia's first railroad linked Ulaanbaatar and Nalayh by a narrow gauge line. A year later the first wide-gauge link was opened between Choybalsan and Soloveysk, on the Soviet border. In 1940 the first air services were introduced. After World War II, further developments occurred. Motor depots were set up in all the *aymags* in the 1950s, vehicle fleets were modernized, and an extensive road- and bridge-building program resulted in a total network of 5,300 miles (8,600 kilometres) of major roads by the 1970s. The construction of the Trans-Mongolian Railway linking Moscow, Ulaanbaatar, and Peking over the 1949–55 period, a major step forward, acted as a stimulus to settlement along its length. In the 1970s air services linked all the *aymags* and almost half of the subsidiary *somons* and state farms. By the early 1970s, railway transport accounted for 71 percent of the freight carried and 30 percent of passenger traffic; motor transport accounted for 28 and 46 percent, respectively; and water and air transport accounted for less than 1 percent of freight carried, although air transport took almost 24 percent of passenger traffic. The annual number of pas-

Mongoltrans

sengers carried exceeded 50,000,000; and freight carriage totalled almost 15,000,000 tons annually. The transport system was expected to undergo further expansion.

(Sh.B./Sh.Nj.)

ADMINISTRATION AND SOCIAL CONDITIONS

Constitutional and political framework. The Mongolian constitution adopted in 1960 states that the country is a Socialist state of workers, *arats* (herdsmen and farmers) organized in cooperatives, and working intellectuals, in the form of a people's democracy. The constitution secures to citizens the right to work, to rest, and to leisure; to material aid in old age and in sickness or disability; and to education. It also safeguards rights of national minorities. It is the parallel duty of citizens to observe the laws, to maintain labour discipline, and to defend the freedom and independence of their country. The national flag incorporates an ideogram known as the *soyombo*, comprised of ancient symbolic elements, including the Yin-Yang principle of opposites. Officially, power is exercised through bodies known as the *khurals* of people's deputies. The Great People's Khural, or Parliament, is the supreme legislative body; and the Council of Ministers is its executive body.

All the *khurals* not only legislate decisions but also are permanently functioning organs, relying in their work on the help and support of the masses. The standing commissions of the *khurals* at every level, set up for every branch of the national economy and culture, take a direct and day-to-day part in their work. By 1972 there were about 1,500 standing commissions in the country, comprising 10,322 deputies, who furthermore involved in their activities vast numbers of the general public. The country is divided into 18 *aymags* (regions), with further local subdivisions into towns, *somons* (districts), and smaller districts known as *hurins*.

Judicial system

Justice is administered through the Supreme Court, which is elected by the Great People's Khural, and through the *aymag* and town courts, which are elected by the corresponding *khurals* of people's deputies. The procurator of the republic, appointed by the Great People's Khural, and the procurators of *aymags* and towns, appointed by the procurator of the republic, supervise the work of the courts.

Mongolia has a one-party system, and the Mongolian People's Revolutionary Party (MPRP), the most experienced of any governing Communist party outside the Soviet Union, is officially characterized as the guiding and directing force of the state. By the 1970s, its members numbered over 56,000. Its most active auxiliary organization, with about 90,000 members, is the Revolutionary Youth League, or Revsomol, which has always played an important role in social and economic transformation. Because one Mongolian in two is under 25, its significance continues to be great. The Mongolian trade unions, with about 220,000 members, are members of another important mass organization: they look after the interests of their members and also play an active role in the implementation of economic and social policy. There are also various cultural and professional bodies (including an important Committee for Afro-Asian Solidarity), as well as a Young Pioneer organization for children. The Organization of Mongolian Women is particularly significant because women had few or no rights in the pre-revolutionary era; and Mongolia was the first Asian country to grant them full equality in the modern sense. Women make up about a quarter of the total number of deputies in the Great Khural, form 70 percent of the medical personnel, about half of all teachers and students, and a third of all research workers.

The place of women

Social services. Educational development has from the first been regarded as one of the most important elements in the building of the new way of life. Until 1940 the main thrust was directed at the eradication of illiteracy, the establishment of a free system of public education, and the creation of a trained intelligentsia. The beginnings were scarcely auspicious, for only a small percentage of the population was literate; the entire country, in 1921, contained but one general educational school and no public secondary school or institution of higher education. By 1924 the country had 16 elementary schools; and by 1940 there were 331 general and seven secondary schools. From the 1940s on, the main educational emphasis passed to the establishment of higher education institutions and the expansion of elementary and secondary facilities. By 1970 there were almost 600 general schools, nearly 20 special vocational schools, and five higher education institutions—the Mongolian State University (established 1942), the Higher Party School, and teachers' training, medical, and agricultural institutes. All school-age children attend rural schools for seven years; and 10 years' urban schooling is compulsory. Illiteracy had been almost completely eradicated by the early 1970s. The ratio of college students per 10,000 inhabitants, always a useful index of national educational attainment, was then 104, exactly double the ratio for France. Mongolian students also were studying abroad, mostly in the Soviet Union. The Academy of Sciences of the Mongolian People's Republic coordinates research institutions, experimental stations, and other scientific establishments and supervises the growing output of work by scholars.

Health services were similarly expanded over the republic's first half century. By the early 1970s, there were about 100 hospital beds for every 10,000 people and one physician for every 550 people—a ratio that compared favourably with many developed countries. Medical treatment is free, and the state also maintains a network of sanatoriums and holiday rest homes. The improved health facilities brought a dramatic rise in population. The total increased by nearly 5 percent between 1940 and 1956 and by over 20 percent between 1956 and 1963. The vast country was still sparsely populated in the 1970s; and the government was encouraging large families by offering various grants and allowances.

Among other social services, the provision of housing has been greatly emphasized. The *gher* remains the main traditional dwelling in the countryside, about 9,000 being produced annually. Increasing urbanization, as well as the introduction of more settled ways in the country, has necessitated the building of modern apartment blocks in the cities and an increasing number of permanent rural structures. Systems of communication were also emphasized. When the republic was founded, these consisted of a horse postal system and a single telephone wire connecting Kyakha and Zamyn-Ude on the southern frontier. The horse post system, dating from medieval times, lasted until 1949. Traditionally, it had quaint features: urgent messages were indicated by a feather; and there were no stamps, although a horse's-hoof symbol placed on each of the corners of a business letter indicated urgent mail, as did inscriptions for "gallop" or "fly, fly" (the latter with the sign of a bird). Motor-transport mail links date from 1924, shortwave communications from 1934–36, and national broadcasting from 1934. A trunk telephone–telegraph link connecting Ulaanbaatar with all the *aymag* centres, established in the late 1940s, was greatly expanded thereafter. In the 1960s, satellite communication systems were used, and in 1967 television broadcasts started.

The horse post system

(B.Sv./D.Dn.)

CULTURAL LIFE

Contemporary cultural life in Mongolia is a unique amalgam of traditional elements—the heritage of centuries—and a growing modern element built up in a scant 50 years. Official policy is to preserve what are regarded as the better elements of the old ways of life, thus avoiding, on the one hand, uncritical adulation, and on the other, an equally uncritical total rejection of old values.

Traditional elements. Mongolian literature evolved a wealth of traditional genres: heroic epics, legends, tales, *yurol* (the poetry of good wishes), and *magtaal* (the poetry of praise), as well as a host of proverbial sayings. These genres are infused by what Mongolians regard as a national characteristic—a good-humoured love of life, with particular fondness for witty sayings and jokes, particularly evident in the image of Dalan Khuldalchi, the hero of humorous folk tales, and in the stories about the *badarchins*, clever but wily wandering monks. The *bataar*

—the popular hero of folk legend—is also a symbolic figure. *Khurchins*—folk poets and singers—carried down the oral epics and ballads; and their mime and gesture gave rise to the popular trenchant satirical vaudevilles, *Sumya Noyon* and *Dunkher Da-Lam*. The religious mysteries, *tsam* and *maidari*, were formerly staged as mass spectacles. Other folk arts include the making of *shirdeg*, richly ornamented felt carpets noted as adorning the entrances to *yurts* by 13th-century European travellers. The Mongolian form of chess, *shatar*, with a stern khan for king; a dog—the cattle breeder's traditional honoured friend—as queen; and a camel as a bishop, has very deep roots and has produced some finely carved chess sets. The ancient faience decoration of glazed earthenware, with exquisite motifs, has been revived. A complicated and dignified ritual still accompanies the traditional offering and acceptance of hospitality in a country where travelling is all-important; and the seating arrangements in the *gher* are likewise carefully arranged. When conversing, Mongolians traditionally place the right palm on that of the left hand, a symbol of mutual esteem, and the same gesture, together with a light bow, expresses gratitude, greeting, or farewell.

The Naadam festival and the Three Manly Sports

The most famous celebration of traditional ways, however, is the annual Naadam festival of the Three Manly Sports, beginning each year on July 11, National Day, and held in all *aymags* and *somons*. The festival has recorded roots going back 2,300 years or more. The first sport is wrestling, prominent in ancient times at religious festivals, and the ritual entry into the arena of several hundred participants, clad in the bright colours of a special tight-fitting costume known as the Dzodog Shudag and simulating the flight of the mythical Garudi bird, is indeed an impressive sight. The contests themselves are also conducted with great ceremony. Titles awarded at the national Naadam festivals are those of Titan, Lion, Elephant, and Falcon. A three-time winner becomes a Darkhan Avraga (Invincible Titan). The second sport is archery; and bowmen vie for the title of Merghen, or Supermarksman, in individual and group contests, shooting at a leather-covered target with weapons of ancient design. Exceptional winners are characterized as Miraculous Archer, Most Scrupulous Archer, and by similar names. The third sport, horse racing, is in many ways the most spectacular because all of the competitors are children, ranging in age from seven to 12. They are highly skilled and wear fine ornamental dresses as they race for about 20 miles across country. National horse-riding competitions for all ages are held during January and February, the Mongolian New Year, and are claimed to date back to the Bronze Age. Marco Polo, a European visitor of the 13th century, describes a gathering of not less than 10,000 white horses held at the behest of the Great Khan. The races are highly competitive, and the riders phenomenally skillful.

Modern elements. Modern sports range from freestyle wrestling (introduced 1962) to motorcycling and gymnastics. A growing number of economic enterprises cater to the various folk arts. The Bogdo Genen Palace Museum has a superb collection of folk art housed in the former winter palace of the khan, built in 1898. The architectural ensemble contains temples housing the famous sculptures of the goddess Tara made by the 17th-century artist Zanabazar. The State Central Museum and related exhibits portray the rich archaeological and paleontological remains of the country. Buddhist relics are exhibited in the Choizin-Lama Temple Museum, built 1903–05. The Erdene Dzu Monastery Museum on the site of Karakorum, Mongolia's ancient capital, is also noteworthy. Each *aymag* now has its own museum of regional studies. The State Public Library contains works of great variety and historical value.

In contemporary literature, the popular poems and short stories of Dashdorjiyn Natsagorj became particularly significant in the 1930s. In the 1940s, literature became more varied in theme and genre, and the autobiographical "Old Scribe's Story" by G. Navaannamzhil became popular. Younger writers in the 1950s and 1960s injected a more contemporary note, attempting to balance psychological and social imagery. The realistic, epic novel continues in popularity.

The State Drama Theatre, founded 1931, shows both Mongolian and classical works, and the State Opera and Ballet Theatre has a deserved reputation. There is a puppet theatre in the capital and internationally known song and dance companies. Practically every community has its own amateur art group, and the Mongolian circus is also very popular. The Mongolkino film studio in the 1970s was making an increasing impact at international festivals: its productions are assisted by the magnificent landscapes and clear air of the country, which help the production of wide-screen epics. Mass radio and television services are of importance because of the great distances in the country, and are now aided by satellite links. About 70 percent of the households have radio sets, and the ownership of television sets is spreading. There are a dozen central and about two dozen local newspapers. The leading official newspaper is *Unen* ("Truth"). There are also about 30 popular and specialist periodicals with total circulation exceeding 1,000,000 copies a year.

THE OUTLOOK

The first half century of the modern republic witnessed a difficult struggle, in which the then younger generation played a major role, to liquidate the economic and cultural backwardness of the country in a historically short period. It is the government's position that, now that a Socialist system of economy has been established, the main task is building a modern industrial–agrarian Socialist state comparable to that of other nations of the Socialist bloc. Although long-term plans ranged into the 1980s, particularly in the growing cities, it was hoped that the completion of the Fifth Five-Year Plan in 1975 would see the fulfillment of the main tasks of this Socialist structure. By 1975 national income was scheduled to rise by about a third, with agricultural production going up by a quarter and industrial output by over a half. Industrial earnings were scheduled to go up by some 8 percent, and agricultural cooperative wage funds were scheduled to go up by almost a half. Housing, as well as the level of cultural and education services, was to receive continued emphasis. While the full working out of these plans lay in the future, there was little doubt that Mongolia had become increasingly important internationally. The country has been a United Nations member since 1961, and a member of the Council for Mutual Economic Assistance, the Socialist trade group, since 1962. Although the bulk of Mongolian trade is still with Comecon, links with capitalist countries were developing in the 1970s, and a growing tourist trade—of well-equipped hunting parties—was making the country better known. As diplomatic and trade links developed, it seemed likely that the unique contribution to the modern world of a developed Asiatic country with a proud tradition would become increasingly known. (Sh.Bi.)

BIBLIOGRAPHY. The rapidly changing nature of Mongolian society renders obsolete many of the publications about the country. The CENTRAL STATISTICAL BOARD OF THE MPR COUNCIL OF MINISTERS has nevertheless produced a statistical survey, *Fifty Years of the Mongolian People's Republic* (1971), which contains a wealth of useful material. The edition is in both Mongolian and English. The illustrated magazine *Mongolia* (monthly since 1971), in English, also contains statistical data as well as feature stories. A series of English-language pamphlets put out from Ulaanbaatar on the occasion of the 50th-anniversary celebrations contain collections of essays on economic, cultural, and geographical topics by experts. The short essay by O. NAMNANDORJ and G. AMAR entitled "National Sports," published in 1971 by the Mongolian Committee of Physical Culture and Sports, is a survey of traditional and modern elements. C.R. BAWDEN, *The Modern History of Mongolia* (1968), contains much material on the contemporary republic from the viewpoint of a British specialist. Another British publication, A.J.K. SANDERS, *The People's Republic of Mongolia* (1968), contains general information culled from the Mongolian and Soviet press. IVOR MONTAGU, *Land of the Blue Sky* (1956), is a travel account. See also the *Memorandum Attached to the Application Submitted by the Government of the Mongolian People's Republic for Membership of the United Nations* (1957).

Mongols

The Mongols comprise one of the principal ethnographic divisions of Asiatic, or Oriental, peoples. Their traditional homeland is centred in Mongolia, a vast plateau in Central Asia now divided politically into an autonomous region of the People's Republic of China (Inner Mongolia) and the Mongolian People's Republic (Outer Mongolia), which lies at the eastern end of what was a great corridor of migration between Manchuria and Hungary throughout history. It is unfortunate, because confusing, that 19th-century physical anthropologists introduced the terms Mongol and Mongolian as descriptive of racial type. The Mongols exhibit a wide range of physical characteristics and should be considered not as a race but as a group of peoples bound together by a common language and a common nomadic tradition.

The geographical origin of the Mongols themselves is the northeast corner of the present Mongolian People's Republic. To the east, the ancient tribal history is mostly that of the Tunguz peoples (including the ancestors of the Manchus) and to the west, that of the Hsiung-nu, or eastern Huns, and their Turkish-speaking successors, whom the Mongols eventually displaced and in part absorbed. As a result of later wars and migrations Mongols are now found in the Mongolian People's Republic, in the Buryat Autonomous Soviet Socialist Republic, in the Chinese Inner Mongolian Autonomous Region which includes a good deal of territory in Manchuria (in Chinese: The Northeastern Provinces), in the province of Ninghsia, on the northern plateau of Tibet, in the Sinkiang Uighur Autonomous Region (Chinese Turkestan), in Afghanistan, in the Tuva Autonomous Soviet Socialist Republic, and the Kalmyk Republic of the U.S.S.R.

ETHNOGRAPHY AND EARLY TRIBAL HISTORY

All Mongols recognize their kinship to each other in varying degrees through legend, written history, and especially language. Dialects vary from east to west more than from north to south, but very few are unintelligible to other Mongols. Historical change in the language is indicated by the fact that reading *The Secret History of the Mongols* (mid–13th century), the oldest major document written in Mongol, is for a Mongol of today like reading Chaucer for a modern Englishman. Pan-Mongolism, the political desire for a reunion of all the Mongols, was always more romantic than practical and today is a dead issue.

Nomadic tradition

The Mongols have always been nomads, though there has also always been some cultivation. But nomadism means the seasonal movement of livestock and camps from one pasture to another, not unfettered wandering. Nomads have a clear concept of the possession of territory, though they sometimes interpret this in socially conflicting ways. Legend and folklore show that among the premodern Mongols the common people considered livestock to be private property and land the collective property of the tribe, while the families of ruling chiefs considered the land as well as the persons of their subjects to be their property.

Traditional society was based on blood relationship traced through a common male ancestor who gave his name to the clan, though evidence exists of a more ancient system of matrilineal descent. Marriage was forbidden between members of the same clan, giving rise to complicated marriage alliances (and also feuds) among the clans. As clans grew and merged into tribes (often inventing a fictitious common ancestor), the most successful families tended to arrogate to themselves claims to "real" ancestry and, at the same time, to control of the tribal territory, while lesser families could claim ancestry only in a vaguer, tribal sense. In this process weak clans fell to a subordinate but not servile status: they owned their own cattle and had their own headmen, but paid tribute to the ruling clan and moved, camped, pastured, and fought under its orders.

Political and military organization was matched to the family–clan–tribe pattern. Every man who could ride and bear arms was both a herdsman and a soldier according to the need of the moment. Raiding other tribes to capture cattle, women, and prisoners was a recognized method of property accumulation. When, however, a tribe rose to notable power, as in the time of Genghis Khan in the 13th century, a decimal form of military organization was adopted, with units of 10, 100, 1,000 and 10,000. Commanders of large units were assigned territories from which they drew the tribute to the supreme khan and mustered their quotas of troops. Mongol society fluctuates between such periods of feudal concentration and those of tribal dispersion.

The first mention in the Chinese chronicles of tribes that can be identified with Mongolia goes back in a shadowy way to the 2nd millennium BC. The first inhabitants of whom there is certainty, however, are the Hsiung-nu, or Huns, about the 4th century BC. It was once thought that they were Turks, or at least Turkish speaking, but in the last quarter of the 20th century the opinion was growing that they spoke a paleo-Asiatic language, represented today by the Ket dialects of the Yenisey Valley in Siberia. The Hsiung-nu created a great tribal empire in Mongolia while China was being unified as an imperial state under the Ch'in (221–206 BC) and Han (206 BC–AD 220) dynasties. After several centuries of war with the Chinese, complicated by civil wars among themselves, the Hsiung-nu confederation broke up. Some of the southern tribes surrendered to the Chinese and were settled within China, where they were eventually absorbed. Some of the northern tribes migrated westward, where they appear in Europe in the 5th century AD as the Huns of Attila. By then, of course, they were a very mixed people.

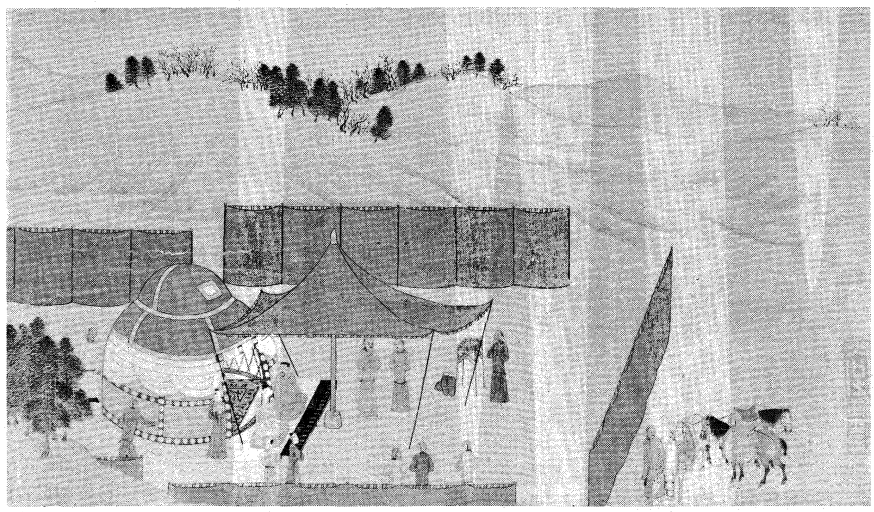
The Hsiung-nu

In Mongolia the Hsiung-nu were succeeded both by Turks, or Turkish-speaking peoples, and by others identified by some scholars as Mongols, or Mongol speaking. There is a lack of convincing archaeological or historical evidence that these groups came to Mongolia from some distant region to fill a void left by the Hsiung-nu departure. Probably they were there all the time as the subjects of the Hsiung-nu, until the breakup of that confederation gave them the opportunity to assert themselves. Among the peoples who have been considered possibly Mongol the most important tribal names are Sienpi (Hsien-pi), who may however have been Tunguz rather than Mongol, recorded in Han dynasty annals, and the Jou–jan (Juan–juan or Geougen) of the 4th to 6th centuries. The latter have been identified by some scholars with the Avars, who migrated into Europe along the plains of the Danube, and were nearly annihilated in Hungary by Charlemagne in the late 8th century.

The Orkhon Turks

According to a legend recorded by the Chinese, the Turks of Mongolia, whose name is recognizable under its Chinese transcription "T'u-chüeh," were a subject tribe ruled by the Jou–jan. The Turks overthrew their masters and soon were in control of all Mongolia, centring their power in the Orkhon Valley in the northern part of the country. The Orkhon Turks were contemporaries of the T'ang dynasty (618–907) in China, and their fortunes rose and fell in counterpoint to periods of T'ang strength and weakness. Comparison of archaeological and historical data, moreover, shows that power in Mongolia was at this time not based simply on levies of nomad horsemen. The khans and great men had fixed headquarters, surrounded by cultivated land that enabled them to breed large, stable-fed horses capable of carrying a man in armour. This situation emphasized a class distinction between the aristocrat on his charger and the herdsman-warrior-archer on his smaller horse. Agriculture also became an element in the economy, and the Uighurs, who came to power after the fall of the Orkhon Turks, enter history as an oasis-centred people.

In the welter of tribes the name Mongol first appears in a tribal list recorded under the T'ang dynasty. It then vanishes, to reappear only in the 11th century, when the Khitan (Kidan, from which comes the word Cathay) ruled in Manchuria and North China, and controlled most of Mongolia. The Khitan, who established the Chinese dynasty of Liao (907–1125), were themselves a Mongol people, but their homeland was in Manchuria rather than in what is now Mongolia. Like other Chinese



A Mongol encampment, showing a central yurt surrounded by felt screens, beside which stand three Mongolian ponies. Detail from the "Wen Chi Scroll," a Chinese handscroll of the Southern Sung period (1126–1279), which illustrates the story of Wen Chi, a Chinese lady captured by the steppe nomads. In a private collection.
By courtesy of Asia House Gallery, New York

dynasties, the Liao exercised its power in Mongolia by playing off the tribes against one another. Liao sources record the existence of a rather shadowy tribal power known in Mongol tradition as Khamag Mongol Uls ("Nation of all the Mongols"), which did not, however, include all of the population who spoke the Mongol language.

When the Khitan fell, their power in China was taken over and extended by the Juchen (Jurchid; Chinese "Ju-chen" or "Nü-chen"), a Tunguz people based farther north in Manchuria. They took the Chinese name of Chin (Kin; "Golden"). In their tribal policy they switched their favour from "All the Mongols" to the Tatars (known in the West as "Tartars," from a medieval pun on *tartarus*, Latin for "hell"). Though Mongols, the Tatars were not part of the tribal league of All the Mongols, centred in the Onon and Kerulen valleys in the eastern half of North Mongolia; the Tatars lived to the east and south of them.

On the whole, though chastened occasionally by punitive expeditions, All the Mongols had been transfrontier allies or auxiliaries of the Khitan-Liao. A contingent of 50,000 Mongols (large for that time) fought on the Khitan side in the last battles of the Khitan Empire. Presumably, this was one reason why the Juchen-Chin transferred their favour to the Tatars, nearer to their frontier. Such alternations, between using the more distant and using the nearer transfrontier and frontier tribes, were frequent in the policies of dynasties in China, and this one had the desired effect of creating a feud between Mongols and Tatars.

Before the era of Genghis Khan a defeated Khitan army had migrated westward at the fall of their Liao Empire. It was led by a prince of the Khitan imperial line, but must have included heterogeneous tribal elements. Moving westward through Mongolia it reached what is now Kazakhstan in the southeast of Soviet Asia and created a new and briefly powerful empire, the Karakitai. It ruled primarily over Turkish-speaking peoples, made up of nomads and city-dwellers in the oases, and the Khitan nucleus had the opportunity to apply its knowledge of how to deal with nomads and its ability in the administration of a bureaucracy.

The Karakitai

THE RISE OF GENGHIS KHAN

Such was the setting in Mongolia when Genghis Khan (his given name was Temüjin) was born, about 1162 (the date now accepted by contemporary Mongol scholars). Genghis was born into a clan that had a tradition of power and rule, being the great-grandson of Khabul (Qabul) Khan, who had been the greatest ruler of All the Mongols. Genghis inherited a feud against the Juchen-Chin dynasty and another against the Tatars, who had betrayed a collateral ancestor of his to the Juchen. His own father was poisoned by Tatars. He also inherited feuds among the ruling clans of All the Mongols, and a feud with the powerful Merkit (Mergid) tribe, from whom his father had stolen his mother.

Genghis was even more deeply a political man than a warrior, and he resorted to war only as an extension of policy by other means. He was orphaned in his teens; his family fell on bad times, and power among the Mongols passed to other clans. Even in such apparently primitive practices as camp-raiding and horse-thieving, he skillfully used ancient customs: marriage alliances; putting himself under the patronage of a stronger prince; making an alliance with Jamuka (later his dangerous rival) by the oath of *anda*, under which men became as if blood-brothers; and recruiting *nökhör* (the modern Mongol term for "comrade"). Unlike the institution of *anda*, which created a fictitious kinship, a man who became a *nökhör* forswore all loyalties of kinship and tribe and declared himself solely "the man" of his chosen leader. Genghis later fell out with his *anda*, but he was never betrayed by a *nökhör*, and his most brilliant generals were *nökhör*.

Genghis broke alliances and betrayed loyalties, but only when he could seem to be acting in "the common cause." By 1206 his success in tribal warfare caused Temüjin to be proclaimed ruler of All the Mongols with the rank of khan and the title of Genghis (Chingis)—a word deriving probably ultimately from the Turkish *tengiz*, meaning "a large body of water, the ocean," although this explanation has not convinced all Mongol scholars. Previous nomads had invaded China, but none had yet ruled the whole of it, chiefly because they had invaded China prematurely, leaving other nomads on their flanks and in their rear. Genghis, however, first united all the "dwellers in felt-walled tents" (*tuurgatan*), probing far back, away from China, to make sure that he controlled all potential nomadic rivals.

Ruler of All the Mongols

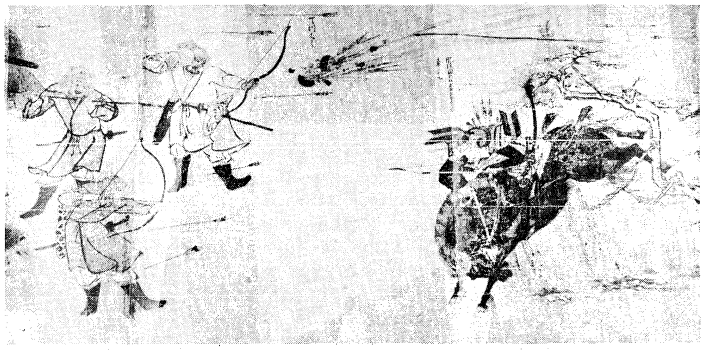
His first move was to bring under control the major tribal groups to the west of him in Mongolia, the Naimans and Kereits (Keraits) with whom he had been alternatingly in alliance and rivalry, as well as the tribes fringing the northern frontier of Mongolia with Siberia. He then turned toward China, where at this time the eastern half of North China, down almost to the Yangtze, was ruled by the Juchen-Chin. In the northwest corner of China and the western extension of Inner Mongolia there was a strange state, that of the Hsi Hsia: its rulers were Tanguts from Tibet and under them there were Turkish merchants who exploited the caravan trade, the cultivators of the oases being Turks and Chinese. China south of

the Yangtze was ruled by the Southern Sung dynasty (1126–1279). Although they had lost North China, the Southern Sung were expanding southward toward Indochina, bringing rich new land under cultivation. Among all of these states there was an interplay of diplomacy, alliances made and broken, and open warfare. The Mongols themselves, far from being ignorant barbarians, understood the game and played it skillfully.

Between 1207 and 1215 the Genghis armies probed deep into North China. He made good use of the Khitan of South Manchuria and North China, whose Liao Empire the Juchen-Chin had overthrown and who were now discontented subjects of the Chin. In 1215 Peking itself (from which the Chin emperor had withdrawn southward) was taken and sacked. Realizing, however, that it was premature to commit his main strength to the conquest of China, Genghis withdrew to Mongolia, leaving one of his best generals, Mugali, to ravage and weaken the country. He himself turned westward. When he had defeated the Naiman, the last of the powerful tribes in Mongolia Proper, the son of the last ruler of that tribe, Küchlüg, fled to Karakitai and married the daughter of its last ruler, whom he then overthrew. In that variegated kingdom, which included Semirechiya in Russian Turkestan and the Kashgarian Oases in Chinese Turkestan (Sinkiang Uighur Autonomous Region), he favoured the Buddhist minority and persecuted Islām, the majority religion. This situation made it easy for the Mongols to defeat him. The Mongol general Jebe (Jeb) proclaimed freedom of religion and forbade massacre and plunder. This policy indicates that the Mongols did not massacre out of sheer savagery but only when they thought it necessary to break the power of an opponent.

Sack of Peking

By courtesy of the International Society for Educational Information, Tokyo



The Japanese champion Takezaki attacking Mongol archers, a detail from the "Mongol Invasion Scroll," *c.* 1293. In the Imperial Household Collection, National Museum, Tokyo.

Taking over the lands of the Karakitai opened the way for Genghis to Khorezm (Khuarezm), the land of the oases along the Syr Darya and Amu Darya in northern Iran. For drawing on the resources of a higher civilization this gave him an alternative to China, and it also secured him against the danger of any other nomadic power organizing, on his flank and rear, a military striking force backed by agricultural and urban resources. This done, he turned back toward China, leaving further campaigning into Russia and the eastern fringes of Europe to his generals and sons. He would not, however, commit his main forces in China until he had dealt with the wealthy Tangut state of Hsi Hsia; it was on this successful campaign in 1227 that he died.

THE SUCCESSOR STATES OF THE MONGOL EMPIRE

Genghis had already dealt with the problem of succession. Each of his four sons was to hold a vassal kingdom: Jochi, the eldest, was given the land from the Yenisey River and the Aral Sea westward "as far as the hooves of Mongol horses have reached"—a wording attributed to Genghis himself; the second son, Chagatai (Tsagadai), received Kashgaria (now the southern part of Sinkiang, China) and most of Mavrannakhar between Amu Darya and Syr Darya; the third son, Ögödei (Ogadai), received Western Mongolia and the region of Tarbagatai (now the northwestern corner of Sinkiang); the youngest, Tolui, inherited the ancient Mongol homeland of eastern Mongolia. Two years later, in 1229, a great Mongol assembly confirmed the succession of Ögödei as the great khan (khagan).

These dispositions made skillful use of ancient traditions. It was the custom among prosperous families that the eldest son, on reaching manhood, was given a wife and his share of the future inheritance; he then moved away and set up his own camp, independent but still allied to his family. The other brothers followed in due order, but each one nearer to the "home camp" than his next older brother. The youngest, as "guardian of the hearth and fire," remained with his parents until their death and received the residual heritage. It was convenient that Jochi could in this way be placed at the greatest distance from the ancient homeland because he got on poorly with his brothers who considered him illegitimate, conceived while his mother was the captive of a hostile tribe. The election of Ögödei as great khan over the head of his elder brother Chagatai (Jochi had already died) did not do violence to nomadic tradition; it was quite acceptable in wartime for the dying ruler to nominate as his successor the son who was considered most able and most acceptable to his brothers.

Tradition of succession

With this first division, further fission was inevitable. Under Batu, the successor of Jochi, there began the formation of the Golden Horde, which ruled or rather drew tribute from the city-states of Russia. In this khanate the Mongols were greatly outnumbered by Turks; the Turkish language soon displaced Mongol, and Islām became the prevailing religion. Because its reservoir of nomad power was in the Kipchak Steppe, the Golden Horde is sometimes known as the Kipchak khanate. By its methods of collecting taxes and tribute it contributed to the rise of the grand dukes of Muscovy, and it was eventually a Moscow-led alliance that broke the power of the Mongols (by then more frequently called Tatars), at the Battle of Kulikovo in 1380. Partly by treachery and partly by guile the Golden Horde was still able to take and sack Moscow, but its power soon disintegrated—an important factor being attacks by Timur, coming from Turkestan.

To the east were the khanates of the House of Chagatai and the Il Khans of Iran (Persia). Like the rulers of the Golden Horde, the rulers of the House of Chagatai considered themselves senior, in genealogy, to the House of Ögödei; they were frequently at odds with the great khan, with each other, and with the Il Khans. On the other hand, the Il Khans (the title itself implies subordination) accepted and supported the authority of the great khans. Like the Golden Horde, again, the House of Chagatai controlled wide pastures and therefore retained a strong nomadic base, while the Il Khans, like the great khans (especially after Kublai [Khubilai] moved his capital into China), were directly affected by the urban influences of an old, highly developed civilization with a rich literary tradition. As in China, this situation led rather rapidly to the passage of real administrative control from Mongol hands into the hands of their subjects. The greatest of the Il Khans was Hülegü (Khulagu, Hulagu), a brother of the great Kublai Khan (*q.v.*) of Mongolia and China, who began the Il Khan tradition of supporting Peking against the House of Chagatai and the Golden Horde.

House of Chagatai

This brief review of the outer khanates helps to explain the subsequent history of the Mongols in Mongolia and the Mongol dynasty in China. As great khan, Ögödei authorized the continuation of Mongol campaigns in Russia and the west and also in China, where the disintegration of the Juchen-Chin dynasty in 1234 had brought the Mongols face to face with the surviving Sung dynasty in the Yangtze Valley. Ögödei was also able to maintain a system of imperial representatives in the appanages of his imperial kinsmen in Central Asia and Iran, but was less able to control the always insubordinate Golden Horde. He died in 1241 and was succeeded, after a stormy regency under his widow Töregene, by his son Güyük (Kuyug) who had already quarrelled with his cousin Batu of

the Golden Horde. Güyük died at Samarkand in 1248, preparing an attack on Batu.

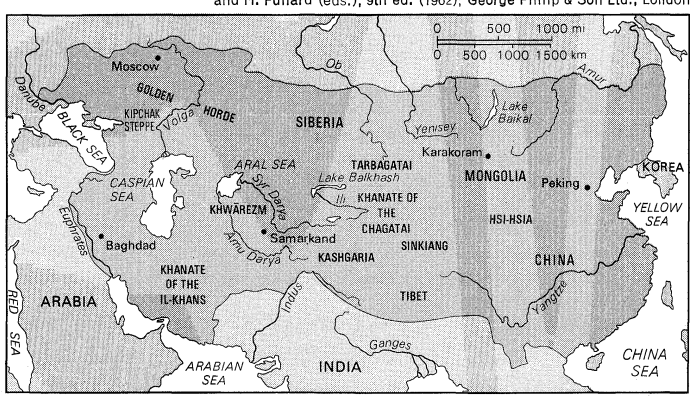
A major change then occurred in the succession. At the next great assembly of the descendants of Genghis Khan, enlarged by the presence of powerful commanders and officials, the great khan chosen (again after much intrigue) was not a son of the House of Ögödei but Möngke (Mungke, Manga), a son of Tolui, the "guardian of the fire and hearth" of the Mongol homeland. This choice was favoured by Batu Khan, and Möngke responded by trying to stabilize and pacify relations among the khanates. Of his brothers he sent Kublai (later Great Khan) to continue the conquest of Sung China and Hülegü to subdue the Ismāʿīli (the fanatical Muslim secret order of assassins) and Baghdad. He was aware of the desire of some of the crusaders for a Mongol alliance against the Saracens, but like Ögödei and Güyük he would not consider this except on terms of the submission of the European rulers and the pope. He himself campaigned deep into southwest China and there died of a fever in 1259.

Kublai Khan

The succession was then disputed between Möngke's second brother, Kublai (Khubilai), and his youngest brother, Arigböge (Ariböx, Arikböge). A third brother, Hülegü, supported Kublai. The dispute was more than a brawl over spoils among barbarian warriors; ideology was involved. Genghis Khan's concept of conquest and rule had been clear: the "people of the felt-walled tents" should remain in the steppes and continue their ancient warrior way of life, drawing tribute from the world of farms, cities, and caravan trade. Kublai and Hülegü, however, favoured moving into the conquered countries and there becoming the new ruling class—even if this meant mingling with the remnants of the conquered ruling class.

In this respect Arigböge was closer to the concept of Genghis than was Kublai, but Kublai prevailed and Arigböge died in honourable captivity. Kublai had himself

Adapted from *Muir's Historical Atlas; Medieval and Modern*, R.F. Treharne and H. Fullard (eds.), 9th ed. (1962); George Philip & Son Ltd., London



The Mongol Empire under Kublai Khan.

proclaimed great khan in Mongolia in 1260. Kublai's reign has been romanticized in the West ever since Marco Polo. Kublai Khan moved the capital from Karakorum (Kharakhorum), which had been built by Ögödei (not Genghis Khan, as is often said), to Peking. He used Mongolia as his base for ascendancy over the other Mongol khanates, but drew his main revenue directly from China. He used foreigners (including Marco Polo and his family) to lessen his dependence on Chinese bureaucrats, but the administrative structure was essentially on the Chinese model; and consequently the Mongols, because they no longer held the real keys to power, lost the throne to the native Chinese Ming dynasty in 1367, only a century after the accession of Kublai and within a decade or two of the end of the Golden Horde and the Il Khanate.

Internecine strife. Although in the first vigour of reconquest the Chinese penetrated deeply into Mongolia and burned Karakorum, they never succeeded in establishing control. Mongol unity was shattered, but Mongols in different regions began to recover. Mongol fission followed several lines. In western Mongolia there arose new lines of chieftains who did not claim descent from Genghis Khan. As a group, these were the Oirat, but at times the names of subgroups or individual tribes, such as the Dzungar (Jüüngar) or the Dörbed (Dörböd), predominated. In the centre, both in Outer and Inner Mongolia, the ruling princes claimed descent from Genghis Khan. In the east, largely in Manchuria, were princes whose ancestor was Khasar, a brother of Genghis Khan. Because there had been bad blood between the two, these princes and the Chingiside princes were always suspicious of each other.

What followed, in the renewed tribal wars and pressure on the frontiers of China in the 15th and 16th centuries, was much more than a resurgent "wave of barbarian invasions." A distinct new period was opening in which all concerned understood that in order to have real power outside the Great Wall of China it was necessary to coordinate nomadic military mobility with towns inhabited by productive artisans, capable of attracting trade from China, and supplied with food by local farming. The lead was first taken by the Oirat, in the far west of Mongolia, who established control over some of the oases of Sinkiang and began to penetrate Tibet. This advance meant that in the regions where the imperial power and economic ascendancy of China under the Ming dynasty (1368–1644) were weakest, the Oirat drew on new resources. Both the Tibetans and the Turkish-speaking oasis people were active merchants, had a literate class whose thinking was independent of the Chinese model, and could keep the records without which a state more advanced than a tribal league was impossible. This stage initiated the long-enduring cleavage between the Oirat and the Khalkha, the main body of what was later to be Outer Mongolia.

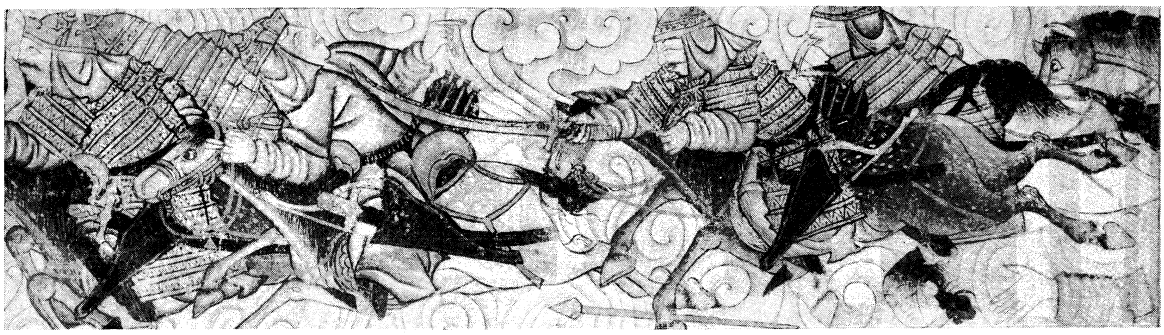
Ascendancy then passed to the Mongols of the Ordos, in the great loop of the Yellow River, under Altan Khan (reigned 1543–83). He exploited a geographical base that enabled him to develop agriculture and trade, to challenge the Oirat in Tibet, and to pressure the Chinese. Meanwhile, the Mongols of the centre, Khalkhas in the north and Chahars (Tsakhars) in the south (there was as yet no "Outer" and "Inner" Mongolia), had lagged behind for lack of a suitably diversified geographical base. The best that they could achieve was a tribal league unification under Dayan Khan, a descendant of Kublai and grandfather of Altan Khan, who was proclaimed khan in 1470 at the age of five and died in 1543. After this, and after the death of Altan Khan, the supremacy over the Mongols of the centre passed to the south to another descendant of Dayan, Ligdan (Legdan, Lingdan) Khan of the Chahars. Using the geographical advantages of the modern Chinese frontier province of Chahar, close to the Great Wall, he tried during his reign (1604–34) to build up a power comparable to that which had been held by Altan Khan. He was too late, however, because of the rise of the Manchus (see below).

Altan Khan

Revival of Buddhism. It was this period that saw the second flowering of Buddhism among the Mongols. In the reign of Kublai, Buddhism in its Tibetan form had been fashionable at court and among some of the Mongol aristocracy, but the people as a whole had not been converted. The new entry of Buddhism was promoted by political considerations. A number of Mongol princes, almost simultaneously, saw that for the kind of power that was now advantageous it was necessary to have not only a religious ethos higher than that of shamanism (whose priests used magic to cure the sick, divine what was hidden, and to control events), but a literate class to provide a bureaucracy. To use the Chinese language meant the risk—as had been proved under the Mongol empire in China—of the absorption of the Mongol ruling class into the Chinese ruling class. Tibet was not strong enough to dominate Mongolia. The Tibetan monastic system had already produced able clerical bureaucrats. Moreover, Tibetan alphabetic writing was easier to use than Chinese hieroglyphs.

Establishment of the Dalai Lama

Thus it was that Altan Khan invited from Tibet a prelate who had claims to primacy in Tibet, but also rivals, and proclaimed him Dalai ("Ocean," "All-encompassing") Lama. Significantly, this title is Mongol, not Tibetan.



Mongol warriors in cavalry pursuit wearing lamellar armour, which covers them from the neck to the elbow and to the lower leg. Miniature from *Jami' at-tawarikh* by Rashīd od-Dīn, 14th century. In the Topkapi Sarayi Museum, Istanbul.
By courtesy of the Topkapi Sarayi Museum, Istanbul

Moreover, a way was found to link church and state. A son of the line of the Tüshētü Khans of Khalkha was conveniently found to be the first "reincarnation" of the line of Jebtsundamba Khutukhtus of Urga. The significance of this device is underlined by the fact that as soon as the Manchus controlled Mongolia, they ruled that no man of the lineage of Genghis Khan could be "discovered" to be a reincarnation or Living Buddha, and also that the Living Buddha of Urga must always be discovered in Tibet. In their rule of Mongolia they thus separated church and state and used them against each other.

At the beginning of the revival of Buddhism in Mongolia there was a great burst of translation of the scriptures from Tibetan (and Sanskrit) into Mongol. The Mongols wanted to use Buddhism as a unifying principle in a new nationalism. When the Manchus won control, however, they threw their support to the use of Tibetan as the "Latin" of the church, further widening the cleavage between clerical and secular authority and bureaucracy. By the end of the Manchu regime there were many monks in Mongolia who were literate in Tibetan, but not in their own language.

THE ASCENDANCY OF THE MANCHUS

The rise of the Manchus, which had such profound effects on the fate of Mongolia, began long before 1644, the year a Manchu emperor was first seated on the throne in Peking. In the late 15th century it was becoming clear that a new barbarian conquest of China was again possible. In competition with the various Mongol princes and tribes already mentioned, the Manchus had the advantage that in South Manchuria, but outside the Great Wall of China, there was a large Chinese population with a number of urban centres and a flourishing trade that, instead of passing by land through the Great Wall, went largely by sea to the promontory of Shantung—to the rear, that is, of the rulers in Peking. These Chinese were somewhat alienated from other Chinese. They had for centuries been accustomed to trading with the barbarians and to farming under the patronage of barbarian princes, and they did not like Peking's periodic attempts to maintain a "closed frontier" along the Great Wall.

The Manchus subjugated these Chinese, but also cultivated their loyalty, and were soon heavily dependent on them not only economically but for military manpower. To balance this dependence they built up a network of alliances with their other neighbours, the easternmost Mongols. Mongol troops took part in the conquest of China; and before they occupied Peking, the Manchus had control of the southern fringe of Mongolia. They organized it as part of their military reserve for the domination of China; this organization is the origin of the institutional and administrative concept of "Inner" Mongolia.

It took the Manchus about a century to add northern or "Outer" Mongolia to their empire, resulting in two Mongolias markedly different from each other, Inner Mongolia being from the beginning much more closely integrated with China.

Meanwhile, the Oirats, under their leading tribe, the Jüüngar, made a belated effort to unite all the Mongols in rivalry with the Manchus. The Oirat were strengthened by their control of a number of the Sinkiang oases, but weakened by rivalries among their chiefs, by the diversion of much of their strength to adventures in Tibet, and by the reluctance of the Khalkha princes to accept the overlordship of princes not descended from Genghis Khan. Led by such great warriors as Galdan (Galdan Boshogt), the Oirat made sweeping campaigns far to the east in Mongolia, but were never quite able to consolidate their gains. In trying to make the Oirat a recognizably distinct nation, the great religious leader, the Jaya Pandita, revised the Mongol alphabet, making it phonetically more accurate, and originated an independent literary tradition.

Attempt at Mongol revival

Unwilling to accept submission to the Oirat as the price of unification, the Khalkha princes rallied more and more to the Manchus who guaranteed their aristocratic privileges and titles in a great convention at Dolonnor, in Inner Mongolia, in 1691. With the added resources of Khalkha, the Manchus were then able to mount a long series of military campaigns in which they annihilated the Oirat power with tremendous slaughter on the scale of genocide.

This conquest, however, was not completed until 1759, and it was complicated by many events, particularly a great revolt against Manchu rule in western Khalkha in the 1750s led by a noble named Chingunjav, who was a coconspirator with an Oirat leader named Amursana, who in turn had first submitted to the Manchus and then rebelled against them. But this was the last period of general warfare involving the Mongols, and it ended with a considerable redistribution of the tribes. Several Khalkha groups that had fled from the Oirat into Inner Mongolia never returned; a few Chahars from Inner Mongolia were settled in Sinkiang as garrisons; numbers of the Oirat group were included in the western part of Khalkha geographically, but not within the tribal organization; some ended their migrations in Ala Shan, at the western end of Inner Mongolia, but not within the Inner Mongolian organization; and some ended theirs far away in the Kokonor-Tsaidam region of Tibet. The most distant Oirat wanderers (mostly Torguuds and Dörbeds) migrated in the early 17th century from the Altai to the Volga, where they took service under the tsars and took part in the Russian conquest of the Caucasus. In 1771 about 70,000 families migrated all the way back to Sinkiang, where they were accepted under Manchu rule and allotted pastures for grazing. The descendants of those who remained on the Volga were known as the Khalimag (Kalmuks).

Mongol dispersion

The ensuing period of peace degenerated into stagnation and economic decline. Chinese camp followers had accompanied the Manchu conquest, and from this grew Chinese control of the caravan trade and of a barter trade exploiting usurious terms of credit. Because Mongol troops were of decreasing use for the control of China, there was no incentive for the Manchus to protect, economically, this source of manpower, and the Manchu authorities relied increasingly on the potentates of the Lama Buddhist church, who were themselves increasing-

ly corrupt, for the control of Mongolia. Chinese colonization began to encroach on the pasture lands of Inner Mongolia and at the end of the 19th century an attempt was made to plant a screen of Chinese colonists along the frontier between Siberia and Outer Mongolia.

THE 20TH CENTURY

In the Russo–Japanese War of 1904–05 both Russians and Japanese enlisted Mongol mercenaries as auxiliaries; from this time on there were Japanese army officers who dreamed of a new Mongol nationalism that could be used against both Russia and China. The Russians were more restrained and were satisfied with Buryatia (Buryat Mongolia). By secret treaties after the war Inner Mongolia east of the meridian of Peking was recognized by Russia as a Japanese sphere of interest.

By 1911, when the Chinese Revolution broke out, unrest was widespread in Mongolia. At the time, the Mongol language and Mongol sources being little known outside Mongolia, most observers thought of Mongolia only in terms of Russian and Japanese intrigue. But a rich documentation now available proves that the unrest was both social and political. The Mongols by this time identified the Manchus with the Chinese, the Chinese with usurious debt, and their own clerical and secular rulers as people who lived in luxury on Chinese loans and passed on the usurious interest to their subjects. The rulers, for their part, saw both the chance of a new government they could control and the danger that if they did not act, they would not be able to control the people.

Independence

Under the leadership of the Jebtsundamba Khutukhtu ("Urga Living Buddha"), the Mongols declared their independence. Uncertain of themselves in world politics, however, they sought to replace Manchu imperial patronage with that of the tsar; but the Russians, because of the secret treaties with Japan and an understanding with Britain about Tibet and Mongolia (which though not secret could hardly be comprehended by the Mongols), would go no further than support of "autonomy," not "independence." This status was ratified after difficult negotiations between the Mongols, the Russians, and the new Republic of China. Union between Inner and Outer Mongolia was similarly frustrated. Some Inner Mongolian leaders saw themselves as the future elite of a united Mongolia because they could draw on an intelligentsia with a knowledge of the language and politics of China, but for the same reason they were distrusted in Outer Mongolia as too Chinese.

This uneasy situation endured with increasing economic distress and social unrest until the Russian Revolution. A Japanese-controlled clique in the Chinese government then sent a military expedition into Mongolia that forced the Mongols to sign a "request" to be taken over by the Chinese Republic. Almost immediately afterward, defeated anti-Bolshevik troops began to retreat into Mongolia. Their most important leader was the "Mad Baron" Ungern-Sternberg, who defeated the Chinese occupation forces and treated the Mongols with unfeeling savagery.

In this period of terror and confusion two secret revolutionary groups, which later merged, were formed by Sükh Baatar, a former trooper and machine gunner in the Mongol forces disbanded by the Chinese and Choibalsan, who had been a boy runaway from a monastery and later a student in Siberia. Choibalsan, though not the principal leader during the lifetime of Sükh Baatar, was in touch with underground Bolsheviks—hiding there as refugees—later massacred by Ungern-Sternberg. Memoirs of the Mongol partisans enlisted by these two show that quite apart from Russian propaganda there was already in Mongolia a geographically widespread and socially embittered demand for radical social and political change. The traditional leaders had been discredited by their inability to handle either the Chinese intervention or the incursions of the anti-revolutionary Russians.

The revolutionaries took the initiative in going to the Bolsheviks for help, which was quickly granted. The remnants of the Chinese warlord forces were driven out and Ungern-Sternberg was handed over to the Bolsheviks for execution. Urga, the capital (now Ulan Bator), was taken by a joint Mongol-Russian column in July 1921, and this is now considered the date of the founding of the present republic, though the first measures of the revolutionary victors were surprisingly moderate. The "Urga Living Buddha" was continued in office, but as a "constitutional monarch," meaning he could sign only documents prepared for him by the new regime.

Sükh Baatar died in 1923 and the Living Buddha in 1924. In Russia Lenin also died in 1924, and in China Sun Yat-sen died in 1925. Mongolia was engaged in a revolutionary process also going forward at different rates of speed in Russia and China. The problem of the Living Buddha was easily solved; no successor was found, and Mongolia was proclaimed a People's Republic—using a wording that exactly follows the Chinese, not the Russian model. The government, and the controlling People's Revolutionary Party, was a coalition of conservative and nationalistic revolutionary elements. Shifts between "rightist" and "leftist" policy ensued, closely affected by the rise of Stalin in the Soviet Union and by the defeat of the Chinese Communists (for the time being) by Chiang Kai-shek. Increasingly important, too, was the Japanese invasion of Manchuria in 1931, followed by their encroachment on Inner Mongolia and North China, and all-out invasion of China in 1937.

Proclamation of the People's Republic

The uncertainty, both internal and external, engendered cliques and conspiracies—some real and others imagined—resulting in an atmosphere of suspicion and fear that cost many lives. The situation began to clear after 1939, when the Japanese, in a thrust toward Siberia, invaded the northeastern corner of Mongolia, testing the Soviet-Mongol alliance. The Mongol border troops fought ferociously and with great professional skill, holding the heights of Nomynkhan and the line of the Khalkyn River until Soviet troops came up. The Japanese defeat was shattering, and it undoubtedly played a major part in their fateful decision not to honour the "Berlin-Tokyo Axis" with Hitler but to make their major effort in the Pacific and Southeast Asia. In Mongolia, the victory dispelled fears that elements of the army might go over to the Japanese; on the contrary, many Inner Mongolian troops recruited by the Japanese went over to the Mongols. The military alliance with the Soviet Union was reconfirmed when the Mongols took part in the short, but devastatingly successful, Soviet campaign against Inner Mongolia and Manchuria in the last two weeks of World War II.

At this time, refugees from Inner Mongolia swarmed into the Mongolian People's Republic. The Japanese had organized a puppet government of Inner Mongolia under Teh Wang (Prince Teh, Demchukdondub). He had, however, tried to minimize Japanese control and to promote Mongol nationalism. When the Chinese Communists came to power, he was condemned as a war criminal but later released.

Postwar plebiscite

Under an agreement made by U.S. Pres. Franklin D. Roosevelt at Yalta, Chiang Kai-shek consented to a plebiscite after the war in Mongolia. The result was overwhelmingly in favour of independence over "autonomy." Full diplomatic recognition did not follow, however, because of a border dispute. Mongolian membership in the United Nations was at first sponsored by Chiang Kai-shek but later opposed by him and by the United States, and it was not until 1961 that Mongolia gained membership. Mongolia now has diplomatic relations with 52 nations, including many in western Europe.

Mongolia's relations with China have been strained because of its unswerving loyalty to the Soviet alliance. Nevertheless, its long frontier with China has been amicably redemarcated. In the meantime, Inner Mongolia has been reorganized as an Autonomous Region within the Chinese People's Republic; but developments in the 1970s indicated that relations were being tightened between the region and the adjacent provinces of China. The Chinese of Inner Mongolia now outnumber the Mongols by perhaps as much as ten to one.

In the Mongolian People's Republic postwar developments were dramatic. Slightly more than 51 percent of all employment became nonpastoral. Coal is mined on a

large scale. Oil was discovered and is refined within the country. Industries include woollen textiles, leather-processing, and meat-packing. Wheat growing has been developed on a large scale, and Mongolia is now a net exporter. Veterinary services—important in a country where livestock outnumber people 25 to 1—are intensively fostered. There is no illiteracy, and industrial employment requires a certificate of eight years of schooling. The urban complex of Ulan Bator, the capital, holds about 25 percent of the population, while rural life is organized under collectives or cooperatives, the members of which are shareholders. The territory of a collective corresponds with that of a county, so that civil life and economic activity are integrated under the Socialist revolution.

BIBLIOGRAPHY. The only good general histories of Mongolia are in Mongol; a history of Mongolia in English, based on work by Mongol and Russian scholars, is being prepared in Moscow under the auspices of UNESCO. For modern history, see C.R. BAWDEN, *The Modern History of Mongolia* (1968), whose author shares the opinion of R.A. RUPEN, *Mongols of the Twentieth Century*, 2 vol. (1964); and G.G.S. MURPHY, *Soviet Mongolia* (1966), that the Mongolian Revolution was largely made in the Soviet Union and exported to Mongolia. For the contrary view that there was widespread support in Mongolia for social and economic as well as political revolution, and genuine Mongol leadership, see OWEN LATTIMORE, *Nationalism and Revolution in Mongolia* (1955), which includes a translation, by the author and URGUNGGE ONON, of the official biography of the revolutionary hero Sükh Baatar; and *Nomads and Commissars: Mongolia Revisited* (1962). For a rounded review, see GERARD M. FRITERS, *Outer Mongolia and Its International Position* (1949). An indispensable handbook for information on contemporary Mongolia is A.J.K. SANDERS, *The People's Republic of Mongolia* (1968).

(O.La.)

Monopoly and Competition

The complexity of market behaviour

This article is concerned with the meaning of the terms monopoly and competition as applied to the complex workings of the marketplace. In popular discussion the issues often become highly simplified. Thus monopoly is usually seen as a kind of octopus devouring the weak, and competition as a band of honest tradesmen serving the public. Economists who analyze actual selling practices find the distinctions among them more difficult to draw. What may be said of the behaviour of sellers in one industry does not necessarily apply to sellers in another. Different industries have different structures, and their structural characteristics determine the behaviour of those who do business. The article begins by defining the fundamental concepts of market structure; it goes on to examine the ways in which the behaviour of sellers differs among industries with different types of market structure; and it ends with an analysis of public policies toward monopoly and competition.

CHARACTERISTICS OF MARKET STRUCTURE

Industries differ, as suggested above, in their structural characteristics. Probably the most critical of these characteristics are (1) the concentration of sellers in an industry, (2) the degree of differentiation among the industry's products, and (3) the ease or difficulty with which new sellers can enter the industry.

Seller concentration. By seller concentration is meant the number of sellers in an industry together with their comparative shares of industry sales. When the number of sellers is quite large, and each seller's share of the market is so small that in practice he cannot, by changing his selling price or output, perceptibly influence the market share or income of any competing seller, economists speak of *atomistic competition*. A more common situation is that of *oligopoly*, in which the number of sellers is so few that the market share of each is large enough for even a modest change in price or output by one seller to have a perceptible effect on the market shares or incomes of rival sellers, and to cause them to react to the change. In a broader sense, oligopoly exists in any industry in which at least some sellers have such large shares of the market, even though there may be an additional number of small sellers. When a single seller supplies the entire output of an industry, and thus can determine his selling price and output without concern for the reactions of rival sellers, a *single-firm monopoly* exists.

Product differentiation. The structure of a market is also affected by the extent to which those who buy from it prefer some products to others. In some industries the products are regarded as identical by their buyers—as, for example, basic farm crops. In others the products are differentiated in some way so that various buyers prefer various products. Notably, the criterion is a subjective one; the buyers' preferences may have little to do with tangible differences in the products but are related to advertising, brand names, and distinctive designs. The degree of product differentiation as registered in the strength of buyer preferences ranges from slight to fairly large, tending to be greatest among infrequently purchased consumer goods and "prestige goods," particularly those purchased as gifts.

Barriers to entry. Industries vary in the ease with which new sellers can enter them. The barriers to entry consist of the advantages that sellers already established in an industry have over the potential entrant. Such a barrier is generally measurable by the extent to which established sellers can persistently elevate their selling prices above minimal average costs without attracting new sellers. The barriers may exist because costs for established sellers are lower than they would be for new entrants, or because the established sellers can command higher prices from buyers who prefer their products to those of potential entrants. The economics of the industry also may be such that new entrants would have to be able to command a substantial share of the market before they could operate profitably.

The effective height of these barriers varies. One may distinguish three rough degrees of difficulty in entering an industry: *blockaded entry*, which allows established sellers to set monopolistic prices, if they wish, without attracting entry; *impeded entry*, which allows established sellers to raise their selling prices above minimal average costs without attracting new sellers, but not as high as a monopolist's price; and *easy entry*, which does not permit established sellers to raise their prices at all above minimal average costs without attracting new entrants.

TYPES OF MARKET STRUCTURE

A system of classification

This discussion of market characteristics suggests a general way of classifying industries according to their market structures:

I. Atomistic competition
  A. With homogeneous products—"pure competition"
  B. With differentiated products—"monopolistic competition"
II. Oligopoly
  A. With homogeneous products—"pure oligopoly"
    1. With blockaded entry
    2. With impeded entry
    3. With easy entry
  B. With differentiated products—"heterogeneous oligopoly"
    1. With blockaded entry
    2. With impeded entry
    3. With easy entry
III. Single-firm monopoly
  A. With blockaded entry
  B. With impeded entry

Under atomistic competition, in which entry is generally easy, there are no barriers to entry. By the same token, product differentiation among sellers is obviously inconsistent with single-firm monopoly.

Comparative importance of each type. The comparative importance of these types of market structure differs among various sectors of the economy. In the manufacturing sector in the United States, which includes about 400 industries, single-firm monopolies are almost completely absent. But in more than half of the manufacturing industries there is enough seller concentration for them to qualify as oligopolies. The remaining industries are more or less atomistic in their market structure. An appreciable degree of product differentiation is found in

about half of the oligopolistic industries and in about half of the atomistic industries. Very strong product differentiation is usually found among oligopolistic industries. Easy entry is typical of atomistic industries, impeded entry of oligopolies. Entry is probably blockaded in a minor fraction of the latter, generally those with very high seller concentration.

The proportions of oligopolistic and atomistic manufacturing industries are about the same in the United Kingdom as in the United States. The incidence of oligopolies is slightly higher in Japan and progressively higher in France, Italy, Canada, and Sweden. Single-firm monopolies in manufacturing are found in a few industries in some of these countries, but they are typically under government ownership.

In the public-utility sector in the United States, single-firm monopolies are typically found in industries supplying gas, electricity, and telephone and telegraph service. Oligopoly, frequently highly concentrated, is typical in the radio and television and transportation industries; entry into these industries is usually very difficult or blockaded. The significance of such structural conditions is lessened, however, by the fact that these industries are subject to various degrees of public regulation. The situation is much the same in other Western countries, except that public utilities are frequently under government ownership.

In the distributive trades (wholesaling and retailing), a number of United States industries are fairly atomistic, while a somewhat larger number are relatively unconcentrated oligopolies in which a few large sellers supply about half the industry's output and a very large number of small sellers supply the remainder. Product differentiation is important. Entry is relatively easy. The service-trade industries in the U.S. display a similar range of characteristics. In the distributive- and service-trade sectors in other Western countries, oligopoly is less frequent and atomistic industries are proportionally more important. The residential-construction industries in the U.S. and elsewhere are relatively atomistic in structure, have significant product differentiation, and are easy to enter. Industries in the agricultural sectors of Western countries generally are typically atomistic in structure, with easy entry. But the significance of these structural conditions is lessened by governmental interference designed to modify the working of market forces.

MARKET CONDUCT AND PERFORMANCE

Sellers' behaviour in different market structures

How do sellers behave in determining their selling prices, outputs, advertising costs, and so forth; and in what ways does this market behaviour differ among industries with different types of market structure? An educated layman might ask, for example, whether sellers cut their selling prices in order to take customers away from each other until some rock-bottom market price is reached just high enough to allow them minimal interest returns on their investments or whether, on the other hand, they agree with each other to set a uniform higher price well above their production costs, sharing the market and reaping excessive profits.

It is helpful to distinguish the related ideas of *market conduct* and *market performance*. Market conduct refers to the price and other market policies pursued by sellers, in terms both of their aims and of the way in which they coordinate their decisions and make them mutually compatible. Market performance refers to the end results of these policies—the relationship of selling price to costs, the size of output, the efficiency of production, progressiveness in techniques and products, and so forth.

Behaviour measuring

Pure competition. Market conduct and performance in atomistic industries provide good standards against which to measure behaviour in other types of industry. The atomistic category includes both pure competition and monopolistic competition. In pure competition, a large number of small sellers supply a homogeneous product to a common buying market. In this situation no individual seller can perceptibly influence the market price at which he sells but must accept a market price that is impersonally determined by the total supply of the product offered by all sellers and the total demand for the product of all buyers. The large number of sellers precludes the possibility of a common agreement among them, and each must therefore act independently. At any going market price, each seller tends to adjust his output to that quantity that will yield him the largest aggregate profit, assuming that the market price will not change as a result. But the collective effect of such adjustments by all sellers will cause the total supply in the market to change significantly so that the market price falls or rises. Theoretically, the process will go on until a market price is reached at which the total output that sellers wish to produce is equal to the total output that all buyers wish to purchase. This way of reaching a provisional equilibrium price is what Adam Smith was referring to when he wrote of prices being determined by "the invisible hand" of the market.

The competitive ideal

If the provisional equilibrium price is high enough to allow the established sellers profits in excess of a normal interest return on investment, then added sellers will be drawn to enter the industry, and supply will increase until a final equilibrium price is reached that is equal to the minimal average cost of production (including an interest return) of all sellers. Conversely, if the provisional equilibrium price is so low that established sellers incur losses, some will tend to withdraw from the industry, and supply will decline until the same sort of long-run equilibrium price is reached.

The long-run performance of a purely competitive industry therefore embodies these features: (1) industry output is at a feasible maximum and industry selling price at a feasible minimum; (2) all production is undertaken at minimum attainable average costs, since competition forces them down; and (3) income distribution is not influenced by the receipt of any excess profits by sellers.

This performance has often been applauded as ideal from the standpoint of general economic welfare. But the applause, for several reasons, should not be unqualified. Pure competition is truly ideal only if all or most industries in the economy are purely competitive and if in addition there is free and easy mobility of productive factors among industries. Otherwise, the relative outputs of different industries will not be such as to maximize consumer satisfaction. There is also some question whether producers in purely competitive industries will generally earn enough to plow back some of their earnings into improved equipment and thus maintain a satisfactory rate of technological progress. Finally, some purely competitive industries have been afflicted with "destructive competition"—the coal industry and the basic agricultural industries, for example. For some historical reason such an industry accumulates excess capacity to the point where sellers suffer chronic losses, and the situation is not corrected by the exit of people and resources from the industry. The invisible hand of the market works too slowly for society to accept. In some cases, notably in agriculture, government has intervened to restrict supply or raise prices. Leaving these qualifications aside, however, the market performance of pure competition furnishes some sort of a standard to which the performance of industries of different structure may be compared.

Monopolistic competition. In the more complex situation of monopolistic competition (atomistic structure with product differentiation) market conduct and performance may be said to follow roughly the tendencies attributed to pure competition. The principal differences are the following. First, individual sellers, because of the differentiation of their products, are able to raise or lower their individual selling prices slightly; they cannot do so by very much, however, because they remain strongly subject to the impersonal forces of the market operating through the general level of prices. Second, rivalry among sellers is likely to involve sales promotion costs as well as the expense of altering products to appeal to buyers. This is a competitive game that all will play but that nobody, on the average, will win, and the long-run equilibrium price will reflect the added costs involved. In return, however, buyers will get more vari-

ety. Third, since not every seller is likely to be equally successful in his sales-promotion and product policies, some will receive profits in excess of a basic interest return on their investment; such profits will come from their success in winning buyers. Monopolistic competition may, like pure competition, include industries that are afflicted with what has been called above destructive competition. This may result not only from a failure to get rid of excess capacity but also from the entry of too many new firms despite the danger of losses.

Monopoly. While single-firm monopolies are rare, except for those that are subject to public regulation, it is useful to examine the monopolist's market conduct and performance in order to establish a standard at the other pole from pure competition. As the sole supplier of a distinctive product, the monopolist can set whatever selling price he chooses provided he is willing to accept the sales that will correspond to that price. Since the market demand will generally be less the higher the price he sets, the monopolist presumably will set that price that produces the greatest profits, given the relationship of production costs to output. By restricting output he can raise his selling price significantly—an option not open to sellers in atomistic industries.

Monopolistic behaviour

The monopolist will generally charge prices well in excess of production costs and reap profits well above a normal interest return on investment. His output will be substantially smaller, and his price higher, than if he had to meet established market prices as in pure competition. The monopolist may or may not produce at minimal average cost, depending on his cost-output relationship; if he does not, there are no market pressures to force him to do so.

If the monopolist is subject to no threat of entry by a competitor, he will presumably set a selling price that maximizes profits for the industry he monopolizes. If he faces only impeded entry, he may elect to charge a price sufficiently low to discourage entry but above a competitive price—if this will maximize his long-run profits.

Oligopoly. Market conduct and performance in oligopolistic industries generally combine monopolistic and competitive tendencies, with the relative strength of the two tendencies depending roughly on the detailed market structure of the oligopoly.

Rivalry among sellers. In the simplest form of oligopolistic industry, sellers are few and every seller supplies a sufficiently large share of the market so that any feasible and modest change in his policies will appreciably affect the market shares of all his rival sellers and induce them to react or respond. For example, if seller A reduces his selling price below the general level of prices being charged by all sellers sufficiently to permit him to capture significant numbers of customers from his rivals if they hold their selling prices unchanged, they may react by reducing their prices by a similar amount, so that none gains at the expense of others and the group has probably reduced its combined profits. Or seller A's rivals may retaliate by reducing their selling prices more than he did, thus forcing a further reaction from him. Conversely, if seller A increases his selling price above the general level being charged by all sellers (thus tending to lose at least some of his customers to his rivals), they may react by holding their prices unchanged, in which event seller A will probably retract his increase and bring his price back to the previous level. But his rivals may also react by raising their prices as much as seller A raised his, in which case the general level of prices in the industry rises and the combined profits of all sellers are probably increased.

Any seller A in an oligopoly will therefore determine whether or not to alter his price or other market policy in the light of his conjectures about the reactions of his rivals. Correspondingly, his rivals will determine their reactions in the light of their conjectures about what seller A will do in response. The process is not likely to bring the industry price level down to minimal average cost (as in atomistic competition). Many different "equilibrium" levels between the competitive and monopolistic limits are possible, depending on further circumstances.

Thus in an oligopoly viable collusive agreements among rival sellers are quite possible. They may be express agreements established by contract or tacit understandings that develop as a pattern of reactions among sellers to changes in each others' prices or market policies becomes customary. In the United States, express collusive agreements are forbidden by law, but tacit agreements or "gentlemen's understandings" are common in oligopolies. In numerous other Western countries, formal collusive agreements (often called cartels if comprehensive in scope) are legal. Whether tacit or explicit, legal or illegal, one may say that oligopolistic prices tend to be "administered" by sellers, in the senses mentioned above, as distinct from being determined by impersonal market forces.

Sellers' dual aims. The varying market performance of oligopolies results from the fact that individual sellers intrinsically have two conflicting aims. One common desire is to establish among themselves a monopolistic level of price (and of selling costs, etc.), which will maximize their combined profits, giving them the largest "profit pie" to divide. But each seller also has a fundamental antagonism toward rival sellers and wants to maximize his own profits even at the expense of theirs. The relative strengths of these conflicting aims—the maximizing of combined profits and the maximizing of individual profits—will likely depend on how concentrated the oligopoly is, because when sellers are fewer and their individual market shares larger, their rivals' reactions are stronger deterrents to independent actions.

Price behaviour in oligopoly

This is why various sorts of market performance are to be expected in oligopolistic industries. When the entry of other sellers is blockaded, collusive or interdependent behaviour may lead to a full monopoly price. If entry is only impeded, the resulting price may be far enough below the full monopoly level to discourage further entry. But prices are not always what they seem. An announced price that is well above cost may be undercut by clandestine price reductions to individual buyers, bringing the average of actual selling prices down somewhat.

If an oligopolistic industry is made up of a "core" of a few large interdependent sellers plus a "competitive fringe" of several or numerous quite small sellers, the competition of the small sellers may induce the large ones to limit the extent to which they raise their prices.

Price behaviour approaching full monopoly pricing seems to be found mainly in oligopolies having very high seller concentration and blockaded entry. Where these characteristics are less pronounced, prices and profits tend to be lower, though they are likely to be somewhat above the competitive level. A few economists maintain, mainly on the basis of abstract theorizing, that oligopolistic prices in general do not significantly differ from atomistically competitive prices, but the bulk of statistical evidence does not support them.

In oligopolies in which product differentiation is important, sales-promotion costs and the costs of product improvement or development will display roughly the same variety of tendencies found in pricing. Where there are a few large interdependent sellers, these costs may be restricted to about the same level as those of a single-firm monopolist; on the other hand, rivalry in sales promotion and product development may be sufficient to raise them higher. Oligopolists may also arrive collusively at relatively high uniform selling prices but simultaneously engage in independent nonprice competition (perhaps more so where seller concentration is lower).

WORKABLE COMPETITION

Since the character of market performance varies among industries along with their market characteristics, efforts have been made to devise some practical standard for identifying the sorts of market structure that engender socially satisfactory performance in a given industry. The term workable competition was coined to denote competition that may be considered as leading to a reasonable or socially acceptable approximation to ideal performance in the circumstances of a particular in-

dustry. The limits of such an approximation are of course debatable, and so the idea of workable competition must remain elusive because it is basically subjective.

Five attributes of workable competition

Without entering into a complex theoretical discussion of the relationship of individual-industry performance to overall welfare, it is plausible to suggest the following principal attributes of workable performance in an industry: (1) In the long term, selling price on the average should be equal to or not significantly above average costs of production, so that profits do not appreciably exceed a normal interest return on investment. Prices should be responsive to basic reductions in costs. (2) In so far as average costs of production are affected by the scales or capacities of plants and firms, the preponderance of industry output should be from plants and firms of the most efficient scale or with closely comparable technical efficiency. (3) The industry should not have chronic excess capacity—*i.e.*, significant plant capacity that is persistently unused even in periods of high general economic activity. (4) The industry's sales-promotion costs should not be substantially greater than needed to keep buyers informed of the availability, characteristics, and prices of products. (5) The industry should be adequately progressive in introducing more economical production techniques and improved products—balancing the costs of progress with the gains.

While the first three of these attributes are easier to appraise than the others, certain generalizations are possible concerning the workability of different market structures: (1) Unregulated single-firm monopolies tend to generate unworkable market performance, mainly in the form of output restriction, prices well above costs, and consequent excess profits. They have undesirable effects on the uses to which resources are put and on income distribution. (2) Oligopolies with high seller concentration and also very high barriers to entry tend toward unworkable performance, like that of single-firm monopoly. In general, however, they do not show significant degrees of technical inefficiency resulting from inefficient plant scales or excess capacity. (3) Oligopolies with fairly high seller concentration but only moderate barriers to entry are also prone to unworkable performance of the sort just mentioned, but not to as high a degree. (4) Oligopolies with only moderate seller concentration and moderate-to-low barriers to entry tend toward workable performance both in price-cost relations and in technical efficiency, except that some of them may have recurrent chronic excess capacity due to periodic overentry. (If cartels are legalized and their provisions are not rigorously controlled by government, the last two categories of oligopoly may have the same sort of unworkable performance as do very highly concentrated oligopolies.) (5) Industries of atomistic structure tend generally toward workable performance unless they suffer from destructive competition as described above.

In industries with significant differentiation of products among sellers—and especially in oligopolies of this sort—there is a tendency for minor but significant fractions of income to be devoted to persuasive (as distinct from informational) advertising and other sales promotion and also to more or less idle variations of product design, with the result that resources are in a sense "wasted" and costs increased.

By the criteria of workable competition, a purely rational society would presumably favour industries with moderate to low seller concentration, moderate to low barriers to entry, and without extreme product differentiation—all this from the standpoint of enhancing overall material welfare. The argument that oligopolistic and atomistic industries generally need legal protection from destructive competition may be discarded on the basis of evidence. Price and other market warfare in such industries has been extremely rare in industrial countries in the last 50 years.

GOVERNMENT POLICIES

Government policies toward monopoly and competition vary considerably among Western countries. These countries differ in their treatment both of cartels or other collusive arrangements on the one hand and of very high or increasing concentration on the other. The actual policies applied are not always spelled out in full by the legislation supporting them; the vigour of administration varies widely and so do the attitudes of the public commissions that administer the laws.

Efforts to maintain competition

The United States has in general the longest standing and the strongest policy of maintaining effective competition in its economy, under a variety of statutes usually designated as antitrust laws. This article will deal with two major parts of the relevant legislation, and the way they have been administered in recent years.

The Sherman Act. The basic law is the Sherman Anti-Trust Act (passed in 1890). It declares illegal every agreement in restraint of interstate trade among competing sellers, and its prohibitions apply not only to formal cartels but also to any agreements to fix prices, limit outputs, share markets, exclude competitors, and so forth. The existence of an illegal collusive agreement need not depend on a written document but may be, and most frequently is, inferred from circumstantial evidence. The reasonableness of such an agreement in purpose or effect does not make it legal; the intent to restrict competition alone suffices to establish its illegality.

The Sherman Act also declares illegal the "monopolization" of any part of trade or commerce in the country. As currently construed by the courts, this means that no firm may gain or maintain a "dominant occupancy" of an industry through "predatory" or "exclusionary" tactics or through normal commercial practices having the same purpose and effect. This prohibition may also be applied to several dominant oligopolists in an industry who jointly and in concert similarly act to eliminate or exclude competitive sellers.

Both major provisions of the Sherman Act are enforceable by the Department of Justice through litigation in the federal courts. It may bring either "criminal" actions, seeking fines and possible imprisonment as penalties for violations, or "civil" actions seeking court orders to terminate violations. The civil remedy for an agreement restraining trade is usually an injunction forbidding continuation of the agreement. Civil remedies for monopolization may be confined to injunctions against various "monopolizing" activities or at the extreme may extend to ordering dismemberment of the offending firm or firms to create a more competitive market structure.

The prohibition against collusive agreements has been fairly vigorously and successfully enforced in recent decades with the result that many such agreements have been terminated, but the scope of enforcement is probably not great enough to reach more than a minor fraction of potential violators. Moreover, simple parallelism of pricing and other action within concentrated oligopolies is alone not sufficient evidence of violation of the law. The major effects of the prohibition of collusive agreements, aside from actually stopping some of them, are two. First, there are no cartelized industries in the United States, so that the public does not suffer from the most effective and potentially pernicious sort of collusive agreement. Second, such collusion as does exist is tacit, driven "underground" to the level of gentlemen's agreements and is correspondingly much less effective in restricting competition than are cartels.

The prohibition against monopolization has been periodically enforced but never intensively, in part because the substantial dominance of an industry by one firm is very rare in the United States, and in part because of difficulties in demonstrating that probable offenders have actually pursued the practices and policies forbidden. In general, successful actions for joint monopolization cannot be brought against a very large majority of oligopolistic industries in the U.S. But a number of actions have succeeded in establishing illegal monopolization, and in some of these the severe penalty of dismemberment of the offending firm or firms has been secured. The major effects of the enforcement of the antimonopolization law have probably been to discourage the growth of single-firm monopolies and inferentially to encourage the development of oligopolies as a substi-

tute. Enforcement has also substantially discouraged the use of predatory and exclusionary tactics to eliminate or exclude competitors.

The Clayton Act. The other most significant procompetition law in the United States is the Celler-Kefauver Act of 1950, amending Section 7 of the older Clayton Anti-Trust Act. It prohibits mergers of firms, or acquisition of one firm by another, whenever the effect may be substantially to lessen competition or to tend to create a monopoly. It is directed against increases in concentration that probably or possibly will restrict competition.

This law has been rather intensively enforced with the result that a large number of important mergers or acquisitions have been dissolved or stopped, and many others have not been attempted. Continued enforcement of the law has been encouraged by judicial interpretations giving it very broad application; relatively minor effects on competition have been considered substantial enough to make the mergers that would produce them illegal, and even mergers of companies in different industries have been forbidden if some substantial lessening of competition might result. The Celler-Kefauver Act has thus served as a major deterrent to increases of concentration by merger. Nonetheless, a large amount of merger activity has not been deterred by the law, especially "conglomerate" mergers of firms in different and quite unrelated industries.

Before leaving the subject of United States policy it should be noted that, in some industries, the federal government has actually endeavoured to restrict competition. These include industries afflicted with chronic distress or destructive competition, as in agriculture, and industries in which large firms are tending to displace small firms, as in the distributive trades. In agriculture, for example, the government has imposed or sponsored price supports and output restrictions of crops in order to raise prices above competitive levels. In the distributive trades, various measures have aimed to reduce competition and lessen the competitive advantages of chain stores and other mass distributors, but with indifferent success. Both sorts of competition-restricting policies have been poorly designed to remedy basic ills afflicting the industries in question, being oriented much more to preserving "competitors" than to preserving competition.

European policies. Other industrialized Western countries follow policies toward monopoly and competition that differ a good deal from those in the United States. They place much less emphasis on competition or tend to legalize agreements in restraint of competition subject to governmental supervision. Such procompetitive policies as they do have are generally much "softer" than the American, and less comprehensive.

How national attitudes differ

These other Western countries, with minor exceptions, do not have laws prohibiting monopoly or monopolization nor laws preventing mergers and acquisitions that would tend to lessen competition. High concentration within industries, and increasing concentration, are not officially discouraged; in a number of countries the public authorities even encourage increased concentration via mergers or otherwise, although some countries have drawn the line at the acquisition of control of their domestic companies by foreign companies.

In some Western countries—*e.g.*, France and especially Italy—government-owned firms have been established in a number of major industries in which they operate in general rivalry with private firms. But whether their price policies check or modify the monopolistic pricing tendencies of private firms is not readily apparent.

The general tendency in these other Western countries is either to accept and recognize formal cartels or, alternatively, to forbid cartels but accept less formal agreements in restraint of trade, and then to supervise and to some extent control these cartels or other agreements through governmental commissions. In France and Italy, for example, formal cartels are permitted and recognized, but a commission investigates them and distinguishes between "good" (acceptable) and "bad" (unacceptable) cartels. Since the applied criteria of what is good and bad are not evident, the extent to which cartelized restriction of competition is reduced is unclear. Sweden permits numerous cartels but requires them to register with a government agency that employs publicity and the pressure of public opinion to combat antisocial cartel practices. Japan permits or requires cartels on a wide variety of grounds and apparently does not closely supervise them. In the United Kingdom, formal cartels are prohibited but looser agreements in restraint of competition are allowed; the latter are investigated by a commission and eliminated "if necessary."

In general, the policies toward cartels and agreements restraining competition in other Western countries are much more permissive than in the United States. Correspondingly, monopolistic or quasi-monopolistic tendencies in pricing are doubtless stronger and more frequent, and the automatic regulatory forces of markets weaker.

TRENDS AND ALTERNATIVES

The stability of present market structures

There seem to be no strong recent trends in Western countries toward either more or less competition within industries. In the United States, the average degree of seller concentration in individual industries has been remarkably stable since the 1930s. For example, roughly a third of the manufacturing industries show no appreciable change in concentration, another third show increases in concentration, and the remaining third show decreases of about the same magnitude. Few of the changes in concentration have been dramatically large. Between 1947 and 1963 the share of business in all manufacturing industries done by the largest 25 to 50 firms increased somewhat, but this is reconcilable with the previous statement on a number of grounds—*e.g.*, the largest firms are located in the largest and most rapidly growing industries. In other Western countries there is little evidence of dramatic increases in concentration in the postwar period. Some of these countries, like Japan and Germany, were in the process of re-establishing prewar concentrations that had been broken down under occupation governments in the immediate postwar period.

Restriction of competition by cartels or other agreements is certainly no more important in the United States now than in the 1930s and perhaps less so because of more intensive enforcement of the antitrust laws since 1938. A significant minor fraction of industrial output is now purchased by departments of government—especially the military—through negotiation with sellers, and this sort of buying practice is probably somewhat restrictive of competition. In a number of other Western countries the postwar tendency to subject cartels to some governmental supervision and regulation probably tended to lessen the restrictive effects of cartels. The same general tendency was observable in countries like the United Kingdom, where stricter supervision and control of looser types of collusive agreements was imposed.

A final question is whether broad changes in governmental policies toward monopoly and competition are desirable. The argument in favour of increasing seller concentration and stronger cartelization in order to avert cutthroat competition and to reduce uncertainty on the part of business firms is not well supported in theory and is discredited by experience, particularly that of the United States. The price paid would be unduly high in terms of misallocation of resources, effects on income distribution, and deterrents to the sort of progress that active market rivalry generates. On the other hand, in that minority of industries chronically afflicted with destructive competition, constructive steps should be taken to eliminate excess capacity and to shift labour to other occupations.

Some critics of American antitrust policies have recommended that highly concentrated oligopolies that persistently generate unworkable performance be dismembered (unless this would destroy the cost-savings of very large firms), thus creating more numerous and smaller sellers. This is probably not politically feasible in the United States or abroad, but it has its attractions.

BIBLIOGRAPHY. E.H. CHAMBERLIN, *Theory of Monopolistic Competition*, 8th ed. (1962), a classic and germinal contribution to the theory of markets, notable for its revival and development of the theory of oligopoly and for its introduction

and development of the concept of product differentiation; J. ROBINSON, *Economics of Imperfect Competition* (1933), a penetrating theoretical analysis of monopolistic and quasi-monopolistic pricing, complementing Chamberlin's work in many important ways; W.J. FELLNER, *Competition Among the Few: Oligopoly and Similar Market Structures*, rev. ed. (1949), a highly sophisticated and stimulating development of the theory of oligopoly; G.J. STIGLER, *The Theory of Price*, 3rd ed. (1966), a superior general textbook treatment of price theory, furnishing background for the somewhat more specialized works previously cited; E.S. MASON, "Price and Production Policies of Large-Scale Enterprise," *American Economic Review*, vol. 29 suppl., pp. 67–74 (May 1939), a germinal essay of great importance that pointed the way to feasible techniques for exploring the relation of market structures to market performance; J.S. BAIN, *Barriers to New Competition: Their Character and Consequences in Manufacturing Industries* (1956), introduction of the notion of barriers to entry and discussion of their character, sources, and consequences as aspects of market structure, including empirical tests of their effects on competition; and *Industrial Organization*, 2nd ed. (1968), a general textbook covering theoretical and empirical aspects of problems of monopoly and competition, including a treatment of American public policies promoting or restraining competition; C. KAYSEN and D.F. TURNER, *Antitrust Policy: An Economic and Legal Analysis* (1959), a superior analysis and criticism of the working of American antitrust policies, coupled with detailed proposals for revision in these policies; C. WILCOX, *Public Policies Toward Business*, 3rd ed. (1966), a high-grade general textbook on public policies toward business, including extensive treatment of policies affecting monopoly and competition.

(J.S.B.)

Monotheism

Monotheism and polytheism

Monotheism and polytheism, the beliefs in one god or many gods, are often thought of in rather simple terms; *e.g.*, as a merely numerical contrast between the one and the many. The history of religions, however, indicates many phenomena and concepts that should warn against oversimplification in this matter. There is no valid reason to assume, for example, that monotheism is a later development in the history of religions than polytheism. There exists no historical material to prove that one system of belief is older than the other, although many scholars hold that monotheism is a higher form of religion and, therefore, must be a later development, assuming that what is higher came later. Moreover, it is not the oneness of god that counts in monotheism but his uniqueness; one god is not affirmed as the logical opposite to many gods but as an expression of divine might and power.

The choice of either monotheism, or polytheism, however, leads to problems, because neither can give a satisfactory answer to all questions that may reasonably be put. The weakness of polytheism is especially revealed in the realm of questions about the ultimate origin of things, whereas monotheism runs into difficulties in trying to answer the question concerning the origin of evil in a universe under the government of one god. There remains always an antithesis between the multiplicity of forms of the divine manifestations and the unity that can be thought or posited behind them. The one and the many form no static contradistinction: there is rather a polarity and a dialectic tension between them. The history of religions shows various efforts to combine unity and multiplicity in the conception of the divine. Because Christianity is a monotheistic religion, the monotheistic conception of the divine has assumed for Western culture the value of a self-evident axiom. This unquestioned assumption becomes clear when it is realized that for Western culture there is no longer an acceptable choice between monotheism and polytheism but only between monotheism and atheism.

THE SPECTRUM OF VIEWS: MONOTHEISMS AND QUASI-MONOTHEISMS

The basic monotheistic view. Monotheism is the belief in the existence of one god or, stated in other terms, that God is one. As such it is distinguished from polytheism, the belief in the existence of a number of gods, and atheism, the denial of the belief in any god or gods at all. The God of monotheism is the one real god that is believed to exist or, in any case, that is acknowledged as such. His essence and character are believed to be unique and fundamentally different from all other beings that can be considered more or less comparable; *e.g.* the gods of other religions. The religious term monotheism is not identical with the philosophical term monism, referring to the view that the universe has its origin in one basic principle (*e.g.*, mind, matter) and that its structure is one unitary whole in accordance with this principle; that is, that there is only a single kind of reality, whereas, for monotheism, there are two basically different realities: God and the universe.

The God of monotheism

God in monotheism is conceived of as the creator of the world and man; he has not abandoned his creation but continues to lead it through his power and wisdom; hence, viewed in this aspect, history is a manifestation of the divine will. God has not only created the natural world and the order existing therein but also the ethical order to which man ought to conform and, implicit in the ethical order, the social order. Everything is in the hands of God. God is holy—supreme and unique in being and worth, essentially other than man—and can be experienced as a *mysterium tremendum* ("a fearful mystery") but at the same time as a *mysterium fascinans*, ("a fascinating mystery"), as a mystery approached by man with attitudes of both repulsion and attraction, of both fear and love. The God of monotheism, as exemplified by the great monotheistic religions—Judaism, Christianity, and Islām—is a personal god. In this respect the one god of monotheism is contrasted with the conception of the divine in pantheism, which may also affirm one god or a divine unity. The god of pantheism, however, is impersonal, rather a divine fluid that permeates the whole world including man himself, so that Hinduism can say: *tat tvam asi*, literally "that is you," where "that" refers to the single, supreme reality or principle.

In monotheistic religions the belief system, the value system, and the action system are all three determined in a significant way by the conception of God as one unique and personal being. Negatively considered, the monotheistic conviction results in the rejection of all other belief systems as false religions, and this rejection partly explains the exceptionally aggressive or intolerant stance of the monotheistic religions in the history of the world. The conception of all other religions as "idolatry" (*i.e.*, as rendering absolute devotion or trust to what is less than divine) has often served to justify the destructive and fanatical action of the religion that is considered to be the only true one.

The concept of idolatry

The symbolic language of the monotheistic belief system has no proper terms of its own in speaking of God that cannot be found elsewhere also. God as Creator, Lord, King, Father, and other descriptive names are expressions found in many religions to characterize the various divine beings; the names do not exclusively belong to the religious language of monotheism. This common language is understandable because the monotheistic conception of God differs essentially only in one respect from that of other religions: in the belief that God is one and absolutely unique. Then, consequently, God is regarded as the one and only Creator, Lord, King, or Father. The conception of a divine Word is also to be found in a large number of religions, in accordance with the widespread belief that creation takes place through the word, or speech, of a god.

The extreme positions. The above is the basic monotheistic view. There is, however, a wide range of positions between exclusive monotheism at one extreme and unlimited polytheism at the other. A survey of the various positions may serve to provide a more adequate picture of the complex reality involved in the monotheisms and quasi-monotheisms.

Exclusive monotheism. For exclusive monotheism only one god exists; other gods either simply do not exist at all, or, at most, they are false gods or demons; *i.e.*, beings that are acknowledged to exist but that cannot be compared in power or any other way with the one and only true God. This position is in the main that of Judaism, Christianity, and Islām. While in the Old Testa-

ment the other gods in most cases were still characterized as false gods, in later Judaism and in Christianity as it developed theologically and philosophically the conception emerged of God as the one and only, and other gods were considered not to exist at all.

Ethical and intellectual monotheism

There are two types of exclusive monotheism: ethical monotheism and intellectual monotheism. In ethical monotheism man chooses one god, because that is the god whom he needs and whom he can adore, and that god becomes for him the one and only god. In intellectual monotheism the one god is nothing but the logical result of questions concerning the origin of the world. In many African religions the one god postulated behind the many gods that are active in the world and in the life of man is little more than the prime mover of Aristotle's philosophy, the ultimate source of motion in the universe. He is the intellectual apex necessitated by the system. In Christian theology, heavily influenced as it is by Greek philosophy, both conceptions can be found, usually together.

Unlimited polytheism. On the other hand, there is the extreme position of unlimited polytheism as, for instance, in the classical religions of Greece and Rome: each god has his own name and his own shape, and these are unalienably his and cannot be exchanged with those of any other god (not counting, of course, those cases in which gods are practically each other's duplicate and only bear a different name). The number of divinities is large and in principle unlimited. There are differences of status and power among the gods, of function and sphere of influence, but they are all equally divine. There is, in fact, an ordered pantheon. In unlimited polytheism, the number of gods that are actually worshipped seldom exceeds a few hundred within one religion, but in theory, as in India, millions and millions of gods may be thought to exist (see also POLYTHEISM).

The middle positions. Between the extremes of exclusive monotheism and unlimited polytheism are the middle positions of inclusive monotheism and henotheism.

Inclusive monotheism. Inclusive monotheism accepts the existence of a great number of gods but holds that all gods are essentially one and the same, so that it makes little or no difference under which name or according to which rite a god or goddess is invoked. Such conceptions characterized the ancient Hellenistic religions. A well-known example is that of the goddess Isis in the Greco-Roman mystery religion that is called after her. In *The Golden Ass* of Apuleius, the goddess herself speaks: "My name, my divinity is adored throughout all the world, in divers manners, in variable customs, and by many names." Then there follows a number of divine names, and this enumeration ends: "And the Egyptians, which are excellent in all kind of ancient doctrine, and by their proper ceremonies accustomed to worship me, do call me by my true name, Queen Isis."

Henotheism, or kathenotheism. Henotheism (from Greek *heis theos,* "one god")—a belief in worship of one god, though the existence of other gods is granted—also called kathenotheism (Greek *kath'hena theon,* "one god at a time")—which literally implies worship of various gods one at a time—has gone out of fashion as a term. It was introduced by the eminent 19th-century philologist and scholar in comparative mythology and religion Max Müller (1823–1900). Many later authors prefer the term

Monolatry

monolatry—which is the worship of one god, whether or not the existence of other deities is posited—to the term henotheism. Both terms mean that one god has a central and dominating position in such a way that it is possible to address this god as if he were the one and only god, without, however, abandoning the principle of polytheism by denying or in any other way belittling the real existence of the other gods, as the above-mentioned forms of monotheism do. Henotheism as a religious concept is at home in cultures with a highly centralized monarchical government. It was especially prevalent in some periods in the history of Babylonia and Egypt.

Alternate positions. *Pluriform monotheism.* The complicated relations that exist between monotheism and polytheism become clear when pluriform monotheism is considered, in which the various gods of the pantheon, without losing their independence, are at the same time considered to be manifestations of one and the same divine substance. Pluriform monotheism is one of the efforts to solve the problem of the coexistence of divine unity and divine pluriformity (multiplicity of forms), which was not recognized by an older generation of scholars, although part of the material was already available. It seems, indeed, that in many parts of the world and in many times religious thinkers have struggled with the perplexing problem of the unity and the pluriformity of the divine.

Nuer and Shilluk religion

The Nuer, a Nilotic pastoral people of the eastern Sudan, venerate a being called Kwoth, the Nuer term for "spirit" (also translated as "God"). He is considered to be the spirit in or of the sky. Like all spirits Kwoth is invisible and omnipresent, but he manifests himself in a number of forms. Each of these manifestations bears a name of its own, but though they are addressed and treated as separate entities, they are essentially nothing but manifestations of the one spiritual being Kwoth and are themselves considered spirits and called *kwoth.* A sacrifice offered to one of these manifestations—*e.g.,* a spirit of air, totem, or place—is not at the same time an offering to another; but all sacrifices, to whatever spirit they are offered, are sacrifices to the supreme Kwoth, or God. Nuer religion is certainly no clear monotheism as it is understood in the Bible and in the Qur'ān (the sacred book of Islām), but neither is it polytheism in the popular sense of the word.

The case of the Nuer is not unique. The related Shilluk people have similar conceptions, and here again the idea of a kind of divine substance that manifests itself in various shapes and under different names is encountered. To give one instance, Macardit is God, but this pronouncement cannot be turned the other way round—it is not permissible to say that God is Macardit. The divine being Macardit represents the dire and fatal aspect of the divinity who orders everything; that is to say, who also sends misfortune and death. In Macardit the contradiction between the creative and constructive and the destructive forces of the divinity is resolved. The positive function of this representation of God lies in the fact that, without diminishing either the power or the justice of the total divinity, it enables man to find an answer for the vexing question of theodicy—the problem of affirming divine justice and goodness in the face of physical and moral evil. That this question is a difficult one, indeed, becomes clear when the reactions of the tribes of Patagonia in a case of death are compared. These tribes believe in a high god, a supreme being, who rules everything and is also responsible for misfortune and death. When someone dies they accuse their god of murder.

Many other instances of pluriform monotheism could be mentioned, and many more presumably still await detection. An interesting pluriform system is that of the Oglala Indians of the central plains of the United States, who venerate 16 gods divided into four groups of four. Each group of four forms one god. Thus there are four gods, but these four gods again are one god, Wakan Tanka—the Great Spirit or the Great Mystery.

Religious dualism. Some religions are in the main dualistic; they view the universe as comprising two basic and usually opposed principles, such as good and evil or spirit and matter. Insofar as the conception of a god and antigod rather than that of two gods is encountered, this kind of religion can be considered as another variation of monotheism. In some Gnostic systems (ancient heresies based on esoteric knowledge and the dualism of matter and spirit), Christianity came near to this idea: the demiurge who created the world and man is considered as an evil being and contrasted to the good god. The most important instance of a dualistic religion is the Persian religion Zoroastrianism as founded by Zoroaster (7th–6th century BC) in which Ormazd (the good god) and Ahriman (the evil god) are each other's opposite and implacable enemies. Dualism, the existence of two contrary and, as a rule, mutually inimical principles, must not be confused with the notion of polarity, in which both principles are mutually dependent so that the one cannot

Zoroastrian dualism

exist without the other. Within the religion of Zoroaster, this notion is also found. In the Zoroastrian variation known as Zurvanism, as it is called after the god Zurvān *akarana* (boundless time), good and evil proceed from one and the same source and in the end they come together again (see also DUALISM, RELIGIOUS).

Pantheism and panentheism. Pantheism and panentheism are not necessarily connected with the notion of either monotheism or polytheism. In both cases the conception of the god or gods is impersonal, which tends, of course, to the conception of one god, of one divine substance, like Spinoza's *deus sive natura,* "god or nature." In pantheism god is immanent, in monotheism god is mostly transcendent, but in polytheism the gods may be either. Pantheism, however, is in most cases more a philosophical than a religious category. Sometimes the term panentheism is used to distinguish between the view that all is in God and that god is in all (see also PANTHEISM AND PANENTHEISM).

Primitive monotheism. In connection with monotheism it is necessary to mention the so-called high gods—the remote gods, usually sky gods, found in many primitive and archaic cultures—because this type of divine being has given rise to the theory of primitive monotheism (*Urmonotheismus*). After the Scottish scholar Andrew Lang (1844–1912) had drawn attention to these gods, the Austrian scholar Wilhelm Schmidt (1868–1954) based on their existence in primitive culture and beliefs the theory that the oldest religion of mankind had been monotheistic and that polytheism as well as magic were later degenerations in the course of the history of a pure primeval religion. This theory, defended with great skill and an enormous mass of ethnological material by Schmidt and his collaborators, has long since been proved unsound and was abandoned even by his own students. The connection postulated between the high gods and monotheism has in most respects obscured rather than illuminated the situation. It is true that in many cultures the particular high god is considered as the creator, the founder of the order of the world, and also in some cultures as the reigning god according to whose will everything now happens, but such a god is rarely considered to be the one and only god that counts. Exclusive monotheism is not to be found in either primitive or archaic religions, according to present knowledge. The high god, however, can become a god of exclusive monotheism when circumstances are favourable, at least if he belongs to the active type of high god and not to the intellectual type, which serves mainly as an idea to answer the questions concerning the ultimate origin of things. (See the distinction above between ethical and intellectual monotheism.) This transformation probably occurred in the case of the Islāmic god Allāh. It seems to be more common, however, even for the active type of high god gradually to disappear behind a host of other, often minor, deities who are more concerned with the daily affairs of mankind.

The Deism of the 17th and 18th centuries is often compared to the conception of high gods as *dei otiosi,* "inactive gods," who have created the world and put it into order but after their work was done retreated from the world and left it to run in accordance with the order installed at the creation (see DEISM). Not all high gods, however, are inactive.

MONOTHEISM IN THE WORLD RELIGIONS

Classical monotheism. *Religion of Israel and Judaism.* There may be some reason to speak of the Old Testament conception of God as monolatry rather than as monotheism, because the existence of other gods is seldom explicitly denied and many times even acknowledged. The passionate importance given to the proclamation of Yahweh as the one god who counts for Israel and the equally passionate rejection of other gods, however, make it truer to speak of the monotheism of Israel; as in what became the Judaic affirmation of faith, "Hear, O Israel, the Lord is our God, one Lord" (from New English Bible) (Deut. 6:4). The eminent Dutch Old Testament scholar Theodorus C. Vriezen writes: "It is striking how the whole life of the people is seen as dominated by Yahweh and by Yahweh alone. Even if one cannot speak of a strictly maintained monotheistic way of thinking, it is yet clear that faith in Yahweh is the foundation of life for the Israelite." Monotheism is not a matter of mathematics—of opting for the number one as against other numbers—but the conscious choice of a person or group committing himself or themselves to one god rather than to any other ones and putting their faith in that one god; Joshua proclaims: "But as for me and my house, we will serve the Lord" (Josh. 24:15). In Israel the ethical aspect was as important as the exclusiveness of their one god; the prophets stressed the ethical elements of an essentially exclusive God. The God of Israel was a jealous god who forbade his believers to worship other gods. In this respect he differed from other gods in the ancient Near Eastern religions who, as a rule, did not put such exclusive obligation on their adherents.

Yahweh, the God of Israel

In later times—beginning in the 6th century BC and continuing into the early centuries of the Christian Era—Judaic monotheism developed in the same direction as did Christianity and also later Islām under the influence of Greek philosophy and became monotheistic in the strict sense of the word, affirming the one God for all men everywhere.

Christianity. Among the three great monotheistic religions, Christianity has a place apart, because of the trinitarian creed of this religion in its classic forms, in contradistinction to the unitarian creed of Judaism and Islām. The Christian Bible, including the New Testament, has no trinitarian statements or speculations concerning a trinitary deity, only triadic liturgical formulas invoking God the Father, Son, and Holy Spirit. It is true that Christianity also has had its Unitarians, such as the 16th-century Italian theologian Faustus Socinus, but this religion in its three classic forms of Roman Catholicism, Eastern Orthodoxy, and Protestantism acknowledges one God in three Persons: God the Father, God the Son, and God the Holy Spirit. According to Christian theology, this acknowledgment is not a recognition of three gods but that these three persons are essentially one, or as the dogmatic formulation, coined by the early Church Father Tertullian (*c.* 160–after 220), has it: three Persons and one substance. This conception was not accepted without contradiction as is proved by theological disputes of the 3rd and 4th century. It is evident that trinitarian speculation greatly resembles the way of thinking of pluriform monotheism. It is, of course, unlikely that there are any historical connections between these phenomena; both, however, try to solve what is more or less the same problem in more or less the same manner. The main distinction is that Christianity as a monotheistic religion restricts itself to three Persons, though primitive religions have no reason to restrict the number of possible forms of the one divine substance. Like other religions that cover a large territory and have a long history, Christianity appears in a multitude of variations: there is Christian pantheism, Deism, and even, paradoxically, Christian atheism, as exemplified in the mid-20th-century Death of God theologies.

Trinitarianism

Islām. No religion has interpreted monotheism in a more consequential and literal way than Islām. According to Islāmic doctrine the Christian dogma of a trinitarian god is a form of tritheism—of a three-god belief. There is no issue upon which this religion is so intransigent as the one of monotheism. The profession of faith, the first of the so-called Five Pillars of Islām (the basic requirements for the faithful Muslim), states clearly and unambiguously that "there is no God but Allāh," and in accordance with this principle the religion knows no greater sin than *shirk* ("partnership"), the attribution of partners to Allāh; that is to say, polytheism, or anything that may look like it—*e.g.*, the notion of a divine trinity. The Qur'ān declares: "Say: He, Allāh, is one. Allāh, the eternal. Neither has he begotten, nor is he begotten. And no one is his equal" (112). This profession of faith in Allāh as the one god is encountered in a more popular form, for example, in the stories of *The Thousand and One Nights*: "There is no god except Allāh alone,

he has no companions, to him belongs the power and he is to be praised, he gives life and death and he is mighty over all things." In only one respect has the uncompromising monotheism of Islām shown itself to be vulnerable; *i.e.,* in the doctrine of the Qur'ān as uncreated and coeval with Allāh himself.

Monotheistic elements in ancient Near Eastern and Mediterranean religions. *Egyptian religion.* Egyptian religion is of special interest with regard to the various topics treated in this article, for in it are found polytheism, henotheism, pluriform monotheism, trinitary speculations, and even a kind of monotheism. Especially in the time of the New Kingdom (16th–11th century BC) and later, there arose theological speculations about many gods and the one god, involving concepts that belong to the realm of pluriform monotheism. These ideas are especially interesting when related to trinitarian conceptions, as they sometimes are. In a New Kingdom hymn to Amon are the words: "Three are all gods: Amon, Re and Ptah . . . he who hides himself for them [mankind] as Amon, he is Re to be seen, his body is Ptah." As Amon he is the "hidden god" (*deus absconditus*); in Re, the god of the sun, he becomes visible; as Ptah, one of the gods of the earth, he is immanent in this world.

Monotheistic reform of Akhenaton

Much attention has been given to the reform of Egyptian religion as effected by the pharaoh Akhenaton (Amenophis IV) in the 14th century BC. This reform has been judged in many ways, favourably and unfavourably; it is, however, clear that Akhenaton's theology, if not fully monotheistic, in any case strongly tends toward monotheism. It is even possible to follow the gradual development of his ideas in this direction. At first he only singled out Aton, one of the forms of the sun god, for particular worship, but gradually this kind of henotheism developed in the direction of exclusive monotheism and even took on the intolerance peculiar to this religious concept. The names of the other gods were to be deleted. This un-Egyptian intolerance was probably the main reason for the speedy decline of this creed.

Babylonian religion. As far as is known, monotheism was largely absent from Babylonian religion. There henotheism seems to have been very important, since a person could choose one god for particular worship as if he were the only god.

Greco-Roman religions. The classic religions of Greece and Rome were in the main purely polytheistic, but in later times tendencies arose, partly stimulated by philosophy and later also by Judaism and Christianity, toward inclusive monotheism. The hymn to Zeus by the Stoic philosopher Cleanthes (*c.* 330–*c.* 230 BC) is the best known document of this process. It praises Zeus as the essence of divinity in all gods, creator and ruler of the cosmos, omnipotent, the giver of every gift, and the father of mankind. In the mystery religions of the Greco-Roman world and in the religious philosophies of later antiquity, such as Neoplatonism, Neopythagoreanism, and others, inclusive monotheism was more or less the rule.

Monotheistic elements in Indian and Chinese religions. The religions of India and China show an astonishing multiplicity of form, but exclusive monotheism, unless imported or stimulated by foreign influences, seems to be absent. All other phenomena treated in this survey of monotheism, however, are to be found in their religions. Inclusive monotheism and pantheism fit very well with the Indian notions of religion, particularly in Hinduism, as is witnessed by the reflections on Brahman, the self of the world, and *Ātman*, the self of individual man. As the *Upaniṣads* say: "Truly, in the beginning existed this Brahman, that only knew itself, saying: I am Brahman." Although in many cases one god, such as Śiva (Shiva) or Viṣṇu (Vishnu), receives nearly all the attention of the faithful, this emphasis never leads to a negation of other gods as such. Jainism does not differ in this respect. Only the religion of the Sikhs, heavily influenced by Islām, can be said to teach a kind of exclusive monotheism.

Buddhism teaches in essence that there are no gods in the full sense of the word. Gods are higher beings, but they belong to the cosmos and are as much in need of salvation as man is.

The ancient Chinese religion of the heaven (Shang Ti T'ien) is comparable to the religions that proclaim a high god as creator and ruler of the world, but monotheism has not resulted from this conception. Neither has Taoism led to monotheism.

BIBLIOGRAPHY. General works include: JOSIAH ROYCE, "Monotheism," in the *Encyclopaedia of Religion and Ethics,* vol. 8 (1928); "Monotheismus und Polytheismus," in *Religion in Geschichte und Gegenwart,* vol. 4 (1913); T.P. VAN BAAREN, *Doolhof der goden* (1960); G. VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation,* 2 vol., 1963); and R. OTTO, *Das Heilige* (1917; Eng. trans., *The Idea of the Holy,* 2nd ed., 1950). For primitive religions, see T. P. VAN BAAREN, *Menschen wie wir* (1964); and P. RADIN, *Monotheism Among Primitive Peoples* (1954); on the high gods, see H. ZWICKER, *Das Höchste Wesen* (1970); and E.O. JAMES, *The Concept of Deity* (1950), the only general systematic study of recent date treating the problems of monotheism. H.R. NIEBUHR, *Radical Monotheism and Western Culture* (1960), is a modern, incisive Protestant theological presentation of absolute monotheism.

(T.P.v.B.)

Monotremata

The Monotremata (monotremes) are a distinctive order of primitive mammals, the only surviving members of the subclass Prototheria.

Monotremes lay eggs and are reptilelike in many other ways but possess such essentially mammalian characters as mammary glands, hair, a large brain, and a complete diaphragm.

GENERAL FEATURES

Only two monotreme types, the platypus (*Ornithorhynchus anatinus,* family Ornithorhynchidae) and the echidnas, or spiny anteaters (*Tachyglossus aculeatus* and *Zaglossus bruijni,* family Tachyglossidae), are known. The earliest monotreme fossils known come from the Australian Pleistocene (2,000,000 or more years ago), and they are essentially the same as the living forms, which range in body length from about 30 to 80 centimetres (12 to 32 inches) and weigh from about one to ten kilograms (two to 22 pounds).

The platypus, found in eastern Australia from north Queensland to Tasmania, is one of the most remarkable of all mammals. It has a ducklike "bill," webbed feet, and a flattened beaver-like tail. When the first stuffed specimens reached England around the end of the 18th century, they were thought to be fakes, made by sticking together bits of different animals.

The echidnas also rank high among the world's most interesting mammals and are readily kept in captivity for physiological studies. Superficially, they are quite unlike the platypus. They are recognized by their sharp-pointed spines and tubelike noses. *Tachyglossus* is found throughout Australia, including Tasmania, and part of New Guinea. *Zaglossus* has an extensive distribution in New Guinea.

Neither the platypus nor the echidnas are readily seen by the casual visitor, but they may be fairly common in some areas. Platypuses are not kept in zoos outside Australia.

The monotremes do not appear to be carriers of diseases harmful to man. Both monotreme types are now wholly protected by law in Australia.

NATURAL HISTORY

Ecology. The platypus is amphibious, with a habitat ranging from alpine waters at an altitude of 1,500 metres (about 4,900 feet) in southeastern Australia to the sluggish subtropical Queensland coastal streams. It digs winding burrows with side branches in the banks of the streams and lakes that it inhabits. These burrows are usually five to ten metres (15 to 30 feet) long but may be as long as 30 metres (100 feet). The entrance is normally above the water level, however during floods it may be submerged.

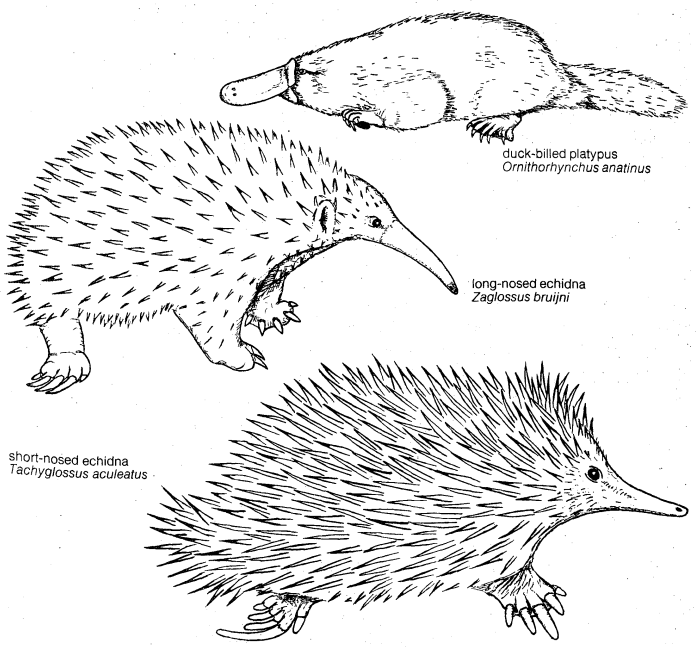


Figure 1: The three extant monotreme species.
Drawing by A.G. Lyne

Enemies of the platypus

The platypus has few natural enemies. It was formerly hunted extensively for its valuable velvet-like fur but for many years has been completely protected by law, and as a result it has increased in numbers. It appears to be relatively free of competition from other species in the water, where it feeds mainly on insect larvae, small crayfish, tadpoles, and other aquatic animals. Some authorities have suggested that the introduced rabbit (*Oryctolagus cuniculus*), which burrows in the banks of streams where the platypus once dwelt in quiet seclusion, is responsible for the disappearance of the platypus from many settled regions.

The echidnas are terrestrial animals, living in forests, scrubland, and deserts. Although usually encountered only by accident, *Tachyglossus* is one of the most widely distributed mammals in Australia. In New Guinea the habitats of *Zaglossus* include the humid forests that are almost continually blanketed by cloud cover. The animal has a known altitudinal range in these forests from about 1,100 to about 2,900 metres (3,600 to 9,500 feet).

Tachyglossus lives largely on ants and termites. The one element common in all the habitats occupied by this echidna is the presence of ants; termites are present in only some of the localities occupied.

Unfortunately, little is known of many aspects of the ecology of the monotremes, such as population density and social behaviour.

Life cycle and reproduction. The platypus breeds once annually, from July to August in the northern part of its range and to October in the southern part. Mating is believed to occur in the water, and it has been suggested that the spurs on the hind legs of the male may be used to hold the female. No spur grip was noted in an observation of copulation recorded by Australian zoologist D.H. Fleay in the course of an extensive study. An unusual courtship precedes mating, in which, among other manoeuvres, the male grabs the female's tail and the pair swim in circles.

In the breeding season, the female platypus digs a special burrow that twists and turns and may be very long. She then builds a nest of grass, leaves, and other plant material in a chamber at the end of the breeding burrow. This special burrow is blocked with earth in one or more places by the female during the period when she is laying and incubating the eggs. Only the left ovary is functional, and the eggs are always found in the left uterus. One to three, usually two, leathery-shelled eggs are laid. When there are two or three eggs they adhere together, but the sticky fluid that covers the eggs when they are laid soon dries and loses its stickiness. The eggs vary from 16 to 18 millimetres (about 0.60 to 0.70 inches) in length and from 14 to 15 millimetres in breadth.

The platypus has been bred in captivity on one occasion only, by Fleay in 1943. His observations suggest that the period from mating to egg laying is probably 12 to 14 days and the incubation period an additional ten to 12 days. The young are not more than about 18 millimetres long at hatching and do not leave the breeding burrow until they are about 17 weeks old. At this age they are fully haired and about 34 centimetres (13 inches) long. Because the female platypus has no teats, the milk is apparently sucked from the skin as it exudes from the ducts of the mammary glands. The life-span of the platypus is said to be ten to 15 or more years; at least one captive animal has lived more than ten years.

More research has been done on the echidna *Tachyglossus* than on the platypus, mainly because it is less difficult to maintain in captivity. In southeastern Australia, echidnas breed once a year from late June to late September. The period between copulation and egg laying is at least 16 days and may be as long as 28 days. Both ovaries are functional.

Egg laying in the echidna

In the breeding season the female echidna develops an abdominal pouch. Egg laying has not been observed, but there is circumstantial evidence that the egg is laid directly into the pouch by the female's curving her body during this process. One egg (rarely two or three) is laid at a time; it is creamy yellow and the shell has a rubbery texture. The eggs of echidnas are slightly smaller than those of the platypus, being from 14 to 17 millimetres long and from 13 to 15 millimetres wide. Within the pouch the egg is not stuck to the hairs but is quite loose and from time to time is located in different positions, apparently retained by the apposition of the lips of the pouch. The incubation period of the egg in the pouch is ten to 10.5 days, and the newly hatched echidna is about 15 millimetres (0.6 inch) long and weighs about 0.38 gram (about 0.01 ounce). After hatching, the young lives in the pouch and imbibes milk by sucking two special areas of the skin—the mammary areolae—where the ducts from the mammary glands open to the exterior on either side of the midline of the pouch. As in the platypus, no teats are present.

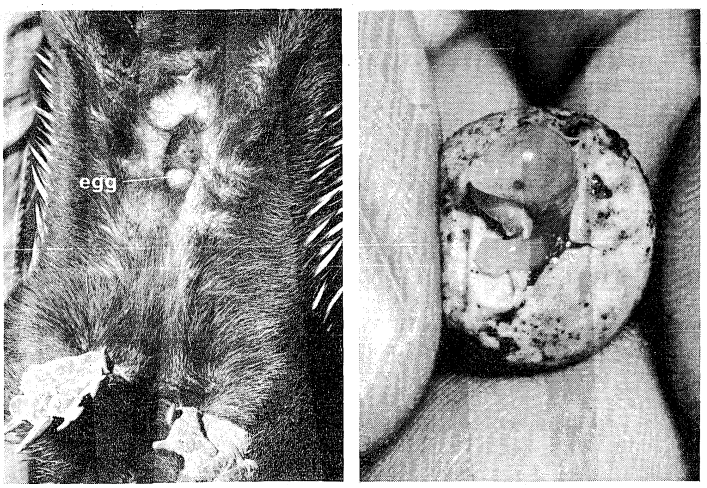
Authorities disagree on whether the young echidna takes milk by sucking or by licking. After detailed observations, an Australian zoologist, Mervyn Griffiths, stated that advanced pouch young take milk by sucking and not by licking it off the hairs in the mammary region, but he considered it possible that newly hatched and early stages of pouch young get their milk by licking it up as it is ejected onto the surface of the skin.

Information on how long the young stay in the pouch and what happens to them when they are cast out is meagre. One of the echidnas studied by Griffiths increased in body weight from 240 grams (about 0.5 pound) to 850 grams (about 1.9 pounds) in 43 days, a mean daily increase in weight of about 14 grams per day. The young was cast out of the pouch when it reached a body weight of 400 grams (about 0.9 pound). Thereafter, the mother's pouch regressed quickly, but she continued to suckle the young for some time. At the time when the young was dropped from the pouch, its eyes were open and the sharp spines projected from the skin. In wild echidnas it has been reported that the mother deposits the young in a burrow when its spines emerge and it is too large to be carried in the pouch.

Tachyglossus has been bred in captivity on only three occasions, once in Berlin and twice in Basel in Switzerland. In captivity *Tachyglossus* has lived up to 50 years, and *Zaglossus* up to about 30 years.

Behaviour and locomotion. Both the platypus and echidnas exhibit extreme specializations for feeding and locomotion.

The platypus emerges from its burrow for feeding usually in the early morning or late afternoon, but is sometimes in the water at other times of the day. When it is burrowing or walking on land the web of the front feet, which extends beyond the claws, is folded under the



(Left) Echidna (*Tachyglossus aculeatus*) held head downward to show egg in pouch. (Right) Echidna hatching.
M.E. Griffiths

Swimming by the platypus

palms. The muzzle and the claws of the front feet are used in burrowing. The platypus swims gracefully and expertly, using the front feet. The hind feet, in conjunction with the flattened tail, are used mainly for stabilization. The tail also assists the platypus in diving. The animal often swims along the surface with only the upper part of the muzzle and a small part of the head and body above the water. When submerged, the eyes and ears are closed by a fold of skin. The sensitive muzzle is an important organ for guiding the platypus while it swims blind, often close to the bottom. It may stir up the mud or move stones to locate its food. It surfaces often to breathe and to chew the prey, which it stores in cheek pouches while it is submerged. It can rest near the bottom for a few minutes by wedging itself under stones or logs. In captivity, the platypus has an enormous appetite, devouring each day the equivalent of half or more of its own body weight.

Echidnas are generally solitary. They appear to be more uncommon than they are because of the inaccessible nature of the places they choose to hide in and because of their remarkable ability to dig vertically into the ground and cover themselves with dirt when disturbed.

Observations on *Tachyglossus* in various parts of Australia indicate that its activity is closely related to the ambient temperature. In north Queensland, they are active primarily at night; but in the more temperate parts of eastern Australia, they are active both day and night. In desert regions they retire into crevices in rock outcrops during the day, where the humidity is higher and they are protected from intense sunlight.

Although echidnas are well adapted for digging, they do not excavate burrows, and they dig primarily to obtain food and escape enemies. They also retreat into hollow logs or under rocks or roots and wedge themselves into crevices and even into slight depressions in such a way that it is practically impossible to dislodge them. They can withdraw their appendages and erect their spines, as does a hedgehog. Echidnas walk with the legs fully extended so that the undersurface of the body is relatively high off the ground and with the hind toes directed outward and backward. They can run and climb well. *Tachyglossus* is extremely powerful and can tear apart rotten logs in search of termites or dig down to ant nests in rocky terrain.

FORM AND FUNCTION

General characteristics. One of the most striking features of the monotremes is that they lay eggs from which the young are subsequently hatched. These eggs are like those of reptiles in that they have large yolks and rubbery shells. Other interesting reptilian features are seen in the structure of the urogenital organs and in the skeleton. In both sexes of the monotremes, as in all reptiles, the intestine, bladder, and reproductive organs all open into a common chamber, the cloaca, with only one external opening. The penis of monotremes is used solely for delivering sperm and not also for urination as in most other mammals. In the skeleton, the shoulder girdle not only retains well-developed coracoid bones, which are reduced to vestiges in other mammals, but also an interclavicle. The pelvic girdle carries epipubic bones, which, as in some reptiles and marsupials, act as supporting structures for the ventral (abdominal) body wall.

Similarities to reptiles

The skull of monotremes has a smooth, rounded cranial portion, terminating in a long rostrum, or snout. Both monotreme types are toothless as adults, and the echidnas are toothless throughout life. The platypus has teeth when young, but they are shed before the animal becomes adult and are replaced by horny pads. The testes are abdominal and there are no teats and no vibrissae (whiskers). The limbs are modified for digging or swimming. The posture of the platypus resembles that of reptiles, especially lizards. In the echidnas, however, the limbs support the body well off the ground, even when the animals are stationary or walking slowly.

An interesting feature of the digestive system of monotremes is that the whole inner surface of the stomach is lined with a cornified (horny) epithelium. There are no glands of any kind. Particularly interesting is the lack of those that produce hydrochloric acid and peptic enzymes, which together initiate the digestion of protein in other mammals. Some other insectivorous mammals also have cornified epithelium covering most of the stomach lining, but there is a small glandular area. It has been suggested that the breakdown of food in the stomach of *Tachyglossus* is assisted by the grinding action of ingested dirt.

Platypus. The platypus has a flattened body covered with dense, short, fine hairs and coarse guard hairs, usually dark brown on the back and much paler on the belly. There is practically no neck. The tail is broad and flat like that of a beaver. It is covered with coarse hair, which is usually worn off on the under surface. The adult male is 50 to 60 centimetres (20 to 24 inches) in total length and the adult female is 40 to 50 centimetres (16 to 20 inches). An average-sized male weighs nearly two kilograms (4.5 pounds). One of the most conspicuous features of the platypus is its elongated and flattened muzzle, which bears a superficial resemblance to a duck's bill. This muzzle is covered with a soft, rubbery, naked, and sensitive skin, which is a blue-gray colour on its upper surface and pale pinkish or mottled below. Dorsal (toward the back) and ventral flaps, or shields, of naked skin extend backward from the muzzle over the adjacent hair. The nostrils are located on the upper surface of the muzzle near its anterior end. The eyes are small and there are no external ears. Both the eyes and ears open into a facial furrow that is closed when the platypus is below the surface of the water. There are cheek pouches in which food is stored until it can be chewed. The limbs are short and the feet webbed. The webbing on the hind feet reaches to the base of the claws, and on the front feet beyond the claws, making extra large paddles for swimming.

Venom apparatus of the platypus

The male platypus has a sharp, movable, horny poison spur on the inner side of each hind limb near the heel. Each spur is about 15 millimetres (0.6 inch) long and connected by a duct to a venom-secreting gland, called the crural, or poison, gland, situated on the dorsal aspect of the upper part of the hind limb. The function of the venom apparatus is unknown. An Australian zoologist, J.H. Calaby, stated that "If, as is generally believed, the crural gland is actively secreting only during and near the breeding season it seems very likely that the chief use of the venom apparatus would be combat between males for territory or females." The poison is not fatal to man but causes intense pain.

Echidnas. Superficially, the echidnas are quite unlike the platypus. They have a rounded body covered on the back and sides with stout, sharp-pointed spines and some hairs. The spines of *Tachyglossus* may measure up to six centimetres (2.5 inches) in length, those of *Zaglossus* up to about three centimetres. The undersurface is covered with coarse hair and is usually without spines. The colour of the spines and hair ranges from light brown to almost

black, depending on species and individual variation. Adult males are larger than adult females. There is a long, naked, and sensitive muzzle, a very short tail, and no neck. In some echidnas a definite external ear is present, but in others it is inconspicuous. The nostrils open near the tip of the muzzle. The mouth is small, only sufficient for the protrusion of the long, sticky tongue. The limbs are short and powerful and have strong claws for digging. In *Tachyglossus* one of the claws of the hind foot is particularly elongated and is used for scratching the skin between the spines.

The echidnas resemble the platypus in that there is a venom apparatus on each hind leg, but the spur and crural gland are smaller. In *Tachyglossus*, the spur (about five to ten millimetres long), which is situated on the inside of the ankle, is present in all males and some females. No observations on the use of the venom apparatus by the echidnas have been published.

Temperature regulation and hibernation. Monotremes are better able to control their body temperature than are reptiles. Except during hibernation, or torpidity of some species, the eutherian, or placental, mammals maintain a high and virtually constant body temperature usually between 36° and 40° C (97° and 104° F). The body temperature of most reptiles varies with that of the surroundings. The body temperature of the monotremes, however, does not quite reach that of the eutherian mammals, being about 31° to 32° C (88° to 90° F). The monotremes are usually described as physiologically primitive in regard to temperature regulation. It has been shown, however, that *Tachyglossus* is an excellent temperature regulator in some situations but less capable in others. It is probably no poorer in this respect than many eutherian mammals (*e.g.*, anteaters, armadillos, and sloths). In cold surroundings the echidna maintains a well-regulated body temperature. Studies of temperature regulation and torpor in *Tachyglossus* have shown that this echidna appears to be a true mammalian hibernator, with the ability to arouse from torpor without the aid of heat from external sources. This ability is lost after repeated periods of torpor in the laboratory at 5° C (41° F). The echidna resembles other hibernators in that torpor is not continuous but is interrupted by frequent arousals. Continuous periods of torpor lasting 9.5 days, with intervening periods of wakefulness ranging from 30 hours to 11 days, have been recorded for the echidna.

The platypus has periods of hibernation or semihibernation, at least in southeastern Australia. In captive animals, Fleay found that the hibernation periods were short and irregular during the cooler months, the longest single period being 6.5 days.

Laboratory studies have shown that the platypus is slightly superior to *Tachyglossus* in temperature regulation. The platypus has sweat glands on the muzzle and over the body surface, whereas *Tachyglossus* has them only in the pouch area. The heat tolerance of *Zaglossus*, which has many well-developed sweat glands distributed over the body, has not been studied. In its natural environment it is unlikely that the platypus will be exposed to high heat loads. The highly insulating coat of the platypus would protect it against heat loss in water and allow it to exploit cold environments.

EVOLUTION, PALEONTOLOGY, AND CLASSIFICATION

The monotremes are without living relatives outside Australia and New Guinea, and, apart from the presence of fossil monotremes in the Australian deposits of the relatively modern Pleistocene Epoch (the last 2,500,000 years or so), little is known of their ancestry. The fossils so far discovered differ from the existing forms only at the species level. It is interesting to note, however, that large echidnas of the genus *Zaglossus*, which now live only in New Guinea, were common and widespread in Australia during the Pleistocene.

Although the existent monotremes are highly specialized in many of their features, their possession of such typically mammalian characters as mammary glands, hair, and a large brain carries the suggestion that such characters are primitive features of early mammals.

Origin of the order

Most authorities regard the order Monotremata as the sole group of the subclass Prototheria and believe that it originated from a line of mammal-like reptiles different from that which gave rise to the other mammals. It has been suggested, because of the similarities between marsupials and monotremes, that they should be grouped together in a subclass of the Mammalia, the Marsupionta, but the idea has not gained wide acceptance. It does not take into account the many similarities between marsupials and eutherian mammals. It has also been postulated that a group of early prototherian mammals, of which the monotremes are specialized derivatives, gave rise to the marsupials and the eutherian mammals, but this view is not accepted by most biologists.

The ducklike muzzle, or "bill," alone distinguishes the platypus from all other mammals. Although platypuses have been divided into several subspecies, based mainly on differences in size, all authorities agree that the family Ornithorhynchidae has only a single genus and species, *Ornithorhynchus anatinus*.

Two genera of echidnas are recognized in the family Tachyglossidae; in *Tachyglossus* (short-nosed echidnas) the elongated muzzle is straight and not greatly prolonged, and in *Zaglossus* (long-nosed echidnas) the elongated muzzle is greatly prolonged and curved downward.

Although Griffiths recognized *Tachyglossus aculeatus* as the only species of the genus, he distinguished six subspecies, some of which he considered of doubtful validity. Some authorities consider that the relatively hairy Tasmanian echidna is a distinct species (*T. setosus*). Some mammalogists who have studied *Zaglossus* in detail have concluded that only one variable species, *Z. bruijni*, should be recognized. It can be seen that much more work on the biology of the monotremes needs to be done to clarify the place these extraordinary animals occupy in nature.

BIBLIOGRAPHY. The following are popular works on monotremes: H. BURRELL, *The Platypus* (1927); D. FLEAY, *We Breed the Platypus* (1944), and "Observations on the Breeding of the Platypus in Captivity," *Victorian Nat.*, 61:8–14 (1944); M. GRIFFITHS, *Echidnas* (1968). Works of a more technical nature are: M.L. AUGEE, E.H.M. EALEY, and H. SPENCER, "Biotelemetric Studies of Temperature Regulation and Torpor in the Echidna, *Tachyglossus aculeatus*," *J. Mammal.*, 51:561–570 (1970); J.H. CALABY, "The Platypus (*Ornithorhynchus anatinus*) and Its Venomous Characteristics," in W. BUCHERL, E.E. BUCKLEY, and V. DEULOFEU (eds.), *Venomous Animals and Their Venoms*, 1:15–29 (1968); W.K. GREGORY, "The Monotremes and the Palimpsest Theory," *Bull. Am. Mus. Nat. Hist.*, 88:1–52 (1947); K.W. ROBINSON, "Heat Tolerances of Australian Monotremes and Marsupials," *Aust. J. Biol. Sci.*, 7:348–360 (1954); K. SCHMIDT-NIELSEN, T.J. DAWSON, and E.C. CRAWFORD, JR., "Temperature Regulation in the Echidna (*Tachyglossus aculeatus*)," *J. Cell. Comp. Physiol.*, 67:63–72 (1966); H.M. VAN DEUSEN and G.G. GEORGE, "Results of the Archbold Expeditions, no. 90, Notes on Echidnas (Mammalia, Tachyglossidae) of New Guinea," *Am. Mus. Novit.* no. 2383 (1969).

(A.G.Ly.)

Monroe, James

As the fifth president of the United States, James Monroe made a significant and lasting contribution to his country's foreign policy in the Monroe Doctrine—a warning to the governments of Europe not to intervene in the affairs of the American hemisphere. He was born on Monroe's Creek, in Westmoreland County, Virginia, on April 28, 1758. His father was of Scottish, and his mother of Welsh descent. At the age of 16 he entered the College of William and Mary, but in 1776 left to fight in the U.S. War of Independence. In 1780 he began the study of law under Thomas Jefferson, then governor of Virginia, and between the two men there developed an intimacy and a sympathy that had a powerful influence upon Monroe's later career.

He was elected to the Virginia House of Delegates in 1782, and though only 24 years of age was chosen a member of the governor's council. From 1783 to 1786 he served in the Congress under the Articles of Confederation, the first constitution of the new nation. During his term he vigorously insisted on the right of the United

Monroe, portrait by E.O. Sully, 1836.
By courtesy of the Independence National Historical Park, Philadelphia

States to navigate the Mississippi River, then controlled by the Spanish, and attempted, in 1785, to secure for the weak Congress the power to regulate commerce and thus remove one of the great defects in the existing central government. In 1786 he married Elizabeth Kortright of New York by whom he had two daughters. Retiring from Congress, he began practicing law at Fredericksburg, Virginia, and was chosen a member of the Virginia House of Delegates in 1787, and (1788) a member of the state convention at which Virginia ratified the new federal Constitution. In 1790 he was elected to the U.S. Senate, where he vigorously opposed Washington's administration; but in 1794 Washington, nevertheless, nominated him as minister to France.

Diplomatic career. It was the hope of the administration that Monroe's well-known French sympathies would secure for him a favourable reception and that his appointment would also conciliate France's friends in the United States. His warm reception in France and his enthusiastic republicanism displeased the Federalists (the party of Alexander Hamilton which encouraged close ties to England) at home; he did nothing, moreover, to reconcile the French people to the Jay Treaty, which regulated commerce and navigation between the United States and Great Britain during the French Revolutionary Wars. Without real justification, the Franch regarded the treaty, as a violation of the French-American treaty of commerce and amity of 1778, and as a possible *casus belli.* Monroe led the French government to believe that the Jay Treaty would never be ratified by the United States, that the administration of George Washington would be overthrown as a result of the obnoxious treaty, and that better things might be expected after the election of 1796 when a new president, perhaps Thomas Jefferson, should take office. Washington, though he did not know of this intrigue, sensed that Monroe was unable to represent his government properly, and, late in 1796, recalled him.

Minister to France

Monroe returned to America in the spring of 1797, and in the following December published a defense of his course in a pamphlet of 500 pages entitled *A View of the Conduct of the Executive, in the Foreign Affairs of the United States.* Washington seems never to have forgiven Monroe for this stratagem, though Monroe's opinion of Washington and Jay underwent a change in his later years. In 1799 Monroe was chosen governor of Virginia and was twice re-elected, serving until 1802.

At this time there was much uneasiness in the United States when Spain restored Louisiana to France by the Treaty of San Ildefonso in October 1800. The Spanish district administrator's subsequent withdrawal of the United States' "right of deposit" at New Orleans greatly increased this feeling and led to much talk of war. Resolved upon peaceful measures, President Jefferson in January 1803 appointed Monroe envoy extraordinary and minister plenipotentiary to France to aid Robert R. Livingston, the resident minister, in purchasing the territory at the mouth of the Mississippi—including the island of New Orleans—authorizing him at the same time to cooperate with Charles Pinckney, the minister at Madrid, in securing from Spain the cession of East and West Florida. On April 18 Monroe was further commissioned as the regular minister to Great Britain.

Louisiana Purchase

Monroe joined Livingston in Paris on April 12, after the latter's negotiations were well under way; and the two ministers, on finding Napoleon willing to dispose of the entire province of Louisiana, decided to exceed their instructions and effect its purchase. Accordingly, on May 2 (antedated to April 30) they signed a treaty and two conventions whereby France sold Louisiana to the United States. The fact that Monroe signed the treaty along with Livingston did not hurt his political career at home, but he is not entitled to much credit for the diplomatic achievement.

In July 1803 Monroe left Paris and entered upon his duties in London; and in the autumn of 1804 he proceeded to Madrid to assist Pinckney in his efforts to define the Louisiana boundaries and acquire the Floridas. After negotiating until May 1805 without success, Monroe returned to London and resumed his negotiations concerning the impressment of American seamen and the seizure of American vessels. As the British ministry was reluctant to discuss these vexing questions, little progress was made, and in May 1806 Jefferson ordered William Pinkney of Maryland to assist Monroe. The result of the deliberations was a treaty signed on December 31, 1806, which contained no provision against impressments and provided no indemnity for the seizure of goods and vessels. Accompanying its signature was a British reservation maintaining freedom of action to retaliate against imminent French maritime decrees. In passing over these matters Monroe and Pinkney had disregarded their instructions, and Jefferson was so displeased with the treaty that he returned it to England for revision.

Monroe returned to the United States in December 1807 and was elected to the Virginia House of Delegates in the spring of 1810. In the following winter he was again chosen governor, serving from January to November 1811, and resigning to become secretary of state under James Madison, a position he held until March 1817. The direction of foreign affairs in the troubled period immediately preceding and during the War of 1812, with Great Britain thus fell upon him. On September 27, 1814, after the capture of Washington, D.C., by the British, he was appointed secretary of war, and discharged the duties of this office, in addition to those of the Department of State, until March 1815.

Presidency. In 1816 Monroe was elected president of the United States. In 1820 he was re-elected, receiving all the electoral votes but one. The chief events of his calm and prosperous administration, which has been called the "era of good feeling," were the Seminole War (1817–18); the acquisition of the Floridas from Spain (1819–21); the Missouri Compromise (1820), by which the first conflict over slavery under the Constitution was peacefully adjusted; recognition of the new Latin-American states, former Spanish colonies, in Central and South America (1822); and—most intimately connected with Monroe's name—the enunciation, in the presidential message of December 2, 1823, of the Monroe Doctrine, which has profoundly influenced the foreign policy of the United States.

Monroe Doctrine

Not until 1852 did the phrase Monroe Doctrine come into use. The "principles of President Monroe," as the message was referred to in Congress, consisted of three openly proclaimed dicta: no further European colonization in the New World, abstention of the United States from the political affairs of Europe, and nonintervention of Europe in the governments of the American hemisphere. In the diplomatic correspondence preceding the proclamation of these principles in the president's message was a fourth dictum not publicly associated with the Doctrine until 1869: the United States opposed the transfer of any existing European colonies from one European sovereign to another.

It is generally concluded that Secretary of State John Quincy Adams was undisputed author of the noncolonization principle of the Doctrine; the principle of abstention from European wars and politics was common to all the fathers of American independence, inherited and expressed by the younger Adams all his professional life; in Cabinet meetings, Adams also urged the dictum of nonintervention in the affairs of the nations of the Western Hemisphere. But Adams had no idea of proclaiming these dicta to the world. Monroe took responsibility for embodying them in a presidential message that he drafted himself. Modern historical judgment considers the Monroe Doctrine to be appropriately named.

On the expiration of his second term Monroe retired to his home at Oak Hill, Virginia. In 1826 he became a regent of the University of Virginia, and in 1829 was a member of the convention called to amend the state constitution. Having neglected his private affairs and incurred large expenditures during his missions to Europe, he was compelled to ask Congress to reimburse him for his expenses in the public service. In 1826 Congress finally authorized the payment to him of $30,000, and after his death appropriated a small amount for the purchase of his papers from his heirs. He died in New York City on July 4, 1831, while visiting his daughter. In 1858, the centennial year of his birth, his remains were reinterred with impressive ceremonies at Richmond, Virginia.

Thomas Jefferson, James Madison, John Quincy Adams, and many other prominent statesmen of Monroe's time all spoke loudly in his praise; but he suffers by comparison with the greater men of his time. Possessing none of their brilliance, he had, nevertheless, to use the words of John Quincy Adams, "a mind . . . sound in its ultimate judgments, and firm in its final conclusions."

BIBLIOGRAPHY

General works: WILLIAM P. CRESSON, *James Monroe* (1946), an extended study, well documented and clear; ARTHUR STYRON, *The Last of the Cocked Hats: James Monroe and the Virginia Dynasty* (1945), an approach that emphasizes the social and cultural background; STUART GERRY BROWN and DONALD G. BAKER (eds.), *The Autobiography of James Monroe* (1959), two long fragments of Monroe's unfinished autobiography printed in their entirety; HENRY ADAMS, *History of the United States,* 9 vol. (1889–91), informative about many aspects of Monroe's presidency.

Specialized studies: BEVERLY W. BOND, *The Monroe Mission to France, 1794–1796* (1907), concise and scholarly; LUCIUS WILMERDING, *James Monroe, Public Claimant* (1960), a brief study of Monroe's attempts after his retirement from the presidency to establish pecuniary claims against the United States and defend his management of public money; GEORGE DANGERFIELD, *The Era of Good Feelings* (1952), a large part is devoted to Monroe's administration; SAMUEL BEMIS (ed.), *American Secretaries of State and Their Diplomacy,* vol. 3 (1927), contains a section on Monroe's career as secretary of state by JULIUS W. PRATT, and vol. 4 (1928) also treats Monroe in a study by DEXTER PERKINS of the secretariat of John Quincy Adams.

(S.F.Be.)

Monsoons

Monsoons are seasonal winds that blow for approximately six months from the northeast and six months from the southwest, principally in Africa and southern Asia. Summer monsoons have a dominant westerly characteristic and a strong tendency to converge, rise, and produce rain. Winter monsoons have a dominant easterly characteristic and a strong tendency to diverge, subside, and cause drought. Both are the result of differences in annual temperature trends over land and sea. Changes are large over land, small over the oceans. The existence of these seasonal winds between East Africa, Arabia, and India was known to the ancient Greeks and Romans. Arab geographers of the Middle Ages gave accurate descriptions of them and spread the use of the term; an Arab pilot with a sound knowledge of the monsoons guided Vasco da Gama from East Africa to India in 1498.

The monsoon has been regarded variously as: (1) a seasonally reversing wind, northeast in winter, southwest in summer as part of a dominant wind; (2) a general circulation of semipermanent cells (wind systems that tend to recur in the same atmospheric location) that are maintained by the seasonal thermal contrasts of the continental and ocean surfaces; (3) an anomaly in the planetary air circulation; (4) an equilibrating flow between two regional pressure patterns; (5) a climatic event that takes place during warm, humid, rainy summers and dry, sunny winters. These characterizations are not necessarily mutually exclusive, but none of them is wholly correct or adequate by itself.

The delimitation of a monsoon region varies with each of the explanations above, and the general relevance of the explanations can best be understood by consideration of the principal monsoons and monsoon regions of the world. This article treats the causes, characteristics, and occurrence of monsoons with respect to monsoon regions (see CLIMATE; WINDS AND STORMS; CYCLONES AND ANTICYCLONES; HURRICANES AND TYPHOONS; THUNDERSTORMS; and PRECIPITATION).

THE INDIAN MONSOON

At the Equator, the area near India is unique in that dominant or frequent westerly winds occur at the surface almost constantly throughout the year; the surface easterlies reach only to 20° N in February, and even then they have a very strong northerly component. They soon retreat northward and drastic changes take place in the upper air circulation. This is a time of transition between the end of one monsoon and the beginning of the next. Late in March the high-sun season reaches the Equator and moves further north. With it go atmospheric instability, convectional (rising, turbulent) clouds, and rain. The westerly Subtropical Jet Stream still controls the flow of air across northern India and the surface winds are northeasterlies. As the high-sun season moves northward during April, India becomes particularly prone to rapid heating because the highlands to the north protect it from any incursions of cold air. In May the southwesterly monsoon is well established over Ceylon. There are three distinct regions of relative upper tropospheric (the troposphere is the lowest part of the atmosphere) warmth, namely (1) above the southern Bay of Bengal, (2) above the highlands of Tibet, and (3) across the still dry trunks of the peninsulas. The relatively warm area above the southern Bay of Bengal occurs mostly at the 500–1,000 millibar level (pressure at sea level is about 1,012 millibars, which corresponds to 14.7 pounds per square inch; 500–1,000 millibars occurs at 6,000–16,000 metres or 20,000–50,000 feet). It does not appear at a lower level and is probably caused by the release of condensation heat (associated with the change from water vapour to liquid water) at the top of towering cumulonimbus clouds along the advancing intertropical convergence.

Atmospheric instability and rapid heating

In May the dry surface of Tibet (above 4,000 metres) absorbs and radiates heat that is readily transmitted to the air immediately above. At about 6,000 metres an anticyclonic cell arises causing a strong easterly flow in the upper troposphere above northern India. The Subtropical (westerly) Jet Stream suddenly changes its course to the north of the anticyclonic ridge and the highlands, although occasionally it may reappear to the southward for very brief periods. This change of the upper tropospheric circulation above northern India from westerly jet to easterly flow coincides with a reversal of the vertical temperature and pressure gradients between 600 and 300 millibars. On many occasions the easterly aloft assumes jet force. It anticipates by a few days the "burst", or onset, of the surface southwesterly monsoon some 1,500 kilometres (900 miles) further south, with a definite sequential relationship, although the exact cause is not known. Because of India's inverted triangular shape, the land is heated progressively as the sun moves northward. This accelerated spread of heating, combined with the general direction of heat being transported by winds, results in a greater initial monsoonal activity over the Arabian Sea (at late spring time), where a real frontal situation often occurs, than over the Bay of Bengal. The relative humidity of coastal districts in the Indian region rises above 70 percent and some rain occurs. Above the heated land the air below 1,500 metres (5,000 feet) becomes un-

stable, but it is held down by the overriding easterly flow. This does not prevent frequent thunderstorms in late May.

During June the easterly jet becomes firmly established at 150 to 100 millibars. It reaches its greatest speed at its normal position to the south of the anticyclonic ridge, at about 15° N from China through India.

In Arabia, it decelerates and descends to the middle troposphere (3,000 metres or 10,000 feet). A stratospheric belt of very cold air, analogous to the one normally found above the intertropical convergence near the Equator, occurs above the anticyclonic ridge, across southern Asia at 30°–40° N and above the 6,000 metres (500 millibars) level. These upper air features that arise so far away from the Equator are associated with the surface monsoon and are absent when there is no monsoonal flow. The position of the easterly jet controls the location of monsoonal rains, which occur ahead and to the left of the strongest winds and behind them to the right. The surface flow, however, is a strong, southwesterly, humid, and unstable wind that brings humidities of more than 80 percent and heavy, squally showers that are the "burst" of the monsoon. The overall pattern of the advance follows a frontal alignment, but local episodes may differ considerably. The amount of rain is variable from year to year and place to place. Most spectacular clouds and rain occur against the Western Ghats, where the early monsoonal airstream piles up against the steep slopes, then recedes, and piles up again to a greater height. Each time it pushes thicker clouds upward until wind and clouds roll over the barrier and, after a few brief spells of absorption by the dry inland air, cascade toward the interior. The windward slopes receive from 2,000 to 5,000 millimetres of rain in the monsoon season: various factors, and especially topography, combine to make up a complex regional pattern. Oceanic air flowing toward India below 6,000 metres is deflected in accordance with the Coriolis effect (moving objects are deflected to the right over the earth's northern hemisphere). The converging, moist oncoming stream becomes unstable over the hot land and is subject to convectional turmoil. Towering cumulonimbus clouds rise thousands of metres, producing violent thunderstorms and releasing latent heat in the surrounding air. As a result, the upper tropospheric warm belt migrates northwestward from the ocean to the land. The main body of air above 9,000 metres (30,000 feet) maintains a strong easterly flow.

Precipitation patterns

Later, in June and July, the monsoon is strong and well-established to a height of 6,000 metres (less in the far north), with occasional thickening to 9,000 metres (Figure 1 top). Weather conditions are cloudy, warm, and moist all over India. Rainfall varies between 400 and 500 millimetres, but topography introduces some extraordinary differences. Cherrapunji, on the southern slopes of the Khāsi Hills at only 1,300 metres (4,300 feet), where the moist airstreams are lifted and overturned, has an average rainfall of 2,730 millimetres (110 inches) in July, with record totals of 897 millimetres (35 inches) in 24 hours in July 1915, over 9,000 millimetres (350 inches) in July 1861, and 16,305 millimetres (642 inches) in the monsoon season of 1899. Over the Ganges Valley the monsoon, deflected by the Himalayan barrier, becomes a southeasterly air flow. By then the upper tropospheric belt of warmth from condensation has moved above northern India, with an oblique bias. The lowest pressures prevail at the surface.

It is mainly in July and August that waves of low pressure appear in the body of monsoonal air. Fully developed depressions appear once or twice a month. They travel from east to west, more or less concurrently with high-level easterly waves and bursts of speed in the easterly jet, causing local strengthening of the low-level monsoonal flow. The rainfall consequently increases and is much more evenly distributed than it was in June (Figure 1 bottom). Some of the deeper depressions become tropical cyclones before they reach the land, and these bring torrential rains and disastrous floods. A totally different development arises when the easterly jet moves farther north than usual because the monsoonal wind rising over the southern slopes of the Himalayas brings heavy rains and local floods. The weather over the central and southern districts, however, becomes suddenly drier and remains so for as long as the abnormal shift lasts. The opposite shift is also possible, with midlatitude upper air flowing along the south face of the Himalayas and bringing drought to the northern districts. Such dry spells are known as "breaks" of the monsoon. Those affecting the south are similar to those experienced on the Guinea coast during extreme northward shifts of the wind belts (as later discussed), whereas those affecting the north are due to an interaction of the middle and low latitudes. The southwest monsoon over the lower Indus Plain is only 500 metres (1,600 feet) thick and does not hold enough moisture to bring rain. On the other hand, the upper tropospheric easterlies become stronger and constitute a true easterly jet stream. Western Pakistan, Iran, and Arabia remain dry (probably because of divergence in this jet) and thus become the new source of surface heat.

Associated dry spells

By August the intensity and duration of sunshine have decreased, temperatures begin to fall, and the surge of southwesterly air diminishes spasmodically almost to a standstill in the northwest. Cherrapunji still receives over 2,000 millimetres (80 inches) of rainfall at this time, however. In September dry, cool, northerly air begins to circle the west side of the highlands and spread over northwestern India. The easterly jet weakens and the upper tropospheric easterlies move much farther south. Because the moist southwesterlies at lower levels are much weaker and variable they are soon pushed back. The rainfall becomes extremely variable over most of the region, but showers are still frequent in the southeastern areas and over the Bay of Bengal.

By early October variable winds are very frequent everywhere. At the end of the month the entire Indian region is covered by northerly air and the "winter monsoon" takes shape. The surface flow is deflected by the Coriolis effect and becomes a northeasterly flow. This causes an October–December rainy season for the extreme southeast of the Deccan (including the Madras coast) and eastern Ceylon, which cannot be explained by topography alone because it extends well out over the sea. Tropical depressions and cyclones are important contributing factors.

Most of India thus begins a sunny, dry, and dusty season. The driest period comes in November in the Punjab; December in Central India, Bengal, and Assam; January in the northern Deccan; and February in the southern Deccan. Conversely, the western slopes of the Karakoram and Himalayas are then reached by the midlatitude frontal depressions that come from the Atlantic and the Mediterranean. The winter rains they receive, moderate as they are, place them clearly outside the monsoonal realm.

Because crops and water supplies depend entirely on monsoonal rains, it became imperative that quantitative, long-range weather forecasts be available. For a forecast to be released at the beginning of June, it is necessary to use, in April, South American pressure data and Indian upper wind conditions (positive correlation) and in May, rainfall in Rhodesia and Java and easterly winds above Calcutta (negative correlation).

THE MALAYSIAN-AUSTRALIAN MONSOON

Southeast Asia and northern Australia are combined in one monsoonal system which differs from others because of the peculiar and somewhat symmetrical distribution of land masses on both sides of the Equator. In this respect, the northwest monsoon of Australia is unique. The substantial water masses between Asia and Australia have a moderating effect on tropospheric temperatures, weakening the "summer monsoon." The many islands (Philippines, Indonesia, etc.) provide an infinite variety of topographic effects. Typhoons that develop within the monsoonal air bring additional complications.

Effects of water areas and topography

It would be possible to exclude North China, Korea, and Japan from the monsoonal domain because their seasonal rhythm follows the normal midlatitude pattern—a predominant outflow of cold continental air in winter, and frontal depressions and rain alternating with fine, dry

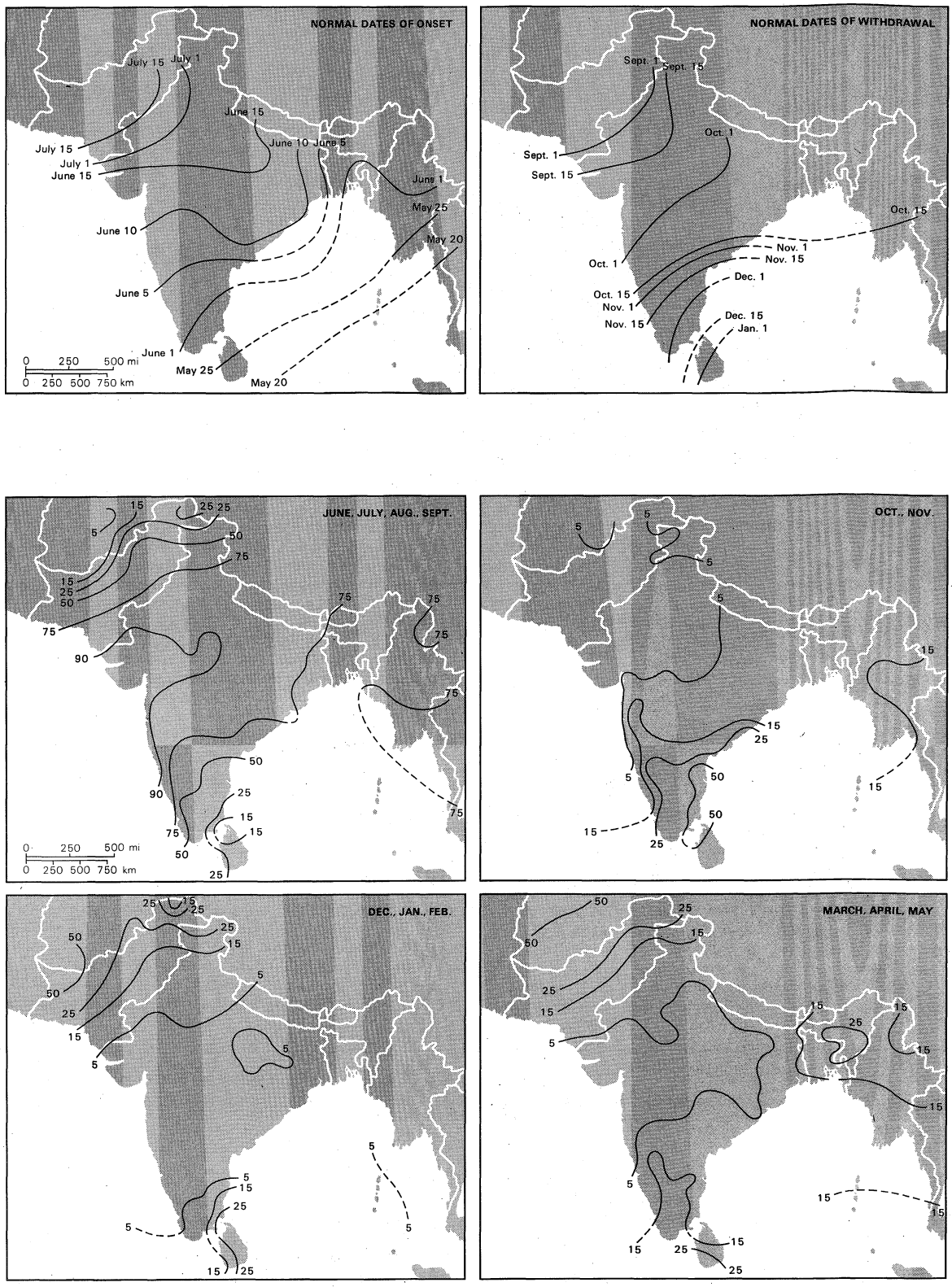


Figure 1: *The monsoonal system in the Indian subcontinent.*
(Top) Normal dates of onset and withdrawal of southwest monsoon. (Bottom) Seasonal distribution of rainfall as percentage of the annual.
From R. Ananthakrishnan and P. Rajagopalachari, India Meteorological Department; with permission from the Director General of Observatories, New Delhi, India

anticyclonic weather in the warm season. On the other hand, the seasonal reversal of wind direction in this area is almost as persistent as that in India. The winter winds are much stronger because of the relative proximity of the Siberian anticyclone. The tropical ridge of high pressure is the natural boundary between these nonmonsoonal areas and the monsoonal lands farther south.

The northern limit of the typical monsoon may be set at about latitude 25° N. Farther north, the summer monsoon is not strong enough to overcome the effect of the travelling anticyclones normally typical of the subtropics. As a result, monsoonal rains occur in June and also in late August and September, separated by a mild anticyclonic drought in July. In South China and the Philippines the trade winds (tropical easterlies) prevail in the October–April (winter) period, strengthened by the re-

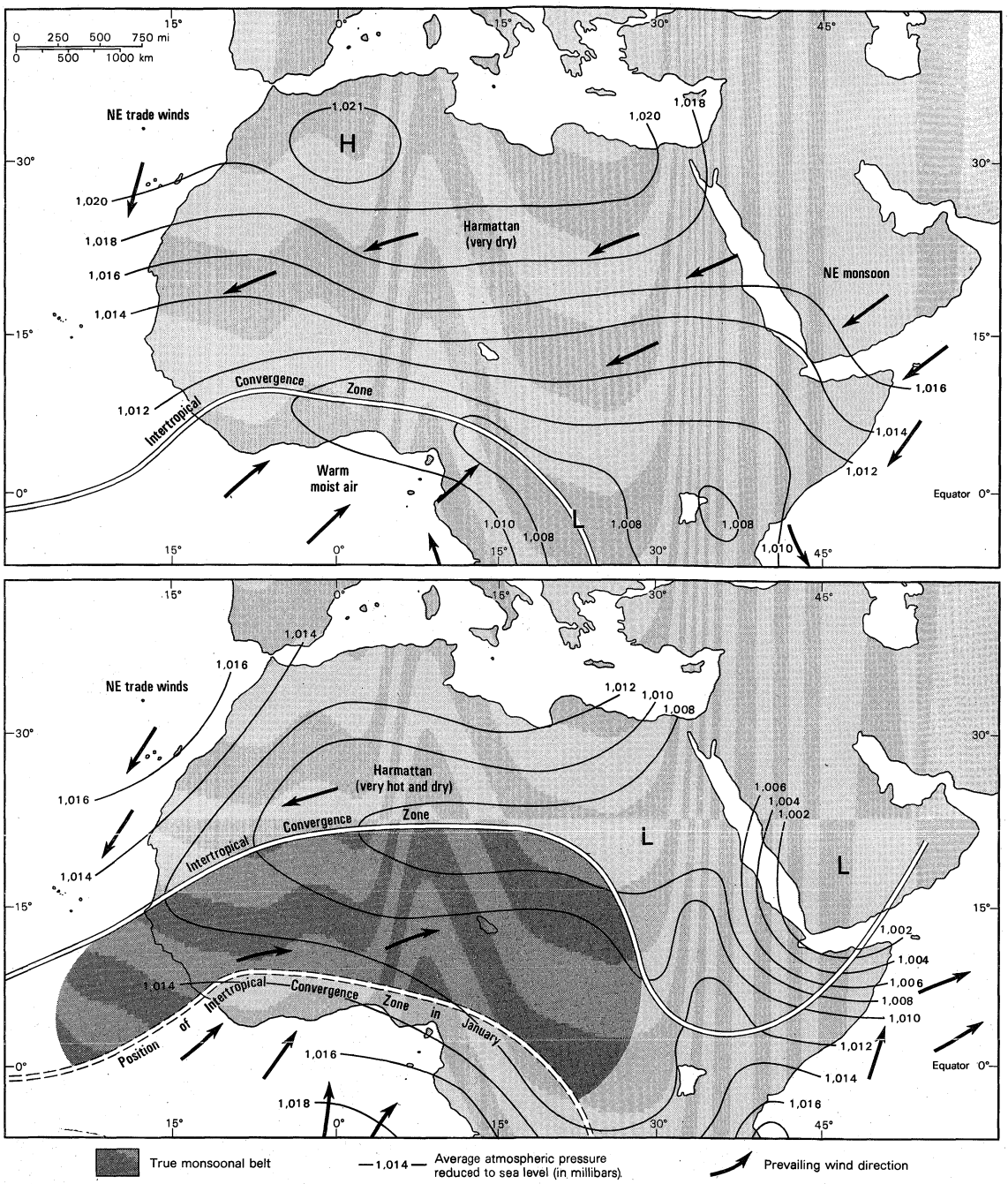


Figure 2: African monsoonal regimes in (top) January and (bottom) July.
From L. Welter and P. Suraud, *Monographie No. 8 de la Meteorologie Nationale* (November 1957)

gional, often gusty, outflow of air from the stationary Siberian anticyclone. Their disappearance and replacement by opposite (southwesterly) winds in the May–September (summer) period is the essence of the monsoon. In any case, these monsoonal streams are quite shallow, about 1,500 metres (5,000 feet) in winter and 2,000 metres (6,500 feet) in summer. They bring rain only when subject to considerable cooling, as happens to the windward anywhere on the steep slopes of the Philippines and Taiwan. On the larger islands there are contrasting effects, the slopes facing west getting most of the rain from May to October and a drought from December to April, whereas the slopes facing east receive orographic rains (those produced when moist air is forced to rise by topography) from September to April and mainly convectional rains from May to October.

In Vietnam and Thailand the summer monsoon is more strongly developed because of the wider expanses of overheated land. The southwesterly stream flows from May to October, reaching a thickness of 4–5 kilometres (2–3 miles); it brings a plentiful but not extraordinary rainfall. November to February is the cool, dry season, and March to April the hot, dry one when the highest temperatures prevail; in the far south the coolness is only relative. Along the east coast and on the eastward slopes more rain is brought by the winter monsoon. In the summer, somewhere between Thailand and Cambodia in the interior, there may be a faint line of convergence between the southwesterly Indian-Burmese monsoon and the southeasterly Malaysian monsoon.

Vietnam, Thailand, and Indonesia

Over Indonesia, because of the water expanses and the low latitude, monsoonal winds are very weak, but their seasonal reversal is very definite. From April to October the Australian southeasterly air flows, whereas north of the Equator it becomes a southwesterly. It generally maintains its dryness over the islands closer to Australia, but farther north it carries increasing amounts of moisture. The northeasterly flow from Asia, which becomes

northwesterly south of the Equator, is already laden with moisture when it reaches Indonesia, bringing cloudy and rainy weather between November and May. The wettest months are December in most of Sumatra and January elsewhere, but rainfall patterns are highly localized. In Java, for instance, at sea level alone there are two major climatic regions, an "equatorial" west with no dry season and a "monsoonal" east with extreme drought in August and September.

Because of its relatively small size and compact shape, Australia shows relatively simple monsoonal patterns. The north shore is subject to a clear-cut wind reversal between summer (November–April, northwesterly) and winter (May–September, southeasterly), but with two definite limitations: first, the northwesterly, rain-bearing monsoonal wind is often held offshore and is most likely to override the land to any depth during January and February; and secondly, even in summer there often are prolonged spells of southeasterly trade winds issuing from travelling anticyclones, separating the brief monsoonal incursions. The Australian summer monsoon is thus typical in direction and weather type but quite imperfect in frequency and persistence. Its thickness is usually less than 1,500 metres (5,000 feet) over the sea and 2,000–2,500 metres (6,500–8,000 feet) over the land.

Much less typical are the marginal monsoonal manifestations. On the northwest coast there frequently is a northwesterly air flow in the summer (December–March), as opposed to the winter southeasterlies, but this stream is very shallow and does not bring any rain; *i.e.*, its weather is not monsoonal although its direction is so. On the northeast coast the onshore air is humid and brings rain, but its direction is only partly modified in summer. It comes in mostly as a northeasterly, while at other times it is mostly southeasterly.

THE WEST AFRICAN MONSOON

The main characteristics of the West African seasons were known over two centuries ago. The southwest winter monsoon flows as a shallow (less than 2,000 metres) humid layer of surface air, overlain by the primary northeast trade wind, which blows from the Sahara and the Sahel as a deep stream of dry, often dusty air. As a surface northeasterly it is generally known as the harmattan, gusty, dry, and thirsty in the extreme—cool at night and scorchingly hot by day. As in a thorough monsoonal development, upper tropospheric anticyclones occur at about 20° N, while the easterly jet stream may occur at about 10° N, much closer to the Equator than they are in the Indian region.

Alternation with the harmattan

The West African monsoon is the alternation of the southwesterly wind and the harmattan at the surface. As seen in Figure 2, such alternation is normally found between latitudes 9° and 20° N. There are northeasterlies farther north all the time, but only southwesterlies farther south. The whole year is more or less dry at 20° N except for erratic rains in the high-sun season. The drought becomes shorter and less complete farther south. At 12° N it lasts about half the year, and at 8° N disappears completely. Farther south a different, lighter drought begins to appear in the high-sun months when the monsoonal southwesterly is strongest. It is due to the arrival of dry surface air issuing from anticyclones formed beyond the Equator, in the Southern Hemisphere, and is thus similar to the monsoonal drought in Java. Like the "break" of the monsoon in southern India, however, it occurs beyond the Equator.

The moist southwesterly stream, particularly frequent between 5° and 10° N, can reach much further north, bringing warm, humid nights and moderately hot but still humid days. The harmattan brings cooler nights, but the extreme daily heating causes a thermal range of 10° to 12° C (18° to 22° F). Even in the daytime, the harmattan may give a sensation of coolness to the human skin as it evaporates moisture from it. The alternation of the two winds is seasonal on the basis of overall frequency, but in fact it varies considerably with the synoptic pressure patterns. The harmattan comes in spells that mostly last from a few days to over a week.

The advancing fringe of the southwest monsoon is too shallow (under 1,000 metres or 3,000 feet) for many thunderstorms and disturbances. They usually occur 200–300 kilometres (125–200 miles) in arrears, where the moist air is deeper (1,000–2,000 metres), but the ground is still hot enough to make it very unstable. The tops of the cumulonimbus may reach 12,000 metres (40,000 feet), well above freezing level (4,200–4,500 metres or 14,000–15,000 feet). The disturbances are usually along a given longitude line that is slightly curved and may in fact form one long line squall. They also reach 12,000 metres or more, travelling steadily westward at 20 to 30 knots (25 to 35 miles per hour). This suggests that they originate in the primary trade wind aloft and as in India are probably related to the tropical easterly jet stream. The southwest monsoon dominates the weather, and clouds and rain abound; the rain is mostly due to coalescence of droplets, the clouds being mostly below 3,500 metres (11,500 feet). The humidity is very high and the daily range of temperature remains around 4° C (7° F).

Seasonal changes

If it were not for the change in wind direction when the southeast trades have crossed the Equator, the monsoon system of West Africa could not be distinguished from the weather system, caused by the seasonal shift in the latitude of the intertropical convergence, as experienced over most of Central Africa. There is a rainy season (in this case, the monsoon season), which on the west coast lasts two to three months at latitude 16° N, three to four months at 14° N, six to seven months at 10° N, eight months at 8° N.

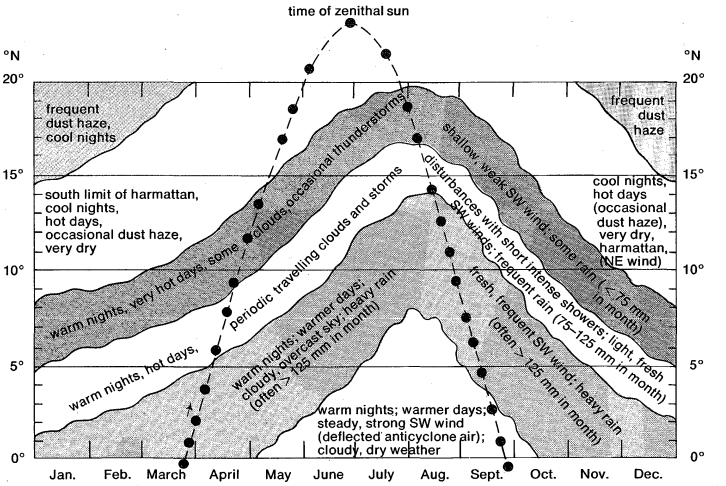


Figure 3: Distribution of climatic characteristics associated with monsoons in West Africa.

On the south coast, which is at latitude 4° N to 6° N, the southwest monsoon (as the intertropical convergence) may occur at any time, but the results are quite atypical for various reasons. In the low-sun season (December–February) the southwesterly is rare and ineffective, and the weather is cloudy but dry. From April to June the midday sun is at its highest and insolation (radiation received at the earth's surface) is most intense. Because the southwest wind is most frequent, the consequent building up of clouds leads to the main rainy season. In July and August (the short drought) cloudy conditions prevail, but the air issues direct from anticyclones further south and is dry, although its direction of flow does not change. Although cloudiness decreases after the second high-sun season in September and October, there is a period of occasional rains just sufficient to constitute a secondary maximum.

Toward the north, conditions are more distinctly monsoonal: by latitude 8° N the two wet seasons have merged into one long "wet" with two subdued peaks, which last approximately 7 to 8 months (March–October). The "dry," which is controlled by northeast winds, lasts from November to early March. There is one rainfall maximum (in August or September) only a short distance farther north although the wet season is only a few weeks shorter.

MONSOONAL TENDENCIES IN EUROPE AND NORTH AMERICA

There are monsoonal tendencies in central Europe, where the average wind direction in summer differs some 30° to 40° from that of the Atlantic, not as a continuous flow but intermittently within frontal depressions, bringing cool, cloudy weather, rain, and thunderstorms. Some see in this climatic pattern a true monsoon, but it is obvious that it is only an "embryo monsoon" that results in weather singularities. The latitude is too high for a true monsoon to arise.

In North America the relatively low latitude and the orientation of the land–sea boundary on the Gulf of Mexico are quite favourable to monsoonal developments. During the summer, low atmospheric pressure is frequent over the heated land; the northeasterly trade winds are consequently deflected to become easterly, southeasterly, or even southerly winds. Texas and the Gulf country generally may be completely overrun by a shallow sheet of oceanic air, which may continue for a long distance inland. The rainfall regime does not reveal any marked monsoonal pattern. There are mostly two, three, or even four minor peaks in the sequence of monthly rainfall totals. In the winter there often occur northers, which are offshore winds caused by the general anticyclonic flow of air from the cold land. Neither the summer onshore wind nor the winter offshore wind is persistent enough to constitute a monsoonal sequence, although monsoonal tendencies are quite evident.

Central American cycle

In Central America a true monsoonal cycle occurs over a small area facing the Pacific, between 5° and 12° N. Not only is there a complete seasonal reversal of the wind, but the rainfall regime is typically monsoonal. The winter period (November to January and March to April according to latitude and other factors) is very dry. The rainy season begins earlier (May) in the south and progressively later farther north, coming at the end of June in southern Mexico. It finishes at the end of September in the north and as late as early November in the south. The result is a rainy season that lasts three months in southern Mexico increasing in duration with decreasing latitude, to last six to seven months in Costa Rica. Latitude for latitude, this is a subdued replica of the monsoon of India.

BIBLIOGRAPHY. INDIA METEOROLOGICAL DEPARTMENT, *Symposium on Monsoons of the World, Delhi, 1958* (1960), is a comprehensive volume containing papers on the surface and upper air climatology of the monsoons, their dynamics, variability, rain and cloud systems, their place in the general circulation of the atmosphere, depressions and other disturbances in the monsoons, and on monsoon forecasting. WORLD METEOROLOGICAL ORGANIZATION WITH THE INTERNATIONAL UNION OF GEODESY AND GEOPHYSICS, *Proceedings of the Symposium on Tropical Meteorology, Rotorua, N.Z., 1963* (1964), ed. by J.W. HUTCHINGS, also contains several important papers on this subject, with special regard to India. Two fundamental papers on monsoons in China may be found in the ACADEMIA SINICA, *Collected Scientific Papers—Meteorology 1919–1949* (1954). P. PEDELABORDE, *Les Moussons* (1958; Eng. trans., *The Monsoons*, 1963), is a series of lectures that presents a clear exposition of earlier views. Most recent and comprehensive is C.S. RAMAGE, *Monsoon Meteorology* (1970).

(J.G.)

Montaigne, Michel de

Michel de Montaigne was a skeptical student of himself and of mankind, who stood for tolerance during the fanatical French religious civil wars of the later 16th century and spoke out for a man-centred morality. To express his views, but perhaps more to determine what they were for himself, he created the essay, which was a new literary form. He called the *Essais* (*Essays*), for which he is known, "a book consubstantial with its author, concerned with my own self, an integral part of my life."

Early career and works. Michel Eyquem was born in the Château de Montaigne, 30 miles east of Bordeaux, on February 28, 1533. His father, Pierre Eyquem, once mayor of Bordeaux, was a prosperous merchant whose grandfather had purchased the seigneury (a feudal territory) of Montaigne with its title of nobility; his mother was descended on her father's side from a family of Spanish-Jewish origin that had recently been converted to Catholicism. The oldest of eight surviving children, Montaigne had a brother and sister who became Protestants.

Montaigne, portrait by an unknown French artist, 16th century. In the Musée Condé, Chantilly, France.
Giraudon

He was brought up gently and was taught to speak only Latin until he was six years of age, when he began to learn French. After seven disappointing years at the Collège de Guyenne in Bordeaux, he may have studied law at Toulouse. From 1554 to 1557 he may have served in the Cour des Aides, a new tax court in Périgueux.

The court was abolished in 1557, and its members were incorporated into the Parlement of Bordeaux, one of eight regional bodies that composed the Parlement of France, the king's highest court of justice. Montaigne remained for 13 years in one of its lower chambers, working conscientiously but without enthusiasm. He was barred from rising higher because he had close relatives in the upper chambers.

Friendship with La Boétie

The one joy of these years was his friendship with another counsellor, Étienne de La Boétie, a bond that proved to be the strongest of his life. A man of letters slightly older than Montaigne, married, a linguist, poet, and respected magistrate, La Boétie was sure of his stoical Catholic views and goals, whereas Montaigne was still a talent in search of a function. Both men pronounced the friendship unique and mysterious; it fulfilled Montaigne's deepest need for complete communication. Indeed, had La Boétie lived, Montaigne might not have written his *Essays*. The friendship continued for four (or perhaps six) years, when La Boétie contracted dysentery and died; Montaigne never ceased to mourn him.

After seeking distraction for two years in various amours, in 1565 Montaigne married Françoise de La Chassaigne, whose Catholic family was prominent in the Parlement. Although fond of women, he accepted marriage unenthusiastically as a social duty. Françoise bore him six girls; all died in infancy except one, Léonor. Otherwise his marriage was not a close bond.

Translator of Raymond Sebond

Montaigne's first publication appeared in 1569: his translation into French, at his father's request, of the 1,000-page "Book of Creatures, or Natural Theology," in Latin, by the 15th-century Spaniard Raymond Sebond (Ramón Sabunde). It sought to demonstrate God's existence and nature and man's duty toward him by analogies between the human and divine drawn from the levels of the creation. Its prologue, full of extravagant claims, was put on the Roman Catholic Church's Index of Forbidden Books. Montaigne's version was never condemned. He made an excellent translation, reshaping Sebond's pedestrian prose and sharply cutting his claims in the prologue.

In June 1568, his father died, leaving Michel the lord of Montaigne. Two years later he sold his Parlement position (in the manner of the time), published most of La Boétie's works with his own dedicatory letters, and retired to his estate, intending to collect his ideas and write.

The "Essays": Books I and II. He spent the next nine years (1571–80) working chiefly on the first two books of *Essays*, but his retirement was not complete, for he continued to make frequent trips to court in Paris, and, in 1574, the Governor of Bordeaux sent him to confer with the army commander in the southwest over a disputed command in one of the religious wars of the time. In 1571 he received the Order of Saint Michael; by 1573 he was a gentleman of the chamber to the Catholic king Henry III and, in 1577, to King Henry of Navarre, who was the future King Henry IV of France and leader of the Protestants. Montaigne, a staunch but moderate Catholic loyalist, may have owed this last distinction to an unsuccessful attempt at mediation, made some time between 1572 and 1576, between Navarre and Duc Henri de Guise, who was to become the leader of the ultra-Catholic Holy League and who would violently oppose succession of the crown to a Protestant.

Content and themes of the *Essays*

The first two books of the *Essays* include 94 chapters; the earliest are short, relatively impersonal, and are strings of anecdotes with brief conclusions. The most interesting among them reveal the problems preoccupying Montaigne at that time: inconsistency, ambition, and, above all, pain and death. We must prepare for death, he writes, as "the goal of our career . . . the necessary object of our aim," and "learn to meet it steadfastly and to combat it." Similarly with pain: "We must resist it and tense ourselves against it." Still influenced by La Boétie, Montaigne now sees life as beset by ills, and he resolves to summon up reason and will to resist them.

The *Essays* of 1574–75 show a growing wariness of stoical solutions to life's problems. In "Of Practice," Montaigne draws on an experience—a dangerous collision on horseback—to help him come to terms with death, and he recognizes that premeditation may only magnify the problem. In the essay "Of Drunkenness," he finds extreme stoical heroism against pain incompatible with consistency and wisdom. He begins "A Custom of the Island of Cea" by rehearsing ancient arguments in favour of suicide but then turns about and attacks it as un-Christian, cowardly, and unnatural.

These chapters lead to Montaigne's attack on Stoicism, and indeed all dogmatism, in his longest chapter by far, the skeptical "Apology for Raymond Sebond," in which he draws on the skepticism of the Greek philosopher Pyrrhon of Elis. He admits the fallibility of the human mind to know anything with certainty and takes as his motto "What do I know?" The "Apology" may be divided into six unequal parts, of which the last five are a counterattack against Sebond's dogmatic critics. In the first part he discusses Sebond, his book, and the two main criticisms of it: that in religion reason cannot supplant faith and that Sebond's reasons for faith are easily overthrown. Second, he shows that animals are the virtual equals of man; third, he demonstrates by man's ignorance of God and of himself that man has no knowledge; fourth, he very briefly warns an unnamed princess, for whom the chapter is written, that what follows—abandoning one's own weapon (reason) to make one's adversary lose his—is indeed a last resort; fifth, he argues that man cannot have knowledge because he is a creature of flux and therefore can know neither flux nor being; and finally, he concludes that not Stoicism but God alone can raise men above humanity.

Content of the "Apology"

The "Apology for Raymond Sebond" is puzzling because it belies its title. It is the most apologetic of apologies; less than one-tenth of the work deals with Sebond at all, and his arguments are often refuted. Some have considered it a perfidious attack on Christianity; others have thought that Montaigne forgot Sebond even while defending him. A likelier view is that Montaigne wrote most of it by 1576 as a treatise on presumption and the vanity of reason and with little if any thought of Sebond. The princess addressed in the warning, recognizable as Marguerite de Valois, was in Gascony in 1578, among Protestants eager to convert her. Knowing Montaigne and having enjoyed Sebond, presumably in his translation, she apparently asked Montaigne to defend him; and Montaigne, never a real admirer of Sebond, probably reluctantly decided to enlarge his earlier piece into the "Apology."

Montaigne's attack on human knowledge leaves knowledge of self unscathed. The "Apology" repeatedly asks: "If we cannot know ourselves, what can we know?" When Montaigne, apropos of recognition of our limitations, wrote, "At least we must become wise at our own expense," he implied that self-knowledge is a possibility, the beginning of wisdom as well as of knowledge. It is in this way that the skepticism of the "Apology" leads him from dogmatic philosophy to the study of himself.

Self-study leads him directly and comfortably to self-portrayal: in style as well as content, the portrait must be faithful, changing as the subject changes. The essays for Montaigne were trials, probings, samplings of the self he is examining and even the power of judgment with which he examines it.

Mood of the later essays

The *Essays* of 1578–80 are optimistic, as Montaigne turned his attention from human limitations to human resources. Intellectually he had freed himself from tutelage in the "Apology"; the test of experience remained. This came in the summer of 1578 with the agonizing pain caused by the kidney stone. His father had suffered from the same illness, and Montaigne had long dreaded such an attack. Eighteen months later, assessing the results, he found that his fear had been worse than the pain and felt he had taken the measure of pain. Whereas he had formerly seen virtue only in struggle, his model now became the effortless virtue of Socrates.

Earlier Montaigne had written a largely negative critique of education. Now, in "Of the Education of Children," he showed what good training could do: not fill the memory but fashion the judgment and thus form a morally independent man who loves virtue and hates vice, not from fear but for themselves.

Travel. After the first two books of *Essays* were published (Bordeaux, 1580), Montaigne went north, presented a copy to Henry III, served in his army at the siege of La Fère, in Normandy, and set out with a brother, two friends, and servants on a 15-month trip (September 1580–November 1581) through France, Germany, Switzerland, and Austria to Italy. His main reasons were his health, heartsickness over the state of France, weariness of domesticity, eagerness to see Rome, and sheer love of travel.

He left a 300-page record of his trip, his *Travel Journal*, not intended for publication and published only in 1774. No masterpiece, it further reveals aspects of Montaigne's self-portrait: his consuming interest in all religions, delight in mechanical gadgetry, and outgoingness.

Mayor of Bordeaux. In September 1581, while at La Villa, in Italy, Montaigne learned he had been elected mayor of Bordeaux. Feeling ill and unfit, he returned to France expecting to decline; but Henry III politely ordered him to accept. His first term (1581–83) was uneventful; the second was not so. The city was loyal to the Catholic king, but intransigent Holy Leaguers held important posts within; Protestant forces were nearby, and a collision between them was always imminent. Montaigne stresses his detachment from his office, but his letters reveal his vigilant efficiency. He was outside the city, as he often had to be, when his term ended in July 1585, and did not return to transfer authority to his successor.

One peaceful year at home followed. Then in July 1586, a large League army besieged Castillon, five miles away. Montaigne was suspect to extremists on both sides. Plague broke out again and spread to Montaigne's estate, driving him and his family to seek asylum for six months before returning home to his *Essays*.

Role as mediator in war of the three Henrys

During this time he attended the conferences at Saint-Brice near Cognac, where Catherine de Medicis summoned him to try—in vain—to persuade Henry of Navarre to abjure his Protestantism and return to court. In October 1587, three days after Navarre's important victory over a royalist army at Coutras, Navarre paid his second overnight visit to Montaigne, who probably helped to persuade him not to expand his victory into rebellion. Three months later (February 1588) Montaigne went to Paris on a secret mission from Navarre to Henry III,

seeking to bring them closer together against the menacing League. Again suspect to both extreme factions, Montaigne was seized on his way and briefly detained by Protestants, and that summer he was imprisoned—again briefly—by Leaguers in the Bastille.

The "Essays": Book III. While in Paris in 1588, Montaigne supervised the publication of the fifth edition of the *Essays*, the first edition to contain the 13 substantial chapters of Book III. His sense of human solidarity had been fostered by the favourable reception of the first *Essays*, by his friendly relations with people of other nations, by being twice chosen to govern Bordeaux, and by his warm feeling for his peasant neighbours who had behaved heroically during the plague.

Almost from the start of Book III, Montaigne proclaims his new sense of unity—in the race and in the individual: "Each man bears the entire form of man's estate. . . . There is no one who, if he listens to himself, does not discover in himself a pattern all his own, a ruling pattern . . . " Now his self-portrait becomes also a portrait of man.

His heightened confidence leads to greater candour and greater concern with public behaviour. He studies himself more than any other subject, for, in so doing, one can learn best to control his conduct. Moreover, self-study is a school of human nature. He adds but does not correct; as an author he felt he must not suppress anything he has written. His ideas follow one another, but sometimes from afar; he loves the free, gambolling, poetic gait. His chapter titles are often allusive, not descriptive; he rejects artificial signposts and an inattentive reader. He is always ultimately a moralist, seeking to understand in order to judge. The highest human wisdom and happiness, he finds, would be to know our exact duty to others and to ourselves. Against the glamour of public action he upholds private life as more demanding, for in it man needs virtue, not its semblance. Private life seeks man's most important knowledge: how to live in conformity with his natural condition. Through self-examination and self-control, it leads to moral independence. Because others cannot know us as we ourselves can, we must learn to be our own judges: "I have my own laws and my court to judge me, and I address myself to them more than anywhere else."

Montaigne's portrait of man

Man, Montaigne now believes, would be happier and better if he accepted himself for what he is. It is a dangerous temptation to try to be more than human. Self-acceptance is no enemy of self-improvement but, indeed, a condition of it. He rarely repents, and he questions the validity of any overall repentance not sent by God. Self-portrayal commits him to practice what he preaches; his rule is to dare to say all that he dares to do. His aim is truth and harmony. Having learned from his illness to accept pain as interdependent with pleasure and to harmonize them, Montaigne rejoices in the legitimate joys of living: "I love life and cultivate it just as God has been pleased to grant it to us." The last pages of the *Essays* are a hymn of gratitude for life.

Last years. Relatively little is known about his last four years. Early in 1588 he met his "covenant daughter," an enthusiastic admirer who became his literary executrix, Marie de Gournay. He followed the King when he fled the revolt (instigated by Henri de Guise) from Paris to Chartres and Rouen. He attended the Estates-General in Blois, where Henry III had Henri de Guise killed. Around the end of 1588, Montaigne returned south and assisted in keeping Bordeaux loyal first to Henry III and then, after his assassination (1589), to Henry IV. Two letters of 1590 to his new king reveal the acknowledgment of an invitation to join the King and a regret that he could not do so because of illness, though he would have been a firm, frank counsellor. He continued to work on his *Essays*. The best account of his death reports that a quinsy left him speechless for his last few days. On September 13, 1592, at the age of 59, he died peacefully hearing mass in his room.

Montaigne had written no new books or chapters of his *Essays* since 1588 but had added about a thousand passages that reveal a further sharpening of his views and greater boldness in five main areas: self-revelation, obscenity, his plan to portray himself, the state of religion in France, and his independent morality. He has learned how self-portrayal forms a man, delineating his inner self more sharply. He writes indignantly against the cruelty and hypocrisy of the civil wars and blames the Catholics as well, though not as much, as the Protestants. In his boldest plea for moral autonomy, he criticizes any virtue that is not all our own, even that motivated by religion and law, as a slave morality.

Some of his earlier views are modified, and this growth of the essays by accretion is one of their most appealing features. His rejection of stoical rigour is sharper. The arbitrariness of the soul, its freedom to shape and colour the reports of the senses, was a proof of our ignorance in the "Apology"; now it proves our capacity for happiness: "Our good and our ill depend on ourselves alone."

Montaigne's view of death changes completely. Whereas he had once called it the goal of man's career, he now pronounces it the end—but not the goal—of life. No longer is pain always to be avoided: to eliminate it would eliminate pleasure "and in fine would annihilate man."

Assessment. In the four centuries since he wrote, views of Montaigne have changed much as he did himself. His contemporaries deplored his self-portrait and admired his stoical sententiae. The 17th century saw mainly the skeptic and *honnête homme*, a gentleman adhering to the rules of good manners. Jean-Jacques Rousseau and later Romantics were drawn to his self-portrait and his free, pre-Neoclassical style. Sainte-Beuve in the 19th century was struck by his natural, independent morality; this and the universality of his self-portrait have impressed the last century of readers.

Translated into English in 1603, Montaigne was promptly adopted by Anglo-Saxon admirers, including Francis Bacon, John Marston, John Webster, William Shakespeare, Sir Richard Burton, Lord Byron, William Makepeace Thackeray, Ralph Waldo Emerson, Robert Louis Stevenson, Aldous Huxley, and Virginia Woolf.

Influence on Anglo-Saxon writers

In the 20th century, he is fully recognized in all his aspects as a great writer, and his public is worldwide. Most of his readers see him as friend, mentor, and master of the essay, of the "art of being truthful," and of the art of living.

MAJOR WORKS

EDITIONS: *Essais de messire Michel Seigneur de Montaigne*, 2 vol. (1580); *Édition seconde, reveue et augmentée*, 2nd ed. (1582); *Cinquiesme édition, augmentée d'un troisième livre et de six cens additions aux deux premiers*, 5th ed., 3 vol. (1588); *Les Essais de Michel seigneur de Montaigne, édition nouvelle, trouvée après le déceds de l'autheur, reveüe & augmentée par luy d'un tiers plus qu'aux précédentes impressions*, ed. with a preface by Mlle de Jars de Gournay, 3 vol. (1595); *Les Essais de Michel de Montaigne, publiés d'après l'exemplaire de Bordeaux, avec les variantes manuscrites & les leçons des plus anciennes impressions, des notes, des notices et un lexique*, ed. by F. Strowski, F. Gébelin, and P. Villey, 5 vol. (1906–33). (TRANSLATIONS): *The Essays or Morall, Politike and Millitarie Discourses of Lo: Michaell de Montaigne . . . Now done into English by . . . John Florio* (1603); *The Essays of Michel de Montaigne*, trans. and ed. by J. Zeitlin, 3 vol. (1934–36); *Essays*, trans. with an introduction by J.M. Cohen, "Penguin Classics" (1958).

OTHER WORKS: *Journal du voyage* (1774). See also *Oeuvres complètes*, ed. by A. Armaingaud, 12 vol. (1924–41) and *The Complete Works of Montaigne*, trans. by D.M. Frame (1957).

BIBLIOGRAPHY

Biographies: DONALD M. FRAME, *Montaigne* (1965); PAUL BONNEFON, *Montaigne, l'homme et l'oeuvre* (1893), an excellent study but now somewhat outdated; ROGER TRINQUET, *La Jeunesse de Montaigne* (1972).

Critical works: HUGO FRIEDRICH, *Montaigne*, 2nd rev. ed. (1967), with emphasis on his intellectual background (in German); HERBERT LUTHY, "Montaigne, or the Art of Being Truthful," *Encounter*, 1:33–44 (1953, reprinted in QUENTIN ANDERSON and JOSEPH A. MAZZEO (eds.), *The Proper Study*, pp. 319–336, 1962); PIERRE VILLEY, *Les Sources et l'évolution des Essais de Montaigne*, 2nd ed. rev. and enl., 2 vol. (1933), establishes the chronology of the *Essays*.

(D.M.F.)

Montana

Although its name is derived from the Spanish *montaña* ("mountainous region"), Montana has an average elevation of only 3,400 feet, the lowest among the Rocky Mountain states. The mountains sweep down from the Canadian province of British Columbia into the western third of the state, into Idaho on Montana's western and southwestern border, and southward into Wyoming. The eastern two-thirds of the state, however, is a gently rolling landscape with far horizons and a big sky, with millions of grazing cattle and sheep, and with only scattered evidences of human habitation. It forms a part of the northern Great Plains, shared with the provinces of Alberta and Saskatchewan to the north, North and South Dakota to the east, and northeastern Wyoming.

Only three states—Alaska, Texas, and California—have an area larger than Montana's 147,138 square miles (381,086 square kilometres), which is, in fact, larger than that of Japan; and only three—Alaska, Wyoming, and Nevada—have a lower population density than Montana's 4.8 persons per square mile. The 694,409 inhabitants counted in the census of 1970 are relatively far from the markets for their cattle, grain, lumber, metals, and petroleum, as well as from the nation's manufacturing and supply centres. Nature is omnipresent, and the state is strongly oriented toward the out-of-doors, toward summer and winter sports, toward hunting and fishing, and toward the long-distance trip for socializing and entertainment or as a cure for the imaginary but chronic prairie- or mountain-born restlessness, cabin fever.

In spite of its northern location, the Montana of the 1970s is very much a Western state, with high-heeled cowboy boots more fashionable than shoes and cattle rustling not merely a memory from the past. The main street of the capital, Helena, is Last Chance Gulch, the city's original name and a reminder of the prospectors who invaded the hills in the 1860s to pan for gold. By 1889, when Montana became the 41st state of the Union, the cattle drive was an institution, and the state had begun to emerge as the copper-mining capital of the nation. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; GREAT PLAINS; and ROCKY MOUNTAINS.)

THE HISTORY OF MONTANA

Settlement. The first white men known to have set foot in Montana were the members of the Lewis and Clark Expedition in 1805–06, though French fur traders were in the area as early as the 1740s. Other trappers and traders followed, setting up forts to trade with the Indians. The only early trading post to become a permanent town was Fort Benton, which was established in 1846 and became an important port on the Missouri River and head of navigation for steamboats from St. Louis. Jesuit missionaries followed the fur traders. Father Pierre-Jean De-Smet, a Belgian, established St. Mary's Mission near present Stevensville in 1841, and Father Anthony Ravalli, an Italian, joined him in 1845.

Gold miners and vigilantes

Miners and adventurers flocked in after rich placer gold deposits were discovered at Bannack in 1862 and Alder Gulch in 1863. A secretly organized gang of holdup men and murderers, known as road agents and led by Sheriff Henry Plummer, was identified late in 1863. The hastily formed Committee of Vigilance destroyed the gang by hanging more than two dozen road agents in a few weeks in 1864. Montana Territory, formerly part of Idaho Territory, was established in that year, with Bannack its first capital and Virginia City, on Alder Gulch, its second. Farmers settled in the Bitterroot, Deer Lodge, and Gallatin valleys to raise grain and vegetables.

Large-scale cattle and sheep raising began with the first big herd of cattle being driven overland from Texas in 1866. The vast grasslands of central Montana were ideal for cattle, but in the severe winter of 1886–87, some owners lost 90 percent of their herds to freezing and starvation. Since that time, stock growers have raised hay for winterfeed.

As pressure from settlers increased, the native Indians fought to protect the hunting grounds that provided the necessities of life for their people. Sioux and Cheyenne won their last major victory in June 1876, at the Battle of the Little Bighorn, also known as the Custer Massacre. A band of Nez Percé under Chief Joseph won a battle in the Big Hole Basin the following year and fled toward Canada, only to be met and defeated by United States troops a few miles south of the "Medicine Line," the international boundary. Montana entered the Union in 1889, with Joseph K. Toole as the first governor.

Growth of the state. As gold mining declined in importance, world demand for copper increased, and the immense ore deposits at Butte, known as "the richest hill on earth," attracted outside capital. From 1895 to 1905 the so-called copper kings—William A. Clark, Marcus Daly, and Frederick Augustus Heinze—fought for control of the mines and of the state government, which became notoriously corrupt.

Economic booms and busts

After 1909, homesteaders poured into the plains country to claim farms of 320 acres. During the "county-busting era," from 1910 to 1925, the number of counties doubled from 28 to the present 56, most of the new ones lying in the grain-growing part of the state in the east and north. The expected population growth did not materialize, however, for, after a few years of bumper wheat crops and high prices, a cycle of drought years brought financial disaster. Hundreds of mortgages were foreclosed and more than 200 banks closed permanently between 1920 and 1926. Thousands of settlers sought livelihoods elsewhere; since then, farming methods suitable for the semi-arid prairie land have been developed.

Thus, Montana was already depressed when the nationwide depression began in 1929. Droughts continued, many mines closed, and markets for forest products were few. The federal government undertook several big irrigation and soil-conservation projects. Fort Peck Dam, a multipurpose project on the Missouri River, was completed in 1940. During World War II both mining and agriculture returned to their boom levels of past years.

Petroleum production, which had begun in a small way in 1915, was expanded in 1951 with major discoveries in the Williston Basin in the northeast. A very rich field at Bell Creek near Broadus was opened in 1967. In that year the state's longest and costliest strike took place against the Anaconda (Mining) Company, the state's largest private employer, lasting 250 days and costing about $34,000,000 in wages.

THE NATURAL AND HUMAN LANDSCAPE

The heavily forested western third of Montana includes the tumbled, rugged Rocky Mountains and the Continental Divide, contrasting with the generally treeless, gently rolling eastern portion, which is broken by buttes, tree-bordered streams, and small, isolated mountain ranges.

Rivers and lakes. The state is unusual in that its waters contribute to three major drainage systems. On the west side of the divide, streams flow into the Clark Fork and Kootenai rivers and then, through the Columbia River, into the Pacific Ocean. East of the divide, the Madison, Jefferson, and Gallatin rivers meet at Three Forks to form the mighty Missouri, which, with its major tributary, the Yellowstone, finally joins the Mississippi. From parts of Glacier National Park, in the northwest, streams drain northward into Hudson Bay by way of the Belly and St. Mary rivers. Cruise boats ply the upper Missouri, notably in the spectacular trip through the Gates of the Mountains, above Helena.

Richness of water resources

Most of the natural lakes are in the Rockies, and some of the most beautiful are in Glacier National Park. Flathead Lake, with a surface area of 197 square miles, is the largest in the state. A new lake was formed in 1959 when an earthquake toppled a mountain into the Madison River northwest of West Yellowstone, the resulting body of water becoming known as Quake Lake or Earthquake Lake. Some large man-made lakes are reservoirs backed up behind multipurpose government dams. Major dams are Fort Peck on the Missouri River near Glasgow; Canyon Ferry on the Missouri near Helena; Hungry Horse on the South Fork of the Flathead near Columbia Falls;

Yellowtail in the Bighorn River Canyon, creating Bighorn Lake, which extends into Wyoming; and Libby Dam on the Kootenai, producing Koocanusa Lake, part of which lies in British Columbia.

Climate. Montanans tend to be defensive about their climate, and they do not react gratefully to expressions of sympathy from New Yorkers, New Englanders, or North Dakotans. The state was called "the icebox of the nation" before Alaska achieved statehood. A temperature of −70° F (−57° C) was registered at Rogers Pass on January 20, 1954, but the cities of Duluth, Minneapolis, and St. Paul, in Minnesota, have lower average temperatures. In July, the warmest month, Montana's average is about 65° F (18° C); in January, the coldest, about 15° F (−9° C). Humidity is generally low, and the same aridity that requires dryland farmers to be ingenious about conserving moisture makes both summer and winter more comfortable than thermometer readings suggest.

Average precipitation is low—about 18 inches in the west and 13 inches in the east—resulting from the position of the Rocky Mountains and the state's interior location on the continent. The Rockies sometimes warm part of the east and north by creating a chinook wind, which melts snow rapidly and gives quick relief from cold. Destructive hailstorms sometimes occur in July and August, damaging unharvested grain.

Soils and vegetation. Montana's soils differ widely. Eastern, prairie soils are mature, underlaid by lime deposits, and productive when enough moisture is available. Many mountain soils are immature and thin because erosion has carried the richer decomposed materials into the valleys. The grasslands east of the Rockies make Montana one of the great range states. Only Texas outranks it in the number of cattle raised on native grasses; Texas and Wyoming exceed it in range-raised sheep. About one-fourth of the state is classified as forest land, and about one-third of this is also suitable for grazing. In general, where there are mountains, there are forests. Most of the trees, and all those cut for timber, are softwood conifers. There are no hardwood forests.

Animal life. Large game animals abound, mainly in the mountains. They include moose, elk, mule and white-tailed deer, grizzly and black bears, bighorn sheep, and mountain goats. Herds of pronghorn antelope run free on the prairie. A few bison, once abundant beyond counting, were saved from extinction, and a herd of between 300 and 400 is maintained on the National Bison Range, near Moiese. A herd belonging to the Crow Indians, which had to be destroyed because of undulant fever, is being rebuilt. There is some trapping of mink, beaver, and muskrat for their fur. Predators include mountain lions, bobcats, and coyotes.

Prevalence of wildlife

Game birds are abundant, including grouse, ducks, pheasant, partridge, and geese. About 700 trumpeter swans, once almost extinct, live in the area of Yellowstone, Hebgen, and Red Rock lakes, forming the largest colony of these birds in the United States. Kokanee and silver salmon, whitefish, and several kinds of trout predominate as attractions for recreational fishermen.

Human inroads. Air travellers can fly over Montana's prairie for miles without seeing a town, a road, or a railroad track. This is grazing country for cattle and sheep. In dryland-farming areas the fields are big; and where wind blows the soil, green swaths of growing grain alternate with brown swaths of fallow ground, on which last year's stubble catches and holds such rain and snow as fall. In the mountains, forests are dark on steep slopes and in deep canyons, with occasional scars of logging roads and patches of timber burned in forest fires.

The main urban settlements developed from mining camps; from trading centres for farmers, stockmen, and forest workers; and from early railroad division points. With improved roads and plentiful automobiles, rural communities have lost stores, banks, and other businesses. Billings, the largest city, boomed when it became the business centre for the state's oil industry. Hydroelectric power was responsible for the growth of Great Falls, the second largest city.

THE PEOPLE OF MONTANA

In the early years of the 20th century, many families from northern Europe settled on homesteads to farm on the eastern prairie, and immigrants from across Europe came to work in the mines. Their descendants have been well assimilated. Place-names such as Belgrade, Glasgow, Havre, Harlem, and Malta do not indicate the national origin of the settlers but were bestowed by railroad officials when the towns were established. If there is a social status in the state, it is conferred less by wealth than by descent from pioneers.

Ethnic minorities. Several Indian tribes, usually on the move in search of game, lived in Montana when the first explorers arrived. Indians now constitute the state's only large racial minority group, with about 27,000 of the about 31,000 nonwhites in 1970. About three-quarters of Montana's Indians live on seven reservations: Blackfeet, Crow, Northern Cheyenne, Rocky Boy's (Chippewa and Cree Indians), Fort Peck (Sioux and Assiniboin), Flathead (Flathead or Salish, Kutenai, and Pend d'Oreille), and Fort Belknap (Assiniboin and Atsina or Gros Ventre). Most of the reservations are poor in natural resources except for grazing land, much of it leased to non-Indian stock growers. The reservations are partially self-governing through tribal organizations, and the federal Bureau of Indian Affairs is the principal agency overseeing Indian activities. Employment opportunities on the reservations are few, and most Indians find it difficult to tear themselves away from traditionally strong family ties to work among strangers. Life is even harder for the landless Indians, who came to the state after the reservations were established and do not have a legal right to live on them. About half of the nearly 2,000 Negroes live in the Great Falls area.

Indian reservations

Religion. About half the people in Montana are affiliated with an organized religious group, with slightly more than half of them Roman Catholics. Through vigorous missionary activity the Mormon Church, centred in nearby Salt Lake City, Utah, has been increasing its membership for several years.

Demography. The population in 1970 represented an increase of only about 3 percent in ten years. Since births outnumbered deaths by almost 78,000 in the decade, it is obvious that many thousands of persons left the state. Only eight cities had more than 10,000 population in the census of 1970: Billings (61,581), Great Falls (60,091), Missoula (29,497), Butte (23,368), Helena (22,730), Bozeman (18,670), Havre (10,558), and Kalispell (10,526).

During the 1960s, population moved not only out of the state but from east to west within it, from farms and ranches to the more industrialized areas. Of the 56 counties, 41 lost population, all but one of them rural. Only four eastern counties showed a gain, and the only significant increase there was 15 percent in Powder River County, due to oil-drilling activity at the Bell Creek field. Twenty-one counties have less than 5,000 inhabitants, and two of these, Golden Valley and Petroleum, have less than 1,000. In the same decade the urban population increased to 53.4 percent from 50.2 percent of the total.

THE STATE'S ECONOMY

Montana's imbalance between dependent persons below 19 and over 65 years of age and those of working age tends to reduce per capita income and tax-paying capacity in comparison with the entire nation. Between 1948 and 1970, per capita personal income declined from more than 110 percent of the national average to only 86 percent. The state government has relatively little power to influence economic growth. Montana specializes in industries that, nationally, are growing slowly if at all. Some tax concessions are offered to new businesses. A 1972 popular vote to legalize gambling was not (1973) implemented by the legislature. Tourism is an important if only seasonal source of employment in the state. Trade unions are strongest and most active in communities where one or two major employers predominate: in Butte, Anaconda, and Great Falls.

Agriculture. The state's largest single source of income is agriculture, with livestock accounting for about two-thirds of cash income, and crops the remainder. Montana's national rank in livestock at the beginning of the 1970s was tenth in beef cows, third in sheep, and 28th in hogs. The state ranked second in the nation in barley and seventh in wheat production, 10th in sugar beets, fifth in flaxseed, and 13th in oats. Following a nationwide trend, the number of farms has been diminishing, while the average size of individual farms has increased for several decades.

Mining. Copper is the leading metal, representing nearly half of the state's metal production. Gold, the basis of pioneer settlement in the 1860s, is now only a by-product of copper mining, along with silver and lead. Among other mineral resources are phosphates, used in manufacturing fertilizers and elemental phosphorus; vermiculite; bentonite; sand and gravel; and gypsum. Montana has enormous low-sulfur coal reserves, estimated at more than 2,000,000,000 tons, that can be strip-mined at low cost. Although visible coal deposits in southern Montana were noted by a military expedition in 1876, these resources have scarcely been touched. Coal is now gaining importance as a fuel to produce electrical power. Legislation adopted in 1971 required mine operators to restore the surface of the mined land for other uses, such as grazing. Petroleum was discovered in commercial quantities at Elk Basin in 1915. The Elk Basin, Kevin-Sunburst, and Cut Bank fields led in production for several years. The great Williston Basin was developed in 1951, but more recently the Bell Creek field in Powder River County has been the most productive.

Exploitation of natural resources

Forestry. Lumbering and forest-products manufacture are vital to western Montana. Of the more than 17,000,000 acres of commercial forest land, more than three-quarters is owned by the federal and state governments. More than 9,000 persons are employed in the forestry industry, including the manufacture of plywood and of pulp and paper products. About half of Montana's forest products are sold to states in the Middle West.

Military presence. The beginning and the ending of big military installations have had profound effects in Montana. Malmstrom Air Force Base, built in 1942, is of continuing economic importance to Great Falls. The economic shock produced in the early 1970s by the construction of an anti-ballistic-missile installation in north central Montana was followed by an opposite upheaval when the project was later abandoned. Closing of the Glasgow Air Force Base in 1968, after 11 years, jolted the economy of that area, and closing of an air force radar base in 1971 lost for Lewistown the benefits of a significant annual payroll.

Transportation. Under Amtrak, administered by the quasi-governmental National Railroad Passenger Service Corporation, the Burlington Northern in 1971 continued passenger service across the northern part of the state but maintained it only provisionally to the six largest cities in the state farther to the south. In-city bus lines operate only in Butte, Anaconda, and Billings.

In 1970 the state had more than 1,200 miles of interstate highways. There were almost 6,000 miles of primary highways and the same amount of secondary roads. Several major airlines and air-taxi lines serve the state, and many small planes are privately owned.

ADMINISTRATION AND SOCIAL CONDITIONS

Constitution. The only copy of Montana's first proposed state constitution was mislaid on the way to the printer in St. Louis, Missouri, in 1866, and its contents must remain a subject of historical curiosity. A second attempt to achieve statehood failed in 1884; but Pres. Grover Cleveland signed an enabling act in 1889, and the third constitutional convention produced a constitution of 21 articles. As the state grew and its people's needs changed, so many amendments were added that in 1970 voters called for a constitutional convention. The new constitution was adopted by the electorate in 1972.

State government. Allegiance is so evenly divided between the Republican and Democratic parties that election patterns are not predictable. Ticket splitting is common. The governor and the lieutenant governor, formerly elected separately, now are elected on one ticket. The state has an open primary; voters need not declare party affiliation, and registration is permanent unless a voter misses a general election.

The political process

The executive branch was reorganized so that in 1973 some 160 state boards and agencies were consolidated into 19 cabinet-style departments, with virtually all appointments directly controlled by the governor. The voting age was lowered to 19 in 1969, and Montana was one of the first states to ratify the amendment to the United States Constitution permitting 18-year-olds to vote.

The two-house legislature, which under the state's new constitution meets every year, was composed, in early 1973, of 50 senators and 100 representatives.

Judges are elected without party designation. The highest court is the five-man Supreme Court. The state is divided into 18 judicial districts, in which 28 district judges serve. On the lowest level are justices of the peace, who serve organized townships, and police judges, whose authority is limited to cases involving municipal ordinances.

Income from taxation on personal property, the most important source of tax revenue, varies widely from county to county because of differences in valuation by county assessors. A major political issue in the legislature in the early 1970s revolved around the choice of increasing the personal income tax or imposing a sales tax. Personal income tax increased in 1971.

Local government. The county is the highest level of local government. Its powers and duties are defined and limited by state statutes. Three elected commissioners are the chief administrators, though a full-time manager may be employed instead. Government in low-population counties is especially expensive and burdensome. Municipal governments, like those of counties, have no inherent powers but derive all their authority from the state. They can, however, enact local ordinances, whereas counties cannot.

Law enforcement. Municipalities have police forces, whereas organized townships have elected constables whose major duty is to serve legal papers. Each county has an elected sheriff, who appoints his deputies and has jurisdiction outside towns and cities. Because cattle rustling is a continuing problem, some sheriffs and deputies act as brand inspectors to prevent the sale of stolen livestock.

Health, education, and welfare. Montanans pay various penalties for the wide-open spaces they enjoy. An increasing number of communities have no physician, dentist, or hospital. Although rural schools continue to consolidate and introduce bus transportation, many children attend small and not always adequate one- or two-teacher country schools. Beaverhead County, larger than the state of Connecticut but with only 8,000 inhabitants, has only two high schools, one with fewer than 100 pupils. School districts are corporate bodies headed by a county superintendent of schools and governed by elected school boards.

The state's system of higher education, chartered in 1893, includes universities at Missoula and Bozeman and four-year colleges at Billings, Butte, Havre, and Dillon. There are church-affiliated private colleges in Great Falls, Helena, and Billings; junior colleges in Miles City, Glendive, and Kalispell; and public postsecondary vocational–technical schools in Great Falls, Billings, Missoula, Helena, and Butte.

Montana's welfare program is state supervised and administered by county departments of public welfare. Because of sparse population and few private social-service agencies, costs are relatively high. Welfare departments are caught between pressures from the public to tighten up the programs and from organized low-income groups to provide wider services. Special problems exist on Indian reservations, where winter allowances are provided to help recipients of public assistance meet clothing and fuel costs in the cold months. Tuberculosis, once endemic among Indians, is now under control.

CULTURAL LIFE AND INSTITUTIONS

The arts. Most artistic activity centres in the cities with colleges and universities, several of which sponsor visits of lecturers and professional artists of various kinds in addition to presenting the work of faculty members and students. Six cities have symphony orchestras that include some professional musicians.

State cultivation of the arts

The Montana Institute of the Arts, founded in 1948, is a grass-roots organization that ties together the scattered, often isolated practitioners of various arts and crafts through publications, an annual festival, and traveling exhibits. The Montana Arts Council, a state agency affiliated with the National Endowment for the Arts, funds dozens of local cultural organizations up to 50 percent with federal money, primarily for music, drama, dance, literature, and the visual arts. Among other activities, it provides short poet-in-residence programs for children in the public schools of small communities. In addition to dramatic performances by students in various colleges, there are summer theatres in more than a half dozen communities; and the Montana Repertory Theater, based in Missoula, tours both inside and outside the state.

Folk culture. Several Indian tribes hold traditional dance ceremonies at which outsiders are welcome. Crow Indians and whites cooperate in an annual re-enactment of the Battle of the Little Bighorn on the Crow Indian Reservation. Rodeos abound, as do games on horseback known as *o-mok-sees* and square-dance groups. In Red Lodge an annual nine-day Festival of Nations, originated to ease tensions among several European ethnic groups of coal miners, has become a proud tradition.

Museums. The Montana Historical Society maintains a fine museum, art gallery, and specialized library in Helena. The Trigg-Russell Memorial Gallery in Great Falls, specializing in the works of the cowboy artist Charles Marion Russell, is the property of the city. Billings has the Yellowstone County Fine Arts Center, operated in a former city jail, and the Yellowstone County Museum. The excellent, small Museum of the Plains Indians is in Browning, the agency town for the Blackfeet tribe. Many communities nurture art galleries and small museums of local historical interest.

Communications and libraries. Montana has 14 daily newspapers and 77 weeklies, 52 AM radio stations, ten FM stations, and ten commercial television stations. Most communities depend on community-antenna television or on booster systems for reception from stations in Salt Lake City, Utah; Spokane, Washington; Williston, North Dakota; and Lethbridge, Alberta. Only about 40,000 of Montana's scattered people do not have access to public-library services. Larger cities and towns all have municipal or county libraries, and nine bookmobiles operated by multicounty federations serve other areas. The Montana State Library, in Helena, lends books to supplement the holdings of local libraries.

Prospects. While admitting that a small population and a scarcity of employment opportunities have some drawbacks, Montanans take great pride in their wide-open spaces. Many persons remain in the state, return to it, or move into it because of the lack of crowding and the recreational advantages—although they might well earn more money elsewhere. There are no traffic-choked streets or highways, and distances are usually described in terms of hours of driving time rather than in miles. There is every likelihood that the population will continue to grow at well below the national average. Montana is one of the few states in the nation about which it may be said that little change appears likely, but this is a condition that will continue to satisfy a large percentage of the residents.

BIBLIOGRAPHY. *The Montana Almanac 1959–60* (latest edition) is the best source of general information, though somewhat dated.

History: M.G. BURLINGAME and K. ROSS TOOLE, *A History of Montana*, 3 vol. (1957); N.P. LANGFORD, *Vigilante Days and Ways* (1958); T.J. DIMSDALE, *The Vigilantes of Montana* (various editions); C.B. GLASSCOCK, *The War of the Copper Kings* (1954); JOSEPH KINSEY HOWARD, *Montana: High, Wide and Handsome*, new ed. (1959); MICHAEL P. MALONE and RICHARD B. ROEDER (eds.), *The Montana Past: An Anthology* (1969); K. ROSS TOOLE, *Montana: A Twentieth Century Portrait* (in press).

Economics: Montana Economic Study, pt. 1, "The Montana Economy," Summary, Bureau of Business and Economic Research, University of Montana (October 1970); "Some Views from Indian Country," special issue of *Montana Business Quarterly*, vol. 8, no. 4 (Autumn 1970).

(D.M.J.)

Montesquieu

Remembered now mainly as a liberal political philosopher, Montesquieu was one of the outstanding Frenchmen of the 18th century. His seminal *L'Esprit des lois* was a major contribution to political theory, but he was as well-known in his own time as a prominent social figure, an acute satirist, and a brilliant stylist.

Early life and career. He was born Charles-Louis de Secondat in the 15th-century château of La Brède, near Bordeaux, on January 18, 1689. His father, Jacques de Secondat, belonged to an old military family of modest wealth that had been ennobled in the 16th century for services to the crown, while his mother, Marie-Françoise de Pesnel, was a pious lady of partial English extraction. She brought to her husband a great increase in wealth in the valuable wine-producing property of La Brède. When she died in 1696, the barony of La Brède passed to Charles-Louis, who was her eldest child, then aged seven. Educated first at home and then in the village, Charles-Louis was sent away to school in 1700. The school selected was the Collège de Juilly, close to Paris and in the diocese of Meaux. It was then much patronized by the prominent families of Bordeaux, and the priests of the Oratory, to whom it belonged, provided a sound education on enlightened and modern lines.

By courtesy of the Academie Nationale des Sciences, Belles Lettres et Arts de Bordeaux, France; photograph, Studio Denis

Montesquieu, oil painting by an unknown artist, 1718. In the Académie Nationale des Sciences, Belles Lettres et Arts de Bordeaux, France.

Charles-Louis left Juilly in 1705, continued his studies at the faculty of law at the University of Bordeaux, graduated, and became an advocate in 1708; soon after he appears to have moved to Paris in order to obtain practical experience in law. He was called back to Bordeaux by the death of his father in 1713. Two years later he married Jeanne de Lartigue, a wealthy Protestant, who brought him a very respectable dowry of 100,000 livres and in due course presented him with two daughters and a son, Jean-Baptiste. Charles-Louis admired and exploited his wife's business skill and readily left her in charge of the property on his visits to Paris. But he does not appear to have been either faithful or greatly devoted to her. In 1716 his uncle, Jean-Baptiste, baron de Montesquieu, died and left to his nephew his estates, with the barony of Montesquieu, near Agen, and the office of deputy president in the Parlement of Bordeaux. The *parlements* were regional courts of law and at this time

were laying some claim to political authority. His position was one of some dignity. It carried a stipend but was no sinecure.

The young Montesquieu, at 27, was now socially and financially secure. He settled down to exercise his judicial function (engaging to this end in the minute study of Roman law), to administer his property, and to advance his knowledge of the sciences, especially of geology, biology, and physics, which he studied in the newly formed academy of Bordeaux. His intellectual horizons might have become fixed in the southwest of France, but Montesquieu had already been attracted by the capital. He had enjoyed what he saw of the small intellectual and social coteries that flourished in Paris, in which new and bold ideas were canvassed and in which young men like himself might meet established figures. Throughout his life his loyalty to Bordeaux was challenged, though never overcome, by the lure of Paris, and Montesquieu was at once the feudal *seigneur* and the witty man of letters.

In 1721 he surprised all but a very few close friends by publishing his *Lettres persanes*, in which he gave a brilliant satirical portrait of French and particularly Parisian civilization, supposedly seen through the eyes of two Persian travellers, Usbek and Rica. This exceedingly successful work mocks the reign of Louis XIV, which had only recently ended; pokes fun at all social classes; discusses, in its allegorical story of the Troglodytes, the theories of Hobbes relating to the state of nature; makes an original, if naïve, contribution to the new science of demography; continually compares Islām and Christianity; reflects the controversy about the papal bull *Unigenitus*, which was directed against the dissident Catholic group known as the Jansenists; satirizes Roman Catholic doctrine; and is infused throughout with a new spirit of vigorous, disrespectful, gay, and iconoclastic criticism. The work's anonymity was very soon penetrated and Montesquieu became famous. The new ideas fermenting in Paris had received their most scintillating expression.

Social progress in Paris

Montesquieu now sought to reinforce his literary achievement with social success. Going to Paris in 1722, he was assisted in entering court circles by the Duke of Berwick, the exiled Stuart prince whom he had known when Berwick was military governor at Bordeaux. The tone of life at court was set by the rakish regent, the Duc d'Orléans, and Montesquieu did not disdain its dissipations. His notebooks and his correspondence hint at a number of amorous adventures, and it was in their background that he produced his *Temple de Gnide* (1725), a mild essay in eroticism written, it is thought, for the royal princess Mademoiselle de Clermont. He frequented the dissolute, gay society of Belébat, the country house of an adventuress called Madame de Prie, who was known also to Voltaire. On a more serious level, he was admitted to the literary salon of Madame de Lambert, regarded by many as the anteroom to the French Academy. He appears to have visited the Club of the Entresol, a kind of political academy the members of which included the Marquis d'Argenson and the abbé Castel de Saint-Pierre, both of them advanced and unconventional political thinkers.

It was during this period that he made the acquaintance of the English politician Viscount Bolingbroke, whose political views were later to be reflected in Montesquieu's analysis of the English constitution. Montesquieu is thought to have read to the Entresol a short work, elegant but superficial, called *Dialogue de Sylla et d'Eucrate*, but if he did so, his reception was in no way memorable. The members of the Entresol and of the salon of Madame de Lambert certainly influenced Montesquieu's developing political ideas, and these received expression in an essay on duty (*Traité général des devoirs*), now lost but clearly, from extracts surviving, inspired by Cicero and the German jurist Samuel von Pufendorf. These extracts he read in 1725 to the Academy of Bordeaux, which was one of the strongest links binding him to Bordeaux. His interest, however, in the routine activities of the Parlement had dwindled. He resented seeing that his intellectual inferiors were more successful than he in court. His office was marketable, and in 1726 he sold it, a move that served both to re-establish his fortunes, depleted by life in the capital, and to assist him, by lending colour to his claim to be resident in Paris, in his attempt to enter the Académie Française. A vacancy there arose in October 1727. Montesquieu had powerful supporters, with Madame de Lambert's salon firmly pressing his claims, but success seemed likely to elude him because of the objection of Cardinal Fleury (who was the most powerful political figure in France at this time) to certain passages in the *Lettres persanes* that were hostile to religion. He succeeded, however, in appeasing Fleury and was elected, taking his seat on January 24, 1728.

Experiences on the grand tour

This official recognition of his talent might have caused him to remain in Paris to enjoy it. On the contrary, though older than most noblemen starting on the grand tour, he resolved to complete his education by foreign travel. Leaving his wife at La Brède with full powers over the estate, he set off for Vienna in April 1728, with Lord Waldegrave, nephew of Berwick and lately British ambassador in Paris, as travelling companion. He wrote an account of his travels as interesting as any other of the 18th century.

In Vienna he met the soldier and statesman Prince Eugene of Savoy and discussed French politics with him. He made a surprising detour into Hungary to examine the mines. He then entered Italy, and, after tasting the pleasures of Venice, proceeded to visit most of the cities of Italy. Conscientiously examining the galleries of Florence, notebook in hand, he developed his aesthetic sense. In Rome he heard the French minister Cardinal Polignac and read his unpublished Latin poem *Anti-Lucretius*. In Naples he skeptically witnessed the liquefaction of the blood of the city's patron saint. He became aware of the works of both the notoriously persecuted historian Pietro Giannone and the then obscure but original philosopher of history Giambattista Vico. From Italy he moved through Germany to Holland and thence (at the end of October 1729), in the company of the statesman and wit Lord Chesterfield, to England, where he remained until the spring of 1731. His account of his visit to England was destroyed by his grandson, and much less is known of it than of the Italian journey.

Montesquieu had a wide circle of acquaintances in England, especially among the Tories and the dissident Whigs. He was presented at court, and he was received by the Prince of Wales, at whose request he later made an anthology of French songs, many of them of a dubious character. He became a close friend of the dukes of Richmond and Montagu. He was elected a Fellow of the Royal Society. He attended parliamentary debates and read the political journals of the day, especially Bolingbroke's *Craftsman*. He became a Freemason. He bought extensively for his library. His stay in England was one of the most formative periods of his life.

Major works. During his travels Montesquieu did not avoid the social pleasures that he had sought in Paris, but his serious ambitions were strengthened. He thought for a time of a diplomatic career but on his return to France decided to devote himself to literature. He hastened to La Brède and remained there, working for two years. Apart from a tiny but controversial treatise on *La Monarchie universelle*, printed in 1734 but at once withdrawn (so that only his own copy is extant), he was occupied with an essay on the English constitution (not published until 1748, when it became part of his major work) and with his *Considérations sur les causes de la grandeur des Romains et de leur décadence* (1734). He had thought of publishing the two together, thus following an English tradition, for, as Voltaire said, the English delighted in comparing themselves with the Romans. Montesquieu was not a professional historian, and his book on the Romans was ill received by the learned. Certainly it was technically defective. It shows a disregard for archaeology surprising in an enthusiastic traveller in Italy, and its documentation is uncritical. In these respects Montesquieu was behind his time. But his comments on the history of Rome are penetrating and vigorous, and his approach to the study of historical causation is particu-

larly interesting. In the distinction that, under the influence of the 17th-century French philosopher Nicolas Malebranche and of Italian writers, he draws between cause and occasion, he finds a formula for the reconciliation of free will and historical determinism. He shows a unifying grasp of the varied aspects of civilization that few before him could rival. The hatred of despotism that characterized the *Lettres persanes* is here flanked by a devotion to liberty as the major political good. This is a mature and well-written book.

Preparation of *L'Esprit des lois*

Montesquieu's literary ambitions were far from exhausted. He had for some time been meditating the project of a major work on law and politics. After the publication of the *Considérations*, he rested for a short time and then, undismayed by failing eyesight, applied himself to this new and immense task. He undertook an extensive program of reading in law, history, economics, geography, and political theory, filling with his notes a large number of volumes, of which only one survives, *Geographica, tome II*. He employed a succession of secretaries, sometimes as many as six simultaneously, using them as readers and as amanuenses, but not as précis writers. An effort of this magnitude was entirely foreign to what was publicly known of his character, for he was generally looked on as brilliant, rapid, and superficial. He did not seek to disabuse the world at large. Only a small number of friends knew what he was engaged in. He worked much at La Brède, devoting himself also to the administration of his estates and to the maintenance of his privileges as a landed proprietor. But he continued to visit Paris and to enjoy its social life. He kept there a second library and also made use of the Bibliothèque du Roi. He attended the Académie, visited the salons, and enjoyed meeting Italian and English visitors. At the same time he persistently, unostentatiously pressed on with the preparation of the book that he knew would be a masterpiece. By 1740 its main lines were established and a great part of it was written. By 1743 the text was virtually complete, and he began the first of two thorough and detailed revisions, which occupied him until December 1746. Now the actual preparation for the press was at hand. A Geneva publisher, J. Barrillot, was selected, further corrections were made, several new chapters were written and in November 1748 the work appeared under the title *De l'esprit des loix, ou du rapport que les loix doivent avoir avec la constitution de chaque gouvernement, les moeurs, le climat, la religion, le commerce, etc.* (*The Spirit of Laws*, 1750). It consisted of two quarto volumes, comprising 31 books in 1,086 pages.

Montesquieu had once before astonished the public when, known as a provincial magistrate, he had produced the satirical *Lettres persanes*. Now, known rather as a wit, he produced the massive treatise comparable with Aristotle's *Politics*. Fame came to him immediately.

L'Esprit des lois is one of the great works in the history of political theory and in the history of jurisprudence. Its author had acquainted himself with all previous schools of thought but identified himself with none. Of the multiplicity of subjects treated by Montesquieu, none remained unadorned. His treatment of three was particularly memorable. The first of these is his classification of governments, a subject that was *de rigueur* for a political theorist. Abandoning the classical divisions of his predecessors into monarchy, aristocracy, and democracy, Montesquieu produced his own analysis and assigned to each form of government an animating principle: the republic, based on virtue; the monarchy, based on honour; and despotism, based on fear. His definitions show that this classification rests not on the location of political power but on the government's manner of conducting policy; it involves a historical and not a narrow descriptive approach.

Analysis of the separation of powers

The second of his most noted arguments, the theory of the separation of powers, is treated differently. Dividing political authority into the legislative, executive, and judicial powers, he asserted that in the state that most effectively promotes liberty, these three powers must be confided to different individuals or bodies, acting independently. His model of such a state was England, which he saw from the point of view of the Tory opposition to the Whig leader, Robert Walpole, as expressed in Bolingbroke's polemical writings. The chapter in which he expressed this doctrine—book xi, chapter 6, the most famous of the entire book—had lain in his drawers, save for revision or correction, since it was penned in 1734. It at once became perhaps the most important piece of political writing of the 18th century. Though its accuracy has in more recent times been disputed, in its own century it was admired and held authoritative, even in England; it inspired the Declaration of the Rights of Man and the Constitution of the United States.

The third of Montesquieu's most celebrated doctrines is that of the political influence of climate. Basing himself on doctrines met in his reading, on the experience of his travels, and on experiments—admittedly somewhat naïve—conducted at Bordeaux, he stressed the effect of climate, primarily thinking of heat and cold, on the physical frame of the individual, and, as a consequence, on the intellectual outlook of society. This influence, he claims, is not, save in primitive societies, insuperable. It is the legislator's duty to counteract it. Montesquieu took care (as his critics have not always realized) to insist that climate is but one of many factors in an assembly of secondary causes that he called the "general spirit." The other factors (laws, religion, and maxims of government being the most important) are of a nonphysical nature, and their influence, compared with that of climate, grows as civilization advances.

Society for Montesquieu must be considered as a whole. Religion itself is a social phenomenon, whether considered as a cause or as an effect, and the utility or harmfulness of any faith can be discussed in complete independence of the truth of its doctrines. Here and elsewhere, undogmatic observation was Montesquieu's preferred method. Sometimes the reader is beguiled by this into the belief that Montesquieu maintains that whatever exists, though it may indeed stand in need of improvement, cannot be wholly bad. It is then that a bold parenthesis or a rapid summing-up reminds one that for Montesquieu certain things are intrinsically evil: despotism, slavery, intolerance. Though he never attempted an enumeration of the rights of man and would probably have disapproved of such an attempt, he maintained a firm belief in human dignity.

In the final books of *L'Esprit des lois*, added at the last moment and imperfectly assimilated to the rest, he addressed himself to the history of law, seeking to explain the division of France into the two zones of written and customary law, and made his contribution to the much discussed controversy about the origins of the French aristocracy. Here he displays not only prudence and common sense, but also a real scholarly capacity, which he had not shown before, for the philological handling of textual evidence.

Reactions to *L'Esprit des lois*

After the book was published, praise came to Montesquieu from the most varied headquarters. The Scottish philosopher David Hume wrote from London that the work would win the admiration of all the ages; an Italian friend spoke of reading it in an ecstasy of admiration; the Swiss scientist Charles Bonnet said that Montesquieu had discovered the laws of the intellectual world as Newton had those of the physical world. The philosophers of the enlightenment accepted him as one of their own, as indeed he was. This, however, had grim consequences also. The intellectual life of Paris was bitterly divided at this time, and alignment with the Philosophes meant incurring the hostility of their enemies. The determinism that could be read into his theory of climate, the objectivity of his attitudes to monarchy and to religion, and his insistence on toleration and on liberty were all controversial, and a variety of denunciatory articles amd pamphlets appeared. Attacks made in the Sorbonne and in the general assembly of the French clergy were deflected, but in Rome, in spite of the intervention of the French ambassador and of several liberal-minded high ecclesiastics and notwithstanding the favourable disposition of the Pope himself, Montesquieu's enemies were successful, and the work was placed on the Index in 1751. This, though it

dismayed Montesquieu, was but a momentary setback. He had already published his *Défense de L'Esprit des lois* (1750). Subtle and good-humoured, but forceful and incisive, this was the most brilliantly written of all his works. His fame was now worldwide.

Last years. Renown lay lightly on his shoulders. His affability and modesty are commented on by all who met him. He was a faithful friend, kind and helpful to young and unestablished men of letters, witty, though absent-minded, in society. It was to be expected that the editors of the *Encyclopédie* should wish to have his collaboration, and d'Alembert asked him to write on democracy and despotism. Montesquieu declined, saying that he had already had his say on those themes but would like to write on taste. The resultant *Essai sur le goût*, first drafted some 25 years earlier, was his last work.

His affection for his country estate at La Brède remained unaltered, and he planned to retire there to continue literary work. But a short illness brought him to his death in Paris on February 10, 1755. The controversies that marked his last years surrounded his death also. He appears to have acknowledged a belief in God, to have been absolved, but to have refused to surrender to the church a notebook of corrections to the *Lettres persanes*.

MAJOR WORKS

Lettres persanes (1721; *Persian Letters, Translated [from the French of Montesquieu] by Mr. Ozell*, 2 vol., 1722); *De la monarchie universelle en Europe* (1734); *Considérations sur les causes de la grandeur des Romains et de leur décadence* (1734; *Reflections on the Causes of the Grandeur and Declension of the Romans. by the author of the Persian Letters, etc.*, 1734); *De l'esprit des loix, ou du rapport que les loix doivent avoir avec la constitution de chaque gouvernement, les moeurs, le climat, la religion, le commerce, etc.* (1748; *The Spirit of Laws, Translated [by Thomas Nugent] . . . with Corrections and Additions Communicated by the Author*, 2 vol., 1750; *The Spirit of Laws, by C. de Secondat, Baron de Montesquieu, translated by Thomas Nugent, revised by J.V. Pritchard, including On the Origin of Inequality, On Political Economy, The Social Contract, by Jean Jacques Rousseau*, trans. by G.D.H. Cole, 1952); *Défense de L'Esprit des loix à laquelle on à joint quelques éclaircissements* (1750).

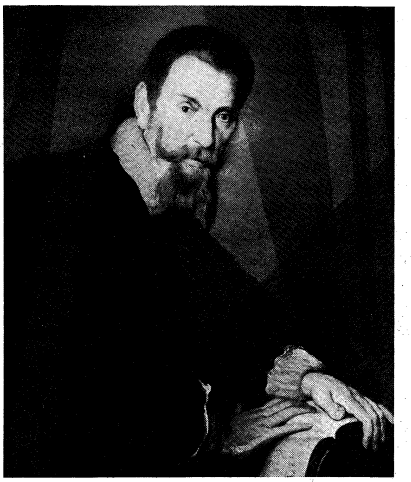
BIBLIOGRAPHY. R. SHACKLETON, *Montesquieu: A Critical Biography* (1961); J. DEDIEU, *Montesquieu, l'homme et l'oeuvre* (1943); the best edition of the *Oeuvres complètes* is by ANDRE MASSON, 3 vol. (1950–55). The "Bibliothèque de la Pléiade" has a useful edition by R. CAILLOIS, 2 vol. (1949–51). The best edition of *De l'esprit des loix* is by J. BRETHE DE LA GRESSAYE, 4 vol. (1950–61); the 18th-century English translation has been reprinted (1949). D.C. CABEEN, *Montesquieu: A Bibliography* (1947), is useful.

(R.Sh.)

Monteverdi, Claudio

Of all the composers who participated in the musical revolution that centred in Italy at the end of the 16th century, Claudio Monteverdi accomplished most in the new genres of opera and oratorio and in writing in the new style of accompanied melody, rather than weaving independent melodic lines in the old polyphonic style. In the late 19th century, historians regarded him as *the* revolutionary who was responsible for the decline of that style, which had culminated its long development in the 16th-century church music of Palestrina; later research, however, showed that Monteverdi was only one of those involved in this trend. He based his work on the belief that music must "move the whole man"; that is, that it must express man's deepest feelings, even in the lightest songs, and that, to do so, it must be joined with words. Even in his earlier madrigals (*i.e.*, unaccompanied secular vocal pieces in a complex polyphonic style), he demonstrated his skill in combining contrasting emotions into a satisfying unit. The same gifts made him a supreme composer of operas in which the characters are as alive as those in the works of Mozart and Verdi.

Early career at the Gonzaga court. Claudio Monteverdi was baptized in the north Italian town of Cremona on May 15, 1567, the son of a barber-surgeon and chemist. He studied with the director of music at Cremona Cathedral, Marcantonio Ingegneri, a well-known musician who wrote church music and madrigals of some distinction in an up-to-date though not revolutionary style of the 1570s. Monteverdi was obviously a precocious pupil, since he published several books of religious and secular music in his teens, all of them containing competent pieces in a manner not far from that of his master. The culmination of this early period occurred in two madrigal books published by one of the most famous of Venetian printers in 1587 and 1590. They are full of excellent, attractive works, somewhat more modern in approach than Ingegneri's, perhaps the result of studying the madrigals of Luca Marenzio (1553–99), the greatest Italian madrigalist of the time, and others. As yet, however, Monteverdi's aim appears to be to charm rather than to express passion; it is exemplified at its best in such a madrigal as the well-known setting of the poem "Ecco mormorar l'onde" ("Behold the Murmuring Sea") by Torquato Tasso, the greatest Italian poet of the late Renaissance.



Monteverdi, oil painting by Bernardo Strozzi (1581–1644). In the Tiroler Landesmuseum Ferdinandeum, Innsbruck, Austria,
By courtesy of the Museum Ferdinandeum, Innsbruck

It is not known exactly when Monteverdi left his home town, but he entered the employ of the Duke of Mantua around 1590 as a string player. He immediately came into contact with some of the finest musicians, both performers and composers, of the time. Most influential on him seems to have been the Flemish composer Giaches de Wert, a modernist who, although no longer a young man, was still in the middle of an avant-garde movement in the 1590s. The crux of his style was that music must exactly match the mood of the verse and that the natural declamation of the words must be carefully followed. Since he chose the highly concentrated, emotional lyric poetry of Tasso and Tasso's rival Battista Guarini, Wert's music also became highly emotional, if unmelodious and difficult to sing. It had an immediate effect on Monteverdi, whose next book of madrigals, published in his first year at Mantua, shows the influence of the new movement on him, though his understanding was imperfect. It represented a complete change of direction for him. The melody is angular, the harmony increasingly dissonant, the mood tense to the point of neurosis. Guarini is the favoured poet, and every nuance of the verse is expressed, even at the expense of musical balance.

Monteverdi's role in the avant garde

The new style and ambience seems to have upset his productivity. Although he went on composing, he published very little for the next 11 years. In 1595, he accompanied his employer on an expedition to Hungary, and four years later to Flanders. About 1599 he married a singer, Claudia Cattaneo, by whom he had three children, one of whom died in infancy. When the post of *maestro di cappella*, or director of music, to the Duke became vacant on the death of Wert in 1596, Monteverdi was embittered at being passed over, but in 1602 he achieved the position, at the age of 35. He published two more books of madrigals in 1603 and 1605, both of which contain masterpieces. The avant-garde manner was now better assimilated into his idiom. While his aim was still

to follow the meaning of the verse in great detail, he solved the purely musical problems of thematic development and proportion. Although the dissonances became more severe and the melody sometimes still more angular, the total effect was more varied in emotion and less neurotic. If Guarini's eroticism stimulated a sensual musical style, Monteverdi often gave his mature madrigals a lightness and humour, seeing the essence of a poem rather than its detail.

It was the advanced musical means, especially the use of intense and prolonged dissonance, that provoked attacks by the conservatives on Monteverdi, who became a figurehead of the avant-garde group. The attacks by a Bolognese theorist, Giovanni Maria Artusi, in a series of pamphlets, made Monteverdi the most famous composer of the age and provoked him to reply with an important aesthetic statement of his view on the nature of his art. He disclaimed the role of revolutionary, saying that he was only the follower of a tradition that had been developing for the last 50 years or more. This tradition sought to create a union of the arts, especially of words and music, so that he should not be judged simply as a composer using conventional musical devices. Moreover, the artwork must be powerful enough to "move the whole man," and this again might mean the abandonment of certain conventions. On the other hand, he declared his faith in another and older tradition, in which music was itself supreme, and which was, in effect, represented by the pure polyphony of such composers as Josquin des Prez and Palestrina. Thus, there were two "practices," as he called them; and this view, which became immensely influential, was to prove the basis of the preservation of an old style in certain types of church music, as opposed to a modern style in opera and cantatas, a dichotomy that can be found well into the 19th century.

Monteverdi's credo

If the madrigals of this time gave him a reputation well outside northern Italy, it was his first opera, *Orfeo*, performed in 1607, that finally established him as a composer of large-scale music rather than of exquisite miniature works. Monteverdi may have attended some of the performances of the earliest operas, those composed by the Florentine composers Jacopo Peri and Giulio Caccini, and he certainly had written some stage music in previous years. In *Orfeo* he showed that he had a much broader conception of the new genre than did his predecessors. He combined the opulence of dramatic entertainments of the late Renaissance with the straightforwardness of a simple pastoral tale told in recitative (*i.e.*, a speechlike vocal style, with the emphasis on the narrative text), which was the ideal of the Florentines. His recitative is more flexible and expressive than theirs, based on the declamatory melody of his madrigals rather than on their theories about heightened speech. Above all, he had a greater gift for dramatic unity, shaping whole acts into musical units, rather than assembling them from small sections. He also showed a sense of matching the climaxes in the drama by musical climaxes, using dissonance, the singer's virtuosity, or instrumental sonorities to create the sense of heightened emotion.

His first opera

A few months after the production of *Orfeo*, Monteverdi suffered the loss of his wife, seemingly after a long illness. He retired in a state of deep depression to his father's home at Cremona, but he was summoned back to Mantua almost immediately to compose a new opera as part of the celebrations on the occasion of the marriage of the heir to the duchy, Francesco Gonzaga, to Margaret of Savoy. Monteverdi returned unwillingly and was promptly submerged in a massive amount of work. He composed not only an opera but also a ballet and music for an intermezzo to a play. Further disaster occurred when the opera, *L'Arianna*, was in rehearsal, for the prima donna, a young girl who had been living in Monteverdi's home, possibly as a pupil of his wife, died of smallpox. Nevertheless, the part was re-cast, and the opera was finally produced in May 1608. It was an enormous success. The score has been lost, except for the famous "Lamento," which survives in various versions and is the first great operatic *scena* (*i.e.*, a scene of especially dramatic effect, usually with arias).

After this enormous effort, Monteverdi returned again to Cremona in a condition of collapse, which seems to have lasted for a long time. He was ordered back to Mantua in November 1608 and refused to go. He eventually returned but thereafter he hated the Gonzaga court, which he maintained had undervalued and underpaid him, though he gained a raise in pay and a small pension for his success with *L'Arianna*. He does not, however, appear to have been uncreative, though the music he wrote in the next year or so reflects his depression. He arranged the "Lamento" as a five-voiced madrigal and wrote a madrigalian threnody on the death of his prima donna, the sestina, published later in the sixth book of madrigals, which represents the peak of dissonant, agonized music in this style. In a more vigorous vein, he wrote some church music, which he published in 1610 in a volume containing a mass in the old style and music for Vespers on feasts of the Blessed Virgin Mary. The mass was a remarkable achievement, a deliberate attempt to show that the polyphonic idiom was still possible when everywhere it was dying. Still more remarkable is the Vespers music, a virtual compendium of all the kinds of modern church music possible at the time—grand psalm settings in the Venetian manner, virtuoso music for solo singers, instrumental music used for interludes in the service, even an attempt to use up-to-date operatic music to set the expressive, emotional words of the Magnificat. Yet, though this music is as "advanced" as possible, Monteverdi makes it an extension of the old tradition by using plainsong tunes—ancient unaccompanied liturgical chants—as the thematic material for the psalms and Magnificats. Above all, this is music of the Counter-Reformation; using all means, traditional and new, secular and religious, it is designed to impress the listener with the power of the Catholic Church, and its Maker, in the face of the challenge of the rise of Protestantism.

The volume containing this music was dedicated to Pope Paul V, and Monteverdi visited Rome apparently to present it in person. He may also have been seeking a new post, in order to leave Mantua, but nothing came of this. Nothing is known of his life for the next two years. Then, Francesco Gonzaga succeeded to the duchy on the death of his father and suddenly dismissed Monteverdi from his service in July 1612 for reasons that are unknown today. Without a job, Monteverdi again went back to his father's house, with his two sons, and remained there for about a year.

Three decades in Venice. When the *maestro di cappella*—that is, the director of music—of St. Mark's in Venice died, Monteverdi was invited to take his place, after an audition of some of his music in the basilica. He finally took up his appointment in the autumn of 1613.

He was appointed largely because the musical establishment of St. Mark's was in need of an experienced director after some years of decline. The last of the native Venetian composers of any distinction, Giovanni Gabrieli, had recently died. Although Monteverdi had not been primarily a church musician, he took his duties extremely seriously and within a few years completely revitalized the music in the basilica. He hired new assistants (including two future composers of note, Francesco Cavalli and Alessandro Grandi), wrote much church music, and insisted on daily choral services. He also took an active part in music making elsewhere in the city, directing the music on several occasions for the fraternity of S. Rocco, an influential philanthropic brotherhood, on the annual festival of its patron saint.

His letters in those early years in Venice reveal a complete change in his state of mind from what it was in Mantua. He felt fulfilled and honoured, well (and regularly) paid, and he seems to have been reasonably prolific. He kept up his links with Mantua, largely because there was little chance of producing opera in Venice, while opportunities came quite regularly from the Gonzaga court. In his correspondence, a philosophy of dramatic music emerges that was not only to mold Monteverdi's later work but also to influence the history of opera in general. The older type of opera had developed, on the one hand, from the Renaissance intermezzo, a

The old and the new opera

short, static musical treatment, often allegorical and with scenery, of a subject from the play with which it was given, emphasizing the wishes of the gods; and, on the other hand, from the pastoral, with its highly artificial characterizations of shepherds and shepherdesses. Monteverdi, however, was increasingly concerned with the expression of human emotions and the creation of recognizable human beings, with their changes of mind and mood. Thus, he wished to develop a greater variety of musical means, and in his seventh book of madrigals, published in 1619, he experimented with many new devices. Most were borrowed from the current practices of his younger contemporaries, but all were endowed with greater power. There are the conversational "musical letters," deliberately written in a severe recitative melody in an attempt to match the words. The ballet *Tirsi e Clori*, written for Mantua in 1616, shows, on the contrary, a complete acceptance of the simple tunefulness of the modern aria. In addition, all this work shows an astonishing renewal for a man of about 50 years of age.

Later musical theories

His attempt to create a practical philosophy of music went on throughout the 1620s, leading to still further stylistic innovations. Following ideas derived from Plato, he divided the emotions into three basic kinds: those of love, war, and calmness. Each of these could be expressed by differing rhythms and harmonies. A further ingredient in his theories was a frank acceptance of realism—the imitating of the sounds of nature in various ways. All these ideas are to be found in his dramatic cantata, *Il combattimento di Tancredi e Clorinda* (*The Combat of Tancredi and Clorinda*), a setting of a section of Tasso's "Gerusalemme Liberata," which was first performed in Venice in 1624. In this work, the rapid reiteration of single notes in strict rhythms and the use of pizzicato—plucking strings—to express the clashing of swords show important steps forward in the idiomatic use of stringed instruments.

These trends were continued in a comic opera, *La finta pazza Licori* (*Licoris Who Feigned Madness*), probably intended for the celebrations of the accession of Duke Vincenzo II of Mantua in 1627. The score of this piece has been lost, but a sizable correspondence survives. At this time, Monteverdi suffered more anxiety since his elder son, Massimiliano, was imprisoned in Bologna, where he was a medical student, for reading books banned by the Inquisition. It took some months before he was finally cleared of the charge. In the same year, 1628, Monteverdi also fulfilled a commission to write music (now lost) for the intermezzi to Tasso's *L'Aminta* and a tournament given in Parma in celebration of the marriage of Duke Odoardo Farnese to Margherita de' Medici. This was in fact the last occasion when Monteverdi was to receive a commission from the northern Italian courts. The War of the Mantuan Succession broke his link with the Gonzagas, who were ruined by the war. Monteverdi may have written an opera for performance in Venice in 1630, but the plague that broke out in that year effectively stopped all musical activities in Venice and the provinces for about 18 months.

Monteverdi and his family seemed to have emerged unscathed, and Monteverdi himself took holy orders during this period. He wrote a grand mass for the thanksgiving service in Saint Mark's when the epidemic was officially declared over in November 1631. The "Gloria" from it still survives and shows him applying some of the theories concerning the diversity of mood suggested by the words. Both this and some other church music probably written about this time, however, show a calm and majestic approach rather than the passion of his earlier years. A book of lighthearted songs and duets published in the following year is much the same. There is also a detached quality about much of the music in the final collection of his madrigals assembled by Monteverdi himself in 1638. A vast retrospective anthology of music dating from 1608 onward, it sets out to display Monteverdi's theories, as its title, *Madrigali guerrieri et amorosi* (*Madrigals of War and Love*), denotes. It contains much fine music, especially the "Lament of the Nymph," another reworking of his mood of 1608, after he lost his wife, and of *L'Arianna*. in terms of later technical devices. It seems to indicate that the agony of the period was never completely forgotten.

Though this collection, put together when Monteverdi was over 70, might seem the end of his career, chance played a part in inspiring him to an Indian summer of astonishing productivity: the first public opera houses opened in Venice in 1637. As the one indigenous composer with any real experience in the genre, he naturally was involved with them almost from the beginning. *L'Arianna* was revived again, and no fewer than four new operas were composed within about three years. Only two of them have survived in score—*Il ritorno d'Ulisse in patria* (*The Return of Ulysses to his Country*) and *L'incoronazione di Poppea* (*The Coronation of Poppea*) —both are masterpieces. Although they still retain some elements of the Renaissance intermezzo and pastoral, they can be fairly described as the first modern operas. Their interest lies in revealing the development of human beings in realistic situations. There are main plots and subplots, especially in *Poppea*, allowing for a great range of characters—the nobility, their servants, the evil, the misguided, the innocent, the good. The music expresses their emotions with astonishing accuracy. Monteverdi shows how the philosophy of music evolved during his early years in Venice could be put to use, using all the means available to a composer of the time, the fashionable arietta (*i.e.*, a short aria), duets, and ensembles, and how they could be combined with the expressive and less fashionable recitative of the early part of the century. The emphasis is always on the drama: the musical units are rarely self-contained but are usually woven into a continual pattern, so that the music remains a means rather than an end. There is also a sense of looking toward the grand climax of the drama, which inspires a grand *scena* for one of the main singers, Ulisse, Nero, or Poppea. At the same time, there are enough memorable melodies for the opera to seem musically attractive.

With these works, Monteverdi proved himself to be one of the greatest musical dramatists of all time. That he was held in the highest esteem by his Venetian employers is shown by their gifts of money in these last years and by their granting him leave to travel to his native city in the last few months of his life. The Venetian public showed its esteem at his funeral, when after his death on Nov. 29, 1643, following a short illness, he was buried in the Church of the Frari, where a monument to him still remains.

MAJOR WORKS

Secular

OPERAS: Seven operas of which only three survive: *La favola d'Orfeo* (1607); *Il ritorno d'Ulisse in patria* (1641); *L'incoronazione di Poppea* (1642). Also surviving, the "Lamento" from *L'Arianna*.

OTHER DRAMATIC WORKS: *Il ballo delle ingrate* (1608), *Tirsi e Clori* (1616), and *La vittoria d'Amore* (1641, music lost), three ballets; *Il combattimento di Tancredi e Clorinda* (1624), a secular oratorio.

MADRIGAL COLLECTIONS: *Canzonette a tre voci, libro primo* (1584); seven books of madrigals (1587–1619); *Madrigali guerrieri et amorosi* . . . (1638), eighth book of madrigals; *Madrigali e canzonette* . . . (1651), ninth book of madrigals, published posthumously; *Scherzi musicali* (1607, 1632).

Sacred

COLLECTIONS: *Sacrae cantiunculae tribus vocibus* (1582); *Madrigali spirituali a quattro voci* (1583); *Sanctissimi Virginis Missa senis vocibus ad ecclesiarum choros ac Vesperae* . . . (1610), which includes *Missa da capella a sei voci* . . . and *Vespro della Beata Vergine* . . . ; *Selva morale e spirituale* (1640); *Messa a quattro voci et Salmi, a una, due, tre, quattro, cinque, sei, sette & otto voci* . . . (1650, published posthumously).

BIBLIOGRAPHY. Monteverdi's music is available in a modern collected (though not always accurate) edition in 16 volumes (some divided into several fascicules) by G.F. MALIPIERO (1926–42, reprinted *c.* 1950–67). A supplementary volume was published in 1966. Monteverdi's ample correspondence is printed in G.F. MALIPIERO, *Claudio Monteverdi* (1930); an English translation of the more important letters is contained in DENIS ARNOLD and NIGEL FORTUNE (eds.), *The Monteverdi Companion* (1968). There is as yet no standard biography in

any language, though the essential documentation of his life is contained in the classic study by E. VOGEL, *Claudio Monteverdi* (1887). There are three accurate biographies in English: H.F. REDLICH, *Claudio Monteverdi: Leben und Werk* (1949; Eng. trans., 1952); L. SCHRADE, *Monteverdi: Creator of Modern Music* (1950); and DENIS ARNOLD, *Monteverdi* (1963).

(D.M.Ar.)

Montevideo

Montevideo, the principal city and the capital of Uruguay and of the department (jurisdictional division) of Montevideo, is on the northern shore of the Río de la Plata Estuary, 120 miles (190 kilometres) east-southeast of Buenos Aires. Perhaps nowhere else in the world does a city predominate in a nation's political, industrial, commercial, and cultural life to the extent that Montevideo does in Uruguay. The city, the largest in the country and one of the great cities of South America, has an area of about 150 square miles (400 square kilometres). Its population in 1967 was 1,280,000. The area of the department, the southernmost, smallest, and most important in Uruguay, is about 200 square miles. The estimated population of the department in 1969 was 1,375,000. The name is derived from the hill (Spanish *monte*) that rises at the entrance to the bay. The hill itself, known as the Cerro (also means hill), surmounted by its light, looms up, a famous landmark, as the city is approached from the sea.

History. Montevideo was founded in 1726 by Bruno Mauricio de Zabala, governor of Buenos Aires, to counteract the Portuguese advance into the area from Brazil. The first families came from Galicia and the Canary Islands, lured by offers of cattle, sheep, and respectability in the form of the title hidalgo. During its early years, Montevideo was mostly a Spanish garrison town. Trade expanded toward the end of the colonial period, and Montevideo's merchants played an important part in securing Uruguayan independence. From 1807 to 1830 Montevideo was alternately occupied by British, Spanish, Argentine, Portuguese, and Brazilian forces, and its trade and population suffered. Independence, which came in 1830, did not bring stability. Uruguay was the scene of complicated interaction of local, Argentine, and Brazilian influences that culminated in the nine-year great siege of Montevideo by a combined Argentine-Uruguayan army from 1843 to 1851. Montevideo defenders were assisted by French and English forces that blockaded Buenos Aires. The result, surprisingly, was that Montevideo flourished during the siege and became the major port of the Río de la Plata. The city's population in 1843 was estimated at 31,197, of whom 15,252 were non-citizen Europeans. The end of the great siege in 1851 and the restoration of normal trade patterns disrupted Montevideo's economy. In the 1870s the gradual development of the Uruguayan interior, spurred by railroad building outward from Montevideo and the installation of modern public utilities in the city gave Montevideo new life. A speculative building boom opened new suburbs. Its collapse in the 1890s slowed the city's growth, but in the 20th century its development has been rapid. A contributing factor has been the port, a government-operated enterprise that was opened in 1901. Formerly, ships had to be unloaded in the harbour by lighters, but after the port's modernization it could dock the largest ships for loading and unloading.

The city's five governments

For the political, economic, and social history of the city in the 20th century, see URUGUAY, HISTORY OF.

The modern city. *The site.* The site of the original city is a low rocky headland projecting westward between the estuary and an almost circular bay that forms the harbour, the headland sloping gently to its western extremity at Punta Sarandí and to the water's edge on either side. The city has expanded eastward over rolling country and northward around the eastern shore of the bay.

Climate. The climate of Montevideo, somewhat blustery in winter, is moderate, with the mean annual temperature around 61° F (16° C). Rainfall is adequate, around 40 inches (1,000 millimetres) a year.

City plan. The harbour district opens out into the Old City (Ciudad Vieja), which once was colonial Montevideo. The buildings there, for the most part, are old, and the streets are narrow; but the Old City is still the financial and business heart of Uruguay. The stock exchange (*bolsa*), the Banco de la República Oriental del Uruguay (Bank of the Republic of Uruguay), head offices of private banks, foreign bank branches, shipping firms, export–import houses, and lawyers locate there. The Old City culminates in the Plaza de la Constitución, with the cathedral on one side and the *cabildo,* the colonial municipal building and for many years the home of the Uruguayan legislature, on the other. Calle Sarandí, a narrow shopping centre, connects the Plaza de la Constitución with the more recent Plaza de la Independencia, Montevideo's major square. The traditional gesture for distinguished foreign visitors is to place a wreath of flowers at the foot of the mounted statue of the hero of national independence, José Gervasio Artigas, in the centre of the plaza. Government House is on the right of the statue, the modernistic Victoria Plaza Hotel on the left. Facing the statue is the Palacio Salvo, the city's first skyscraper. Leading off the Plaza de la Independencia is the Avenida 18 de Julio, lined with shops, theatres, cafés, and business houses. Part way up the Avenida 18 de Julio, the Avenida Agraciada, cut diagonally through city blocks, connects with the marble legislative palace, the most lavish public building in Uruguay. Avenida Agraciada is Montevideo's modernistic showplace. At the end of Avenida 18 de Julio is an obelisk at the entrance of Parque José Batlle y Ordóñez, location of the football stadium.

The Old City

Transportation. The port is the hub of the city's international transport. It is also served by the airport at Carrasco. Four railroads converge on the city, and roads lead to other principal cities. Internally, buses, Montevideo's basic public transportation, taxis, and private cars, although directed by one of South America's few traffic-light systems, crowd the streets and avenues.

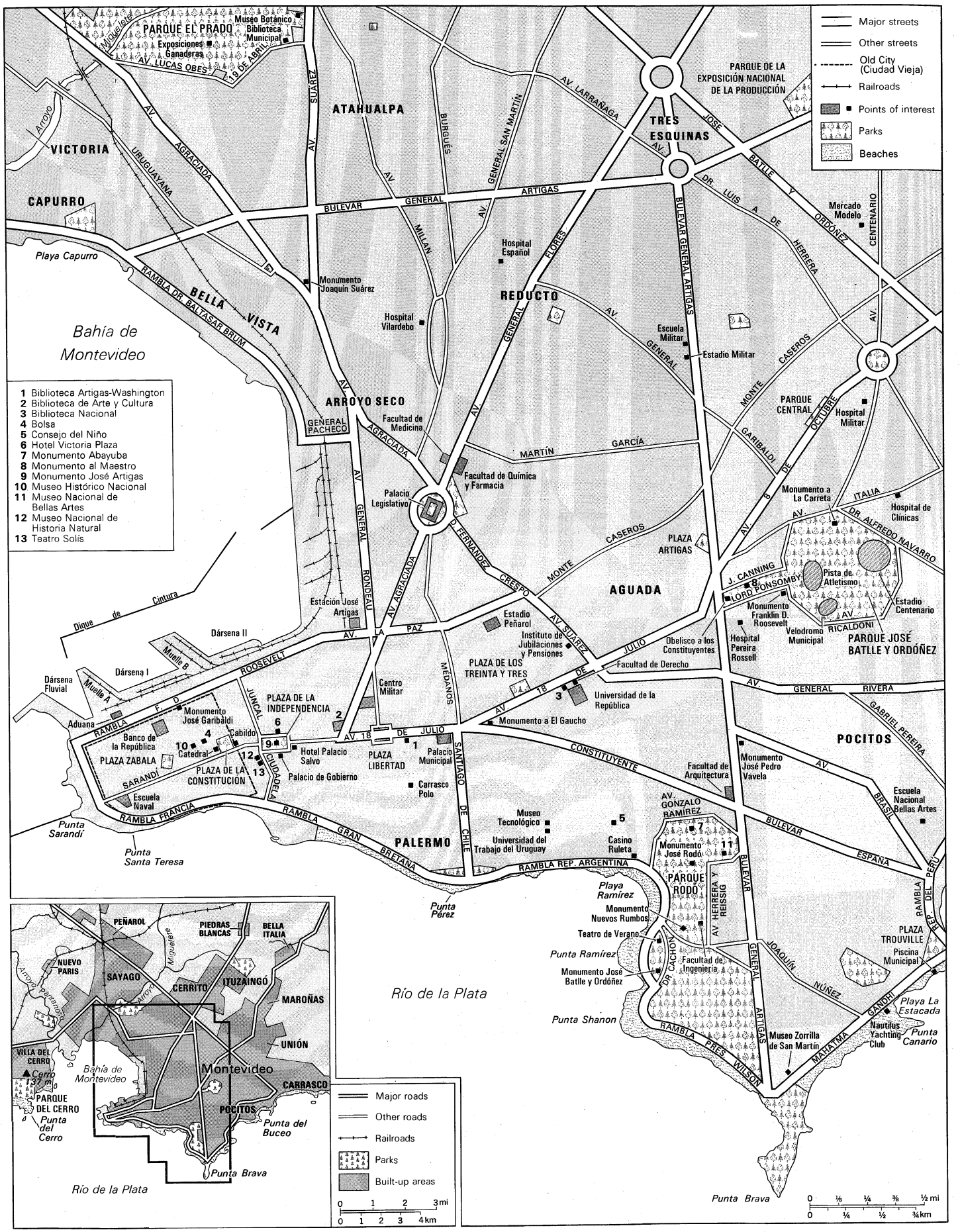
Population. The population of the department of Montevideo in 1860 was 58,000. By 1908 the population was 300,000, and by 1963 (the next census) it was about 1,200,000, representing 46 percent of the nation's total. The population in 1969 was 1,375,000. Because of immigration from the interior, it is growing at an annual rate of 16.7 per thousand, more than twice as fast as the rest of the country. Population density in the department is more than 6,900 per square mile.

In the 1800s Montevideo attracted immigrants of many nationalities, predominantly Spanish and Italians but also including Basques, Frenchmen, Navarrese, and Canary Islanders. The present population is predominantly white, mostly the descendants of the 19th- and 20th-century European immigrants.

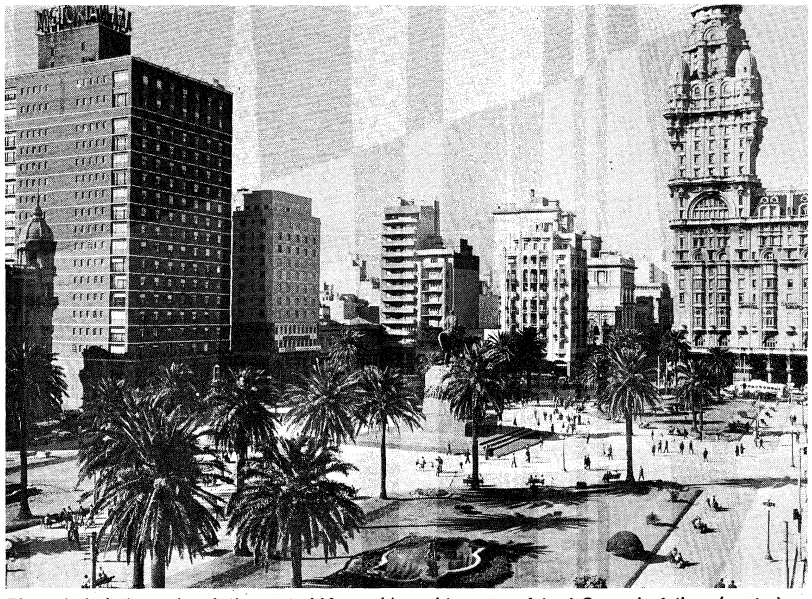
Economy. The port of Montevideo handles the vast bulk of Uruguay's foreign trade. Most of the country's manufacturing industry is there, too. The steady influx of migrants from the countryside, coinciding with the serious economic crisis of the 1950s and 1960s, meant high unemployment (estimated at almost 8 percent in 1971), strikes, the inability of the government to finance municipal services, and general deterioration of the city. Electric power failures and bus stoppages were taken for granted. Frustration with economic deterioration resulted in the formation of Montevideo's revolutionary urban guerrilla group, the Tupamaros, in the 1960s. Yet, Montevideo's size gave Uruguay a metropolis with urban cultural and economic opportunities far exceeding those a country with such a small population otherwise could afford.

The products of the interior are shipped for direct use, processing, and export. About 90 percent of Uruguay's exports are wool, meat, and hides from its ranches. Numerous establishments in the capital process wool, and refrigerated packing plants in the Cerro district are equipped to prepare meats. In addition to export-oriented processing industries, Montevideo possesses the largest portion of Uruguay's internally oriented and economically depressed consumers' goods industries. Textile, shoe, soap, match, and clothing factories are scattered throughout the city. Wines and dairy products, mostly for local consumption, also are produced in Montevideo.

Principal exports



Central Montevideo and (inset) its metropolitan area.



Plaza de la Independencia in central Montevideo with statue of José Gervasio Artigas (centre) and the Hotel Palacio Salvo (right).
Carl Frank

Uruguay's well-known state enterprises—under-capitalized because of the nation's galloping inflation, overstaffed as the result of political patronage, and operating under serious deficits—the ANCAP (Administración Nacional de Combustibles, Alcohol y Portland) oil refineries and cementworks, railway shops, and electric-power system are concentrated in Montevideo. Public employees are a large segment of the city's consumers. Much of the city's economy revolves around providing varied services for its large population, which is notably spread out territorially. The city's growth has urbanized and industrialized that part of the department lying outside the city. Nevertheless, truck gardens, vineyards, dairy and poultry farms, orchards, and commercial flower growers send products into the city.

Government. As the capital of the country, Montevideo is the seat of the principal institutions, agencies, and offices of the national government. The city itself is governed by an elected alcalde (mayor) and a municipal council that also has jurisdiction over the department.

Public utilities and services. Montevideo is one of the few South American capitals in which the public water supply, brought from the Río Santa Lucía is totally safe to drink from the tap, but the city's rapid growth has outstripped expansion of the water-pipe system.

Social services

Public health, sanitation, and welfare services have been among the best in South America as measured by such standards as the number of persons per physician, hospital beds per thousand, and infant mortality, but severe economic depression in the 1960s and early 1970s strained all agencies' abilities to cope with the growing needs.

Education. Primary and secondary education is free and compulsory but has not been able to keep up with growing needs. Higher education in Uruguay is available only in the capital, although plans for a university in the north were under way in the early 1970s. The Universidad de la República was founded in 1849. It has ten schools, including the schools of humanities and sciences, engineering, and medicine. The Universidad del Trabajo del Uruguay provides vocational training through industrial and night schools.

Cultural facilities. Montevideo has had theatres since the Casa de Comedias opened in 1795, followed by the San Felipe Theatre. The Solís Theatre, still in existence, was opened in 1856. The city also has the Museo Histórico Nacional (1900), Museo Nacional de Historia Natural (1837), the Museo Nacional de Bellas Artes (1911), and the Biblioteca Nacional del Uruguay (National Library), as well as the many plazas, buildings, and other historic sites.

The British, who occupied the city in 1807, published the first newspaper in Montevideo, *The Southern Star.* The city has had newspapers ever since, especially newspapers and other publications presenting the views of all political parties and factions. The media are now strongly supplemented by radio and television broadcasts.

In addition to Batlle y Ordóñez Park, site of the football stadium, recreational facilities include many other parks, as well as a line of beaches extending out to the Atlantic, where new resorts continue to arise.

BIBLIOGRAPHY. GEORGE PENDLE, *Uruguay,* 3rd ed. (1963), a perceptive analysis of Uruguay, including descriptions of Montevideo; MARVIN ALISKY, *Uruguay: A Contemporary Survey* (1969), references to Montevideo within a general discussion of Uruguay; RUSSELL H. FITZGIBBON, *Uruguay: Portrait of a Democracy* (1954), contains a chapter on Montevideo, written before the impact of the economic crisis; *Serie Montevideo* (1971), eight 60-page pamphlets in Spanish on Montevideo, its history, social problems, population, art and architecture, and literature.

(M.I.V.)

Montfort, Simon de

Simon de Montfort, earl of Leicester and leader of the baronial revolt against Henry III of England, was the son of Simon de Montfort l'Amaury, leader of the crusade against the heretical Albigenses. Born about 1208 at Montfort in the Île-de France, Simon was wholly French by birth and education. On coming of age, he renounced to his eldest brother Amaury his claims on the family lands in return for the sole right to revive the Montfort claim to the English earldom of Leicester. This claim derived from his father's mother, Amicia, sister of Robert IV (died 1204), the last Beaumont earl of Leicester, whose lands had been divided between Amicia and her younger sister Margaret, countess of Winchester. King John had recognized Simon's father as earl (*c.* 1205) but had deprived him as a French subect (1207), and the Montfort claim had then lapsed.

Simon came to England in 1229 and, helped by his cousin Ranulf, earl of Chester, the tenant of the confiscated estates, obtained the honour of Leicester and did homage to Henry III in 1231, though he was not formally styled earl of Leicester until April 1239. He speedily became one of Henry's favourites, receiving an annual fee of 500 marks to compensate for the divided inheritance and exercising the hereditary stewardship at the coronation of Queen Eleanor (Eleanor of Provence; 1236). Henry arranged for his sister Eleanor to marry Simon on January 7, 1238, thus breaking Eleanor's earlier vow of chastity and offending the English noblemen, who

Simon de Montfort, seal, 1258. In the British Museum (Additional Charter 11296).
By courtesy of the trustees of the British Museum

were not consulted. Henry's brother, Richard, earl of Cornwall, led an angry baronial protest, and Henry, alarmed, turned against Simon and Eleanor, driving them from England (August 1239). Simon went on crusade (1240–42) with Richard, with whom he was now reconciled, and won great prestige among the lords of the Latin Kingdom of Jerusalem, who asked their absentee king, the emperor Frederick II, to appoint Simon as his viceroy there. Returning to England, Simon joined Henry's disastrous invasion of France (1242), winning distinction by covering Henry's escape after his defeat at Saintes. Reconciled with Henry, and accepting an unfavourable settlement of Countess Eleanor's dower claims, Simon now made Kenilworth Castle (a royal grant) his headquarters. He cultivated the friendship of the radical reformer Robert Grosseteste, bishop of Lincoln, and took Robert's friend, the Oxford Franciscan Adam de Marisco, as spiritual director. Although regarded as a king's man, Simon was one of the committee of 12 appointed to handle the acute crisis of 1244 between Henry and his angry barons. He also took part in many important embassies to the French, papal, and imperial courts, and as a result he won many influential friends.

The influence of Grosseteste

In 1248 Henry asked Simon to pacify the English-held duchy of Gascony, in southwestern France. Simon, eager to join Louis IX's crusade, accepted reluctantly, stipulating for full powers as regent for seven years, without fear of recall and with full refund of expenses incurred. Treating the Gascon nobles as faithless rebels outside the law, he ruthlessly crushed the revolt and restored order; the Gascons appealed to Henry, accusing Simon of illegal procedure and oppression, and threatening to renew their revolt. The matter was complicated by Simon's personal contest with Gaston de Béarn, the leading rebel. Henry, frightened, recalled Simon for trial on the rebels' charges; the English magnates acquitted him (1252) and he returned to Gascony to suppress the renewed revolt, but Henry now terminated his lieutenancy. Simon, accepting a partial financial settlement, withdrew to France, though Henry had to implore his help in his own campaign against the rebels in 1253. Such was Simon's international reputation that when Louis IX's mother, Blanche of Castile, died (November 1252) while Louis was still on crusade, the French magnates invited Simon to succeed her as regent.

Henry's behaviour over Gascony, though not wholly unjustified, convinced Simon that Henry was unfit to rule, and the king's disastrous undertaking, at Pope Innocent IV's behest, to conquer Sicily for his son Edmund, strengthened this conviction. Discussions with Grosseteste, Marisco, and other Franciscan intellectuals had fired Simon's mind with visions of a new order in both church and state, and he joined the other leading English barons in forcing upon Henry the revolutionary Provisions of Oxford (June 1258). The reformers began well, but by October 1259 divisions appeared between the conservative wing, led by Richard de Clare, earl of Gloucester, that sought only to limit abuses of royal power, and the radical element, led by Simon, that sought to bind the entire baronage to observe the reforms forced upon the king and his officers. Simon exacerbated the quarrel by his arrogant vehemence and put himself in the wrong by attempting to use Henry's subordination to secure settlement of his own and his wife's justifiable personal claims on Henry. Henry, allying with the Gloucester faction, shattered baronial unity early in 1260, and Simon emerged leading the extremist defenders of the reforms. By October 1261 Henry had isolated Simon, who went abroad; but the king's annulment of the Provisions, after he had received papal absolution from his oath to observe them, revived general disaffection (1262), and Simon returned (April 1263) to lead a rebellion that restored the Provisions (July 1263). But baronial unity had vanished, and despite passionate support from the lesser barons, the county knights, the men of London and the Cinque Ports, and many clergy, Simon was forced to accept arbitration by Louis IX (December 1263). By the Mise of Amiens (January 1264) Louis totally annulled the Provisions and all consequent reforms: Simon rejected the award and after unsuccessfully attempting direct negotiations, defeated Henry at Lewes (May 14, 1264), capturing Henry and his son, the lord Edward.

Rebellion against Henry III

Simon then governed England by military dictatorship, striving unsuccessfully for a legal basis of consent, both by negotiations with Henry's supporters and by calling representatives of both shires and boroughs to Parliament (1265) to counterbalance his lack of baronial support. But his monopolization of power alienated his chief ally, the young Gilbert de Clare, earl of Gloucester, who joined the royalist Marcher lords and secured the Lord Edward's escape at Hereford (May 1265). By rapid and skillful manoeuvring, Edward isolated Simon behind the Severn, destroyed at Kenilworth (August 1) the large army coming to his rescue, and trapped Simon's little force at Evesham (August 4, 1265), slaying Simon and most of his followers.

The most outstanding English personality of his day, Simon is remembered as an early advocate of a limited monarchy, ruling through elected councillors and responsible officials, and of parliaments including county knights and burgesses as well as the great nobles.

BIBLIOGRAPHY. F.M. POWICKE, R.F. TREHARNE, and C.H. LEMMON, *The Battle of Lewes, 1264* (1964), contains a long study of Simon de Montfort. See also C.H. KNOWLES, *Simon de Montfort, 1265–1965* (1965), a critical review of historians' ideas on de Montfort; and M.W. LABARGE, *Simon de Montfort* (1962), a good, general biography. CHARLES BEMONT, *Simon de Montfort, comte de Leicester* (1884), contains a valuable appendix of original documents, which has not been included in the English translation, *Simon de Montfort, Earl of Leicester* (1930).

(R.F.Tr.)

Montgolfier, Jacques-Étienne and Joseph-Michel

The Montgolfier brothers, sons of a wealthy 18th-century French paper manufacturer, invented the hot-air balloon, conducted the first free flights, and experimented with hydrogen-filled balloons. Their flights were spectacular applications of new concepts in physics and chemistry. Modifications and improvements of the basic Montgolfier design were incorporated in the construction of larger balloons that, in later years, opened the way to exploration of the upper atmosphere.

The Montgolfier family, originally from Frankenthal, Germany, settled in the 13th century in Ambert, Auvergne, where they introduced the paper industry. A member of the family had learned papermaking in Damascus during the preceding century while being held there as a slave after one of the crusades. In 1695 the family transferred the industry to the small town of Vidalon, near Annonay, in southern France. Joseph and Étienne were two of the 16 children of Pierre Montgolfier, whose paper factories insured the financial support of their balloon experiments.

The family's origins

Joseph-Michel Montgolfier was born on August 26, 1740, at Annonay, where, early in life, he showed talent

Joseph-Michel (left) and Jacques-Étienne Montgolfier, miniatures on ivory by an unknown artist, end of the 18th century. In the Musée Carnavalet, Paris.
By courtesy of the Musee Carnavalet, Paris; photograph, © Jacques Buchholz

for chemical and mechanical experimentation; he developed, for example, a new technique for making the dye Prussian blue. After assisting his father in the family firm, in about 1770 he took over the management of a branch of the firm at Rives, where he introduced improvements in papermaking. He also invented a special air pump for removing paper from molds and a vacuum distillation apparatus. In 1771 he married his cousin, Thérèse Filhol, who, through her sound business sense, was an important asset to him. Since he was adept at new inventions, Joseph's attention eventually turned to the project that brought him fame rather than to day-to-day problems of industrial production. Undoubtedly a man of genius, he could solve physical or mechanical problems even though he lacked extensive formal education in science and mathematics; he even tried to build a steam-operated vehicle. Moreover, he developed basic insights in physics, such as the principle that motion is conserved when transmitted through two or more contiguous bodies.

Born at Annonay on January 6, 1745, Jacques-Étienne Montgolfier was educated in physics, chemistry, and architecture in Paris, where for a time he assisted in designing new churches and factories; he was also employed by a famous wallpaper firm. Although devoted to his brother, Étienne was much different in temperament. Charming, good-natured, and amiable in disposition, he made friends with the intellectuals of the day. In 1772 Étienne's father summoned him home to manage the paper plant at Vidalon; there he introduced a procedure for crushing paper pulp and began the production of vellum and drawing paper. These improvements soon attracted the attention and recognition of the inspector general of French industry.

It is not known exactly what event or idea led the Montgolfier brothers to entertain the idea of aerial navigation. But Joseph began the experiments, in which he was quickly joined by his younger brother, that resulted in the hot-air balloon. Often in collaboration, but sometimes working separately, the brothers conducted many different tests aimed at keeping heavy objects aloft. Having learned of the "inflammable gas" that was lighter than air (called hydrogen by Antoine Lavoisier in 1789), they found that it could easily inflate a paper or cloth container. But they could not prevent its escape through the thin, porous material; the gas also was expensive and cumbersome to prepare, and no one yet realized how dangerous it was to use. Joseph also thought of electrically charging large, air-filled containers so that an "electric fluid" would thereby be formed around them; this weightless zone would increase the buoyancy of the container, thus causing it to rise. Although he may not have realized that hot air expands and becomes lighter in density, Joseph nevertheless acted on the observation that the products of combustion generally rise. It is not known whether or not he thought of smoke as a special kind of gas; but in November 1782, in Avignon, he devised a means to inflate a small silk container with hot air by suspending it over a fire. He was then able to demonstrate in public the ascension of a three-to-four-cubic foot (0.08-to-0.1-cubic metre), oblong-shaped balloon that he had constructed of silk. At once he notified his brother Étienne in Annonay. In the coming months they built and launched even larger hot-air balloons with considerable success.

The first flights

On June 5, 1783, the two brothers launched, in the presence of local government officials in Annonay, a spherical balloon made of linen and paper; 30 feet (nine metres) in diameter, when inflated, and open at the bottom to receive hot air from a fire on the ground, the balloon rose about 6,000 feet (about 2,000 metres) and remained aloft for ten minutes. This event caused great astonishment in France. An official record was sent to the Academy of Sciences in Paris, which immediately invited the Montgolfier brothers to make a special demonstration. Their craft was later called a Montgolfière, or fire balloon.

While making preparations in Paris, Étienne, on August 27, 1783, assisted at the launching of the first hydrogen-filled balloon, which was invented by Jacques-Alexandre-César Charles, who also is noted in chemistry for his law relating the volume of a gas to its temperature. Charles had prepared the silk envelope by covering it with varnish, thus solving the problem of porosity that had baffled the Montgolfiers. His type of balloon was henceforth called a hydrogen balloon (and also a Charlière). On September 12 Étienne attempted to launch his own hot-air balloon for the Academy of Sciences, but it was destroyed on the ground by wind and rain. He then hurriedly constructed a new model made of reinforced linen. On September 19, 1783, in the courtyard of the palace at Versailles, academicians, and throngs of curious spectators (including Benjamin Franklin) witnessed the inflating and launching of an elaborately decorated, almost spherical balloon. About 57 feet high, it carried three passengers —a sheep, a rooster, and a duck—suspended beneath in a wicker basket; it floated for about 8 minutes and landed safely about two miles from the launch site. These first balloons could remain aloft only for short periods, since they were inflated over a fire. Weighted at the bottom for stability, they would quickly rise when released, and then, as the hot air dissipated, would float gently Earthward. Étienne then constructed a balloon to carry human passengers, and many of its features were incorported in later designs. Its large envelope of 79,000 cubic feet (2,300 cubic metres; inflated) was oval-shaped; suspended beneath the open neck, 16 feet (five metres) in diameter, was a circular gallery for carrying passengers, a stove for providing the hot air, and fuel of sticks and straw. Beginning on October 12, 1783, he conducted a series of methodically graded trial ascents from a Parisian garden, with the balloon secured to the ground by ropes between upright poles.

On October 15, he invited the physicist Jean François Pilâtre de Rozier to board his craft, which, tethered to the ground, rose about 80 feet (25 metres). This was, to the best of modern knowledge, the first time that man left the Earth. On October 19, another passenger accompanied Pilâtre de Rozier on further trial ascents, rising to a height of about 324 feet (99 metres). On November 21 occurred the first free flight by man: guided by Pilâtre de Rozier and François Laurent, marquis d'Arlandes, Étienne's balloon was launched from the jardins de la Muette in the Bois de Boulogne in Paris. After floating for about 25 minutes at low altitude over the city, it landed safely at La Butte-Aux-Cailles, a few miles to the south. (Étienne Montgolfier himself never made a flight.)

Meanwhile, in October 1783, Joseph, who was attempting to solve the problem of steering, conceived the idea that a propulsive force could guide a ship when afloat so that it need not drift at the mercy of the wind; hot air might be forced from one or more openings in the side of the ship opposite from the direction in which the ship was to go. Somewhat later he designed and partially constructed a large, lenticular-shaped, hot-air balloon that

could be more easily pointed in any desired direction and so float at will in the air. Joseph was invited to Lyons to supervise the construction of a giant balloon, named "Le Flesselles," after the government official who sponsored the project. It was 102 feet (31 metres) in diameter, 126 feet (38 metres) high, and 700,000 cubic feet (20,000 cubic metres) in volume, with a gallery large enough to hold six people, fuel, water buckets, and a stove to provide the hot air. After preliminary inflations and tethered ascents, made difficult by the cold weather, a manned flight took place on January 19, 1784, with Pilâtre de Rozier assisting Joseph Montgolfier at the controls. They were accompanied by five dignitaries, the balloon remained aloft 16 minutes; but when the fabric suddenly split, there was a rapid escape of hot air, which caused the balloon to descend abruptly, and Joseph was slightly injured. This was his only flight.

Later accomplishments

The Montgolfier brothers' remarkable balloons quickly attracted attention in Europe and gained for them national recognition. Joseph was named a member of the Academy of Sciences and Étienne a corresponding member of the National Institute in December 1783. King Louis XVI raised their father to the peerage, further honouring the brothers with this hereditary title. During the French Revolution, Joseph used his new authority to protect numerous persons from quarrelling factions and thus saved many lives. Étienne was named administrator of a new provincial subdivision, where he served his constituency with considerable independence and also saved numerous persons from execution and the family firm from confiscation.

In 1792 Joseph invented the hydraulic ram (a completely new device that forces running water to a higher level), and in 1801 he was summoned to Paris to assist in organizing the Conservatoire des Arts et Métiers, where he served as instructor and administrator. Étienne died on August 2, 1799, while enroute home to Annonay from Lyons. Joseph survived until June 26, 1810; he died at Balaruc-les-Bains, following a stroke.

The Montgolfiers' fire balloons drew attention to the technical problems involved in constructing lighter-than-air vehicles: how high a balloon of a given size could go; control of inflation, ascent, and descent; reinforcement of the envelope; and the interaction of volume, temperature, wind, and altitude. Many manned ascents were made with both hot-air and hydrogen balloons.

Short-lived was an attempt to combine a Montgolfière and a hydrogen balloon, in which a hot-air balloon that was used to regulate position and direction was suspended beneath a hydrogen balloon that provided buoyancy. The first flight of this craft ended in a fiery crash on June 15, 1785, and took the life of Pilâtre de Rozier, its inventor, who was attempting to cross the English Channel. Technical and safety problems received great attention in the subsequent development and construction of hydrogen balloons, which in turn opened the way to the scientific exploration of the upper atmosphere.

BIBLIOGRAPHY. For biographical information on the family, see J.B.J. DELAMBRE, *Notice sur la vie et les ouvrages de M. Montgolfier* (1810); and JOSEPH MARIE DE GERANDO, *Notice sur M. Joseph Montgolfier, . . . le lue en séance générale, le 11 mai 1814* (1814, reprinted 1883); FELIX DIGONNET, *L'Invention de l'aérostation à Avignon en 1782 et les premières ascensions dans cette ville* (1906), detailed on Joseph and Etienne, with information also on their families; LEON ROSTAING, *La Famille de Montgolfier, ses alliances, ses descendants* (1910, reprinted 1960); and the well-illustrated COMTE HENRY DE LA VAULX and PAUL TISSANDIER, *Joseph et Etienne de Montgolfier* (1926). The eyewitness account of BARTHELEMY FAUJAS DE SAINT-FOND, *Description des expériences de la machine aérostatique de MM. de Montgolfier*, 2 vol. (1783–84), should be consulted for a study of the first balloon flights. The technical problems encountered by the Montgolfiers, and those that followed them, are discussed by GASTON TISSANDIER in *Histoire des ballons et des aéronautes célèbres*, 2 vol. (1887–89). For a more recent study, see the fully illustrated CHARLES DOLLFUS and HENRI BOUCHE, *Histoire de l'aéronautique* (1932).

(Ch.Do.)

Montreal

The most populous city of Quebec Province and of Canada, Montreal is one of the world's most uniquely situated cities. The present city proper occupies about one-third of Île de Montréal (Montreal Island), the largest of the 234 islands of the Hochelaga Archipelago, one of three archipelagos near the confluence of the Ottawa and St. Lawrence rivers. Built around and up Mont-Royal, which rises 763 feet (233 metres) above sea level (some 660 feet above the island shores), its 58-square-mile (150-square-kilometre) area is interspersed with several independent cities and towns of the 462-square-mile metropolitan area that covers Montreal and other islands, as well as both shores of the St. Lawrence. Montreal is the major seaport on the St. Lawrence River and Seaway, lying between the navigable waters of the open Atlantic Ocean to the east and of the Great Lakes to the west. The population of the city, according to the 1971 census, was 1,214,352; and of the metropolitan area, 2,743,208, representing growths of 2 and 30 percent, respectively, during the preceding decade.

Montreal is said to be the second largest French-speaking city (French and English are Canada's two official languages) in the world (after Paris), a boast more or less belied by its increasingly cosmopolitan character. A stroller through various districts of the city might come to believe himself in Italy, China, or Greece, as he might in New York City or San Francisco. In some areas of this "second largest French-speaking city in the world," however, a French-speaking wanderer or worker must express himself in English to be understood fully. This phenomenon reflects decades of dominance over Montreal's economic life by the English-speaking minority. It also accounts in large measure for the growing agitation within the French community of the city and province for greater economic and political concession from the national government.

In spite of politics and unrest, Montreal remains a city of great charm, of vivacity, and of gaiety, one of the most appealing in North America, as well as one of unquestioned modernity in its physical appearance and way of life. Thus it was chosen as the site of the International World Exposition in 1967—Expo 67. No great objections were raised when the event was perpetuated for the public each summer under the title Man and His World —Terre des Hommes in French.

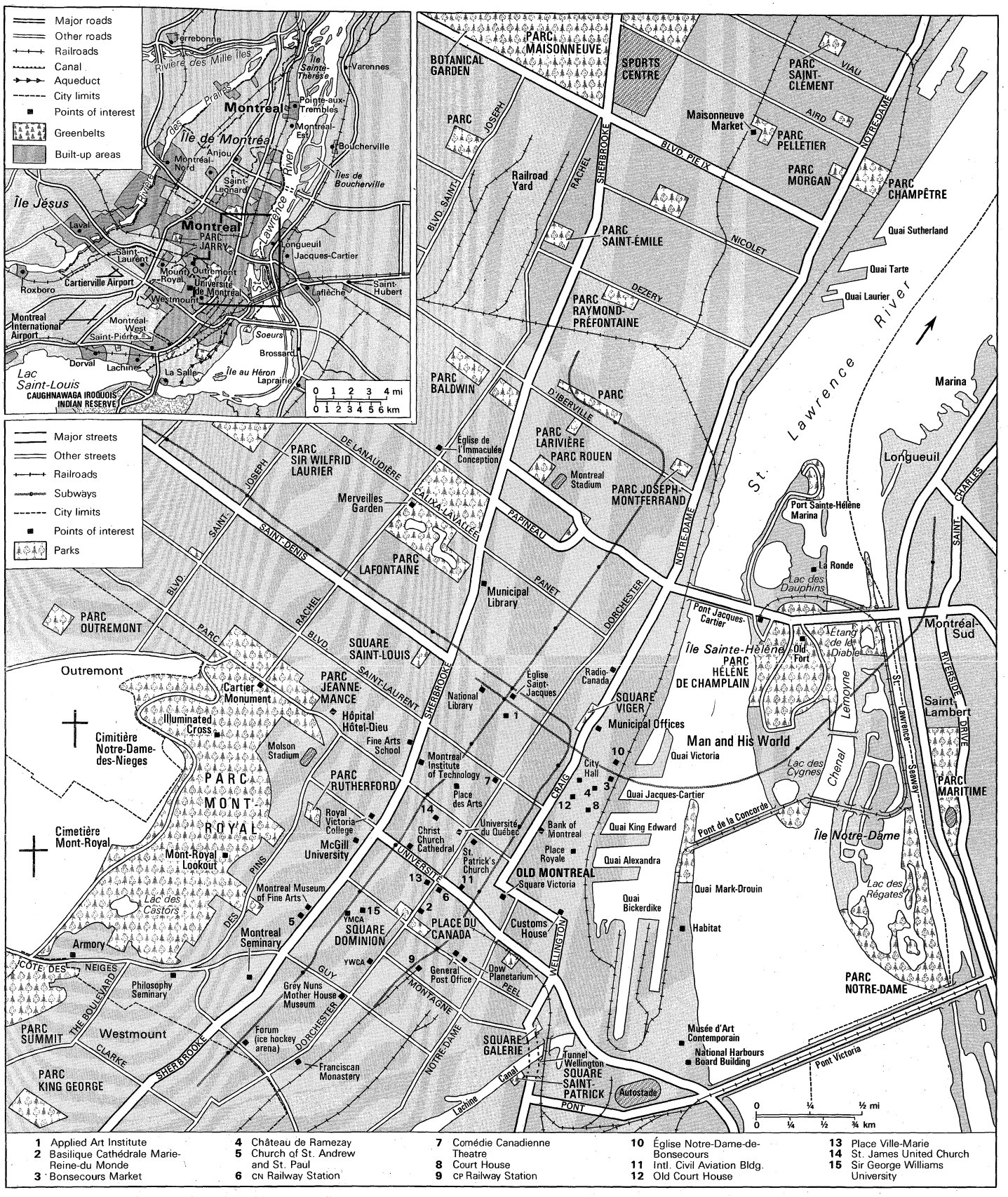
The history of Montreal. *The early settlement.* The site of Montreal was called Hochelaga by the Huron Indians when Jacques Cartier, a French navigator and the discoverer of Canada, visited it in 1535–36 on his second voyage to the New World. More than 1,000 Indians welcomed him on the slope of a mountain that he named Mont Réal, or Mont-Royal. More than half a century elapsed before other Frenchmen returned, this time with Samuel de Champlain, the founder of Quebec city. Hochelaga, it seems, had disappeared, replaced on the shores of the St. Lawrence by a settlement that Champlain called Place Royale.

The founding fathers

It was not until May 6, 1642, that Paul de Chomedey, sieur de Maisonneuve, founded today's Montreal. He built dwellings, a chapel, a hospital, and other structures, protecting the settlement against Indian attack with a stockade. He named the aggregate Ville-Marie.

The community was granted its first civic charter by King Louis XIV in 1644, and Chomedey became its first governor. The first hospital, the Hôtel-Dieu, was founded in 1645 by Jeanne Mance and the first school for girls in 1653 by Marguerite Bourgeoys. Almost immediately a society of priests, Les Messieurs de Saint-Sulpice, took charge of education for boys.

The real development of Montreal began during the first half of the 18th century. Land grants were made, and farming was developed outside the original fortifications. Colonization was initiated under the French seigniorial system, in which a landowner leased portions of his holdings to numerous farming families. For long, Montreal was a base for explorers and traders, and by the end of the 18th century outlying settlements—Saint-Henri and Lachine to the west and Longue-Pointe and Pointe-aux-



1 Applied Art Institute
2 Basilique Cathédrale Marie-Reine-du Monde
3 Bonsecours Market
4 Château de Ramezay
5 Church of St. Andrew and St. Paul
6 CN Railway Station
7 Comédie Canadienne Theatre
8 Court House
9 CP Railway Station
10 Église Notre-Dame-de-Bonsecours
11 Intl. Civil Aviation Bldg.
12 Old Court House
13 Place Ville-Marie
14 St. James United Church
15 Sir George Williams University

Central Montreal and (inset) its metropolitan area.

Trembles to the east—had taken roots, later to become part of the city or of the Montreal Urban Community.

By 1672 the population of Montreal had reached 1,200, but it did not obtain city status for another 120 years. The city surrendered peacefully to British forces in 1760 and, with all of New France, became part of the British North American empire in 1763. In November 1775, Montreal was occupied by American Revolutionary forces, who retreated in the spring following the abortive siege of the city of Quebec by Benedict Arnold and thus failed to secure Canada for the new United States. In 1796, Canada's first public library was opened in Montreal, and in

Central Montreal, looking eastward down Dorchester Boulevard with the Jacques-Cartier Bridge in the background.
By courtesy of the city of Montreal

the following year daily postal service was established between Montreal and the United States.

Beginnings of the modern city. In 1809, John Molson —entrepreneur, brewer, banker, carrier, etc.—linked Montreal and Quebec by water with the first Canadian steamship. The Bank of Montreal, the country's first, was founded in 1817, and the Lachine Canal, forerunner of the St. Lawrence Seaway, was started in 1821. In 1825, Molson, the "Montrealer *par excellence*," provided his city with a splendid theatre, and gas lighting appeared by 1838. A Board of Trade was founded in 1822, and three years later Montreal became the capital of Canada until 1849. On April 25 a mob put fire to the Parliament building, possibly on the ground that it had lost its vocation. The year 1847 saw telegraph links with the cities of Quebec and New York; 1853 a shipping service between Montreal, Liverpool, and the Continent; 1856 a railroad to Toronto; 1858 a transatlantic cable to Europe; and 1861 the city's first horse-drawn tramways.

Sequence of civic firsts

Fires destroyed hundreds of buildings in the early 1850s, and an economic slump provoked numerous bankruptcies in 1857. The present Confederation of Canada was proclaimed in 1867, and ten years later the city had its first labour strike and first telephone conversation with Quebec; it had its first electric lighting in 1882, electric tramways in 1892, and the first automobile along its streets and moviehouses along its sidewalks in 1903.

The 20th century. By 1900, Montreal's population reached 270,000, and it began to annex several cities, towns, and villages on its outskirts. It purchased St. Helen's Island (Île Sainte-Hélène) in 1908, the site, with two neighbouring man-made islands, of Expo 67. Montreal's famous ice-hockey teams, the Canadiens was founded in 1909, and by the early 1970s it had won more world championships than any other professional team. In 1922 several mergers gave birth to the Canadian National Railways Company (CNR), which, like the Canadian Pacific in 1881, established its head office in Montreal.

The world wars gave impetus to the economic life of Montreal, as they did to most industrial centres of North America, and in January 1947 the United States Senate authorized the nation to enter into a joint venture with Canada for building the St. Lawrence Seaway. In 1959 the need for a Montreal Metropolitan Corporation was recognized by the Quebec provincial government (which, like others in Canada, has exclusive jurisdiction over municipalities). The next year saw the opening of Metropolitan Boulevard, a throughway encircling Montreal. In 1962 construction was started on Le Métro, the Montreal subway that was supervised by engineers from the Paris Métro and inaugurated six months before the opening of Expo 67. With the growing recognition of Montreal as a major world centre following this universally acclaimed exposition, it became the first non-United States city to be awarded a major-league-baseball franchise. Its players first took the field in 1969 as the Montreal Expos.

The contemporary city. *Physical plan.* The layout of Montreal has been affected throughout its history by the river and the natural rises of the terrain from it, including the slopes of Mont-Royal. These factors produced a general southwest–northeast pattern of growth, and since its incorporation the city has been divided by St. Lawrence (or Saint-Laurent) Street (or Boulevard) into western and eastern sections. Such designations as street, avenue, and boulevard are virtually interchangeable in Montreal, and the ensuing confusion is enhanced by the occurrence of the place-names of thoroughfares and other city and suburban sites in both French and English.

Modernization of the city

Since about 1958 the centre of Montreal has been transformed significantly, abetted by post-World War II prosperity, by preparations for Expo 67 and the exposition itself, and by an administration intent upon grand designs. Built with the aid of U.S. capital and developers, Place Ville-Marie comprises a 45-story cruciform building and many underground shops, restaurants, and theatres, linked to which are nearby skyscrapers with similar underground complexes. Together, this new downtown provides the Canadian metropolis with an answer to New York City's Rockefeller Center, with an underground commercial, culinary, and artistic life among the most advanced in the Western Hemisphere. Similar change has spread throughout the city, often obliterating historic landmarks, but the city has preserved the historic centre known as Old Montreal. There, seekers after nostalgia stroll amid reminiscences of the past that are in striking contrast to the city's overall momentum into the vanguard of urban modernity.

The people of Montreal. The 1871 population of Montreal proper—about 133,000—increased some tenfold in the following century, but the pace of metropolitan planning in the early 1970s promised ongoing growth in addition to annexation of adjacent areas. Immigration from abroad practically ceased during World War I, but a steady flow of people continued from other parts of Canada and from the U.S. Although the birth rate among Canadians of French descent dropped markedly, immigration from the Continent reduced the percentage of Montrealers of British descent after World War II. French-speaking citizens currently account for more than 60 percent of the population, with the English proportion increasingly eroded by Jews, Italians, Germans, Greeks, central Europeans, and others. Religious affiliations generally follow ethnic traditions, with Roman Catholicism by far the dominant faith.

Social and political tensions

On the other hand, newcomers to Montreal quickly learn that English is the more practical tongue, for in Montreal as in most of Canada—except Quebec city and smaller centres of Quebec Province—English is the primary language of commerce and industry. The economic upper classes are mostly old Montreal-bred English-speaking families, with a sprinkling of French and others. The middle classes are more mixed, whereas the lower economic stratum continues to be made up mainly of French- and Irish-Canadians and of recent immigrants. Thousands of blacks have immigrated from the U.S. in recent decades, most of them settling in the lower part of Montreal-Ouest. The instabilities of Montreal and of Quebec Province as a whole are largely the result of the continuing sociolinguistic separateness of and economic disparity between the two major ethnic groups. Majority political power has been achieved by the French community, but equivalent weight in other areas has been slower to arrive.

The city's economy. Montreal is a corporate-headquarters city, especially for the largest Canadian banks, railroad lines, and insurance companies, as well as for the International Civil Aviation Organization, the trade association of the world's major airlines. Its importance as a shipping and an industrial centre often is overlooked, however. The centuries-long colonial fur trade that penetrated the new continent to its westernmost territories formed the city's earliest commercial ventures, while soapmaking, brewing and distilling (among John Molson's main civic legacies), and wood and leather fabrication are long-time Montreal industries. Innumerable other products of a modern manufacturing economy emanate from the city's factories, along with millions of gallons of processed petroleum each day from Montreal East refineries.

Shopping areas abound throughout the city, in the more remote residential sections as well as throughout the supermodern underground city of streets and shops that has aided Montrealers to carry on in spite of the five to ten feet of snow dumped on the city each winter. This deluge of snow from November through April is among the most significant factors in Montreal's life, costing the city an average of $5,000,000 to $8,000,000 annually to remove it from the approximately 1,000 miles of streets. In early March 1971 the entire city was paralyzed, forcing police to use snowmobiles throughout the entire city and Les Canadiens to postpone a game for the first time in their history.

Municipal and regional administration. From the present city hall, sixth in the city's history. Montreal is governed by a mayor and a 52-member council, all elected for four years, and a six-man executive committee selected by the council. A Montreal Urban Community replaced the Montreal Metropolitan Corporation in 1969. Its responsibilities, for the whole of Montreal Island and Bizard Island, to the north, includes assessment and tax collection, traffic control, water and sewage services, police and fire protection, antipollution activities, and the like. The Urban Community is governed by a general council of 82 members (54 for Montreal proper, 28 for the suburbs) and an executive committee of 12 members (7 from Montreal, 5 from the suburbs).

The entire metropolitan area, as defined by the Canadian Bureau of Statistics and reaching beyond the Urban Community, embraced 23 cities (including Montreal), 34 towns, nine villages and parishes, and one Indian reserve. The changes in the region's economic structure and increasing business and white-collar employment, intensified by the growing modernization of the city, have brought about some of the main administrative problems in recent years. The French-speaking majority, most of whom were relegated to blue-collar jobs for decades, reacted to this historical pattern with growing hostility.

Public transportation. The perennial grumbling of Montrealers about municipal transportation is more an exercise of democratic rights than a reflection of reality. Compared with other large cities, bus and Métro lines allow easy movement throughout the area at relatively low cost. The 400 Métro cars serve three lines running a total of 21 miles under the city and to Longueuil on the south shore of the St. Lawrence. Each of the 22 stations is different in architectural design and artistic decor. Public transportation dates from 1847, and by 1868 buses mounted on sleighs replaced rail cars during the winter. By 1894 the entire system had been electrified and the last horsecars withdrawn from service.

The modern Métro

Educational and health services. As throughout Quebec, a dual school system for Roman Catholic and Protestant students is supported from the public treasury. The language of instruction—French or English—rigidly follows the religious division.

Montreal is probably the outstanding city of Canada in terms of higher education. McGill University (founded 1821) and Sir George Williams University (1929) offer mainly English-language instruction, whereas the Université de Montréal (1876) and the Montreal branch of the Université du Québec (1968) serve the French-speaking population. The city has some 30 general hospitals, and there are faculties of medicine in McGill and Montréal universities.

Cultural life and institutions. With its Place des Arts, completed in the 1960s, its museums, public libraries, art galleries, bookshops in most European languages, symphony orchestra, publishing houses, theatre companies, and free public lectures at the universities, Montreal must be accounted a major culture centre.

The Place des Arts is a complex of concert and theatre halls in downtown Montreal. Art instruction is given, among other places, at the Museum of Fine Arts. Besides a conservatory of music, faculties or schools of music offer instruction at the universities. The Man and His World exhibition, the site of Expo 67, has a museum of contemporary art.

Some 40 years ago only three real bookshops in French and as many in English existed, but today they are found everywhere, in all districts and shopping centres; and "new Canadians," as recent immigrants are called, can buy books, reviews, and magazines in their native languages. The Civic Library has several branches, and more than 25 special libraries are scattered throughout the city. Publishing houses, English and French, are prospering.

In addition to being readers of books, Montrealers are great readers of newspapers. On weekdays they may choose from among four morning French-language newspapers and one in English; in the afternoon there is one in each language. On Sundays they have a choice among four in French, one in English, and several special-feature journals. Foreign newspapers, reviews, and magazines—mostly from the U.S., France, and Britain—are widely available, in addition to some 60 weekly or monthly scandal and gossip sheets. Montreal has about 20 television and radio stations, besides the headquarters of the French-language network of the Canadian Broadcasting Corporation—Radio-Canada in French.

In addition, Montrealers find time for sports, both as spectators and as practitioners. Hockey and baseball are foremost, but other indoor and outdoor sports have many adherents. In winter the slopes of the Mont-Royal are covered with women, men, and children. Montreal has also two very popular racetracks, and golf courses sur-

round the city. In 1976, Montreal was to be the site of the international Summer Olympic Games.

BIBLIOGRAPHY. Firsthand accounts are the narrative journals of JACQUES CARTIER and CHAMPLAIN, the Jesuit *Relations* and the *Histoire de Montréal* of DOLLIER DE CASSON (1672). Standard works include: W.H. ATHERTON, *Montreal*, 3 vol. (1914); STEPHEN LEACOCK, *Leacock's Montreal*, rev. ed. by J. CULLITON (1963); and K. JENKINS, *Montreal: Island City of the St. Lawrence* (1966). JOHN IRWIN COOPER, *Montreal: A Brief History* (1969), is a good summary although some interesting details are omitted. Also of interest is the article on Montreal in the *Encyclopedia Canadiana* (1957; rev. 2nd ed., 1967). The numerous publications of the MONTREAL ECONOMIC RESEARCH BUREAU, including the annual *Abridged History of Montreal*, are both informative and reliable.

(C.Fe.)

Moon

The Moon, like the Earth and the other terrestrial planets (Mercury, Venus, and Mars), has a solid body, nearly spherical in shape. The Moon's radius is about 1,738 kilometres (1,080 miles), only 0.2725 that of the equatorial radius of the Earth; this difference, with the fact that the Moon's mass is only $\frac{1}{81.3}$ of the Earth's mass, is largely responsible for the difference between the atmospheres of the two bodies. The Moon is a natural satellite of the Earth, but it is sometimes described as a planet, because it is large compared to most known, natural, planetary satellites and, because of its size, has many of the properties of a small planet such as Mercury. The escape velocity of a particle from the surface of the Moon is about 0.213 that from the Earth, and, if past temperature conditions were much like present ones, the gas molecules and atoms of any early lunar atmosphere would have slowly escaped in a time short compared with the Moon's lifetime (see below).

The Moon's mean density is about 3.34 grams per cubic centimetre, close to the density of the Earth's mantle; but studies of rocks returned from the Moon's surface indicate that the Moon's interior composition is not the same as that of the Earth's mantle and that in the Moon there is only a very small (possibly no) core. Many of the rock samples retrieved from the large, dark, relatively smooth areas of the face of the Moon called the maria are best described as basalts, but they are quite uncharacteristic of any known terrestrial rock or meteorite. They are about the colour of graphite—charcoal gray.

The lunar rocks differ from those on Earth

The whole Moon reflects less than one-tenth of the light that falls on it. The low reflecting power of the Moon had been known for a long time from estimates of its overall albedo. "Albedo" is a Latin word meaning "whiteness." While it is relatively easy to measure the relative brightness, or reflecting power (normal albedo), of flat surfaces illuminated and viewed normally in the laboratory, it is more difficult to assess the overall Bond albedo of a spherical planet. The brightness of the Moon varies through its phases because of the roughness of its surface and the resulting variable amount of shadow. The Full Moon, when the Sun, Earth, and Moon are in the same line, is ten times as bright as the Moon at First Quarter, the position of the satellite when it has completed one-quarter of its orbit around the Earth. When ¾ of the orbit is completed, the Moon is said to be in the Last Quarter.

This article is divided into the following sections:

I. The motion of the Moon
  The apparent motion of the Moon
    The month
    Optical and physical librations
  The actual motion of the Moon
    Pre-Newtonian history
    Analytic theory
    Numerical lunar theory
    Interpretation of small irregularities in the motion
II. The mass and gravitational field of the Moon
  Underlying theory
  Discovery of the lunar mascons
III. The physical nature of the Moon: selenology
  Observations from the Earth and from space vehicles
    Limits of resolution
    Small-scale details
    Results from manned lunar landings
    Data from closeup photographs
  Types of lunar feature
    Lunar craters
    Lunar lineaments
    Temporary or transient changes on the Moon
    The photometric function
    Lunar polarization studies
    Infrared (heat) studies
    Analysis of returned radar signals
    Physical experiments from orbiting spacecraft
  Theories of origin of the surface features
    Volcanic theory
    Impact theory
IV. The origin and evolution of the Moon
  The evolution of the Moon's orbit
  Evidence from the composition and physical properties of the Moon
    Chemical analysis of lunar samples
    Formation of the maria and highlands
    The thermal history of the Moon
    Correlation of physical and chemical data
  The "circumterrestrial swarm" theory

I. The motion of the Moon

THE APPARENT MOTION OF THE MOON

The lunar orbit is nearly circular, with the Earth as centre, the average eccentricity being only 0.0549, but it is more accurately described as an ellipse focussed on the Earth. In relation to the Sun, however, the Moon sweeps in a sinuous path, because, as it orbits the Earth, the centre of gravity of the Earth–Moon system (considered as a double planet) travels in an elliptical orbit around the Sun.

The month. The Moon's motion around the Earth may be judged by reference to the Sun or to the stars. The Sun appears to move each year across a background of fixed stars simply because all Earth-based observers are moving about the Sun once every 365¼ days, approximately. The Moon also moves across the sky in front of the remote stellar background, but its motion appears to be faster, because it completes its orbit of the Earth in a fraction of a year. A celestial meridian is an imaginary line in the celestial sphere, fixed relative to the stars, similar to a meridian denoting longitude on Earth. The Moon will be observed to cross any particular celestial meridian every 27.32166 days, which is the sidereal period of the Moon, or sidereal month. An observer on Earth sees only one-half of the lunar globe, and, because the Moon also rotates about its axis in 27.32166 days, this half is always approximately the same. Its axial rotation and sidereal period are said to be synchronous.

By the time the Moon has completed one revolution of the Earth relative to the stars, the Earth–Moon system has itself moved a fractional distance along its orbit around the Sun. Therefore, the time between successive Full Moons or successive New Moons (called syzygies) is a little longer than the time between successive lunar crossings of the imaginary celestial meridian. The period between one Full Moon and the next Full Moon, called the synodic period, synodic month, or lunar month, is 29.5306 days. The synodic period is measured using the Sun–Earth–Moon alignment as reference.

Direction of revolution

Viewed from the north, the Moon revolves in counterclockwise fashion around the Earth, as does the Earth around the Sun. The senses of rotation of the Earth and the Moon on their axes are also the same—counterclockwise if viewed by an observer in the northern celestial hemisphere. Because the Earth spins from west to east in a time that is short compared with the synodic month, the Moon, like the Sun and stars, rises in the east and sets in the west, even though its motion relative to the stars is from west to east.

If the orbital plane of the Moon were in the ecliptic (the plane of the Earth's orbit around the Sun), there would be a lunar eclipse (caused by the Earth's shadow) at the time of every Full Moon. The lunar orbit is inclined at 5°09′ to the ecliptic, and the ascending and descending nodes of the Moon's orbit—the points at which the Moon crosses the ecliptic into the northern and southern celestial hemispheres, respectively—regress (move clockwise

if seen from the north) to complete a whole revolution after 18.6 years. The interval of time between the Moon's successive crossings of a given node is referred to as the tropical month. The inclination of the Moon's orbit to the ecliptic causes lunar eclipses to be less frequent than they would be if the inclination were zero; the regression of the nodes, caused principally by the disturbing gravitational force of the Earth's equatorial bulge, leads to a repetition of eclipse phenomena every 18.6 years. This period is known as the saros (see ECLIPSE, OCCULTATION, AND TRANSIT for further details).

Earth–Moon distance

The mean Earth–Moon distance is 384,400 kilometres (about 238,900 miles), but, because of the elliptical shapes of the orbits of both the Moon and the Earth, the actual distance between them varies from 363,000 kilometres to 406,000 kilometres (226,000 to 252,000 miles). At its closest approach to the Earth in any one circuit, the Moon is said to be at perigee: at its most distant it is at apogee. The anomalistic month is the time interval between two consecutive perigee or apogee passages.

Optical and physical librations. The Moon rotates on its axis at a nearly uniform rate, but its orbital motion is much less uniform. Its orbit speed near perigee is greater, and that near apogee less, than its mean orbital speed of 1.03 kilometres (0.64 mile) per second. Hence, to a patient observer on the Earth, the Moon appears to oscillate as the orbit speed loses or gains on the constant rotation and shows more or less of its eastern or western limbs (apparent edges of the disk) at different positions in its orbit. When the Moon is north of the ecliptic, more of the lunar southern hemisphere can be seen; when to the south, more of the northern polar regions of the Moon. Furthermore, the observer can move from pole to pole, or from east to west, on the Earth, adding a small amount to the total, about 57 percent, of lunar surface visible from the Earth. The apparent oscillations of the Moon are known as optical librations and can amount to Moon-centred angular displacement of about 6.8 degrees in latitude and about 8° in longitude.

The Moon also has a very small but real oscillation, known as the physical libration, about its centre of mass. The physical libration amounts only to a few hundredths of a selenocentric (*i.e.*, moon-centred) degree and arises largely because the distribution of mass in the Moon is not quite that of spherical symmetry. There is more mass in a unit section around the Earthward-directed axis of the Moon than there is along either of the other principal lunar axes in the plane of the sky. This is why one face of the Moon persists in pointing toward the Earth: the gravitational pull of the Earth on the excess mass tends to restore the Moon to its mean orientation with respect to the Earth following any small movement away from that orientation.

The harvest-Moon phenomenon

The so-called harvest Moon is the Full Moon occurring on or nearest to September 23, the autumnal equinox. At that time, the apparent motion of the Moon is such that a terrestrial observer in the Northern Hemisphere may witness a minimum delay in the time of the Moon's rising on successive evenings, and, if the Moon is full, the phenomenon can be notable. (G.Fi.)

THE ACTUAL MOTION OF THE MOON

The mathematical treatment of the Moon's actual geocentric motion is called the lunar theory. Customary astronomical terminology distinguishes between an analytic procedure, called a theory of general perturbations, and a numerical method, called a theory of special perturbations. Because of the Moon's nearness to Earth and of its rapid angular motion, the lunar theory has long played a major role in formulating and testing dynamical theory.

Pre-Newtonian history. In ancient times, the Moon's phenomena were of great religious importance. Thus, considerable attention was paid to its motions; and the observations made were of an exactitude unsought in recording terrestrial phenomena. The Babylonians devised numerical methods for predicting eclipses and heliacal risings. The Greek astronomer Hipparchus in the 2nd century BC discovered the eccentricity of the lunar orbit, the motions of perigee and node, and the inclination of the orbital plane to the ecliptic. He explained the motion as being uniform in a circular orbit with the Earth placed eccentrically, and this was adequate to represent the observations then available. The numerical values of the parameters were derived from eclipse observations, some dating to Chaldean times. Although his results were superior to previous attempts, the restriction to eclipse observations led Hipparchus to an incorrect value of eccentricity, through inability to distinguish the major inequality (departure from uniform motion, true value 6.29 degrees), due to eccentricity, from the greatest solar perturbation. The latter inequality, called evection (1 ° 27), was discovered by the Greek astronomer Ptolemy (flourished mid-2nd century AD). Little significant improvement in the knowledge of the lunar motion was secured until the 16th-century Danish astronomer Tycho Brahe discovered the other great inequality caused by solar attraction, called the variation. This term, whose amplitude is 0 ° 66, vanishes when the Moon is new, full, or in quadrature. The explanation of these inequalities by adequate treatment of the Moon's motion required the discovery of universal gravitation.

Analytic theory. The analytic treatment of the lunar motion is carried out in two stages. In what is termed the "main problem," it is assumed that the Sun, Earth, and Moon behave as point masses and that the barycentre, or centre of gravity, of the Earth–Moon system describes a fixed ellipse around the Sun. After the solution of the main problem, the remaining effects are introduced as small corrections to the main solution. These effects include the perturbing forces of the planets acting on both Earth and Moon, the forces caused by the irregular shapes of Earth and Moon, corrections due to use of a non-inertial coordinate system, and relativistic corrections.

The solution of the main problem gives the coordinates as harmonic series in which the arguments are linear combinations of the time-variant angles l, l', F, and D. The periods of revolution of the arguments l, F, and D are the anomalistic, draconic, and synodic months. That for l' is the anomalistic year in the Earth's heliocentric motion. The series coefficients are power series in these parameters: e (= 0.055) the eccentricity and γ (= 0.045) the sine of half the inclination of the lunar orbit, m (= 0.075) the ratio of average angular velocities of Earth and Moon, α (= 0.0025) the ratio of average geocentric distances of Moon and Sun, and e' (= 0.017) the eccentricity of the Earth's orbit. The values of these parameters are determined from observation. The perturbations developed in the second stage of the analysis introduce the planetary orbit periods and the motion of the Moon's node into the arguments, as well as additional secular effects in the harmonic terms of the solution to the main problem.

A purely algebraic method is elegant and permits rigorous use of improved parameter values but is very laborious. It is more common to adopt numerical values for one or more of the parameters (particularly m) from the outset, relying on algebraic approximations to the effects of changes in these parameters.

Isaac Newton based his theory of gravitation largely on the comparison of the motion of Moon with that of bodies falling freely near Earth. He succeeded in proving that the principal periodic inequalities, as well as the motions of node and perigee, were caused by the solar attraction. Although his work produced new inequalities that were not previously deduced from observations, the motion of perigee was only about half the observed value. Extant manuscripts indicate that this disparity was later reduced to about 8 percent. Newton's theory was published in geometrical form in his *Principia*, but the method of fluxions (developed later into the integral calculus) was probably used in the actual derivation. No substantial advance was made beyond Newton's work until 60 years later, and further improvements followed.

In the 19th century, the outstanding lunar theories were those obtained by Peter Andreas Hansen, by Charles Eugène Delaunay, and by George William Hill and Ernest William Brown, many such advances being now known by

the names of the distinguished scientists responsible. Hansen's theory was the most nearly numerical of the analytic theories and was of greater accuracy than any other before Brown; it was used internationally from 1862 to 1922. Corrections by Simon Newcomb were applied after 1883. Hansen's work has served as the basis for orbit computations for artificial satellites.

The most complete algebraic development of the lunar theory until the current time was the monumental work of Delaunay, which was never adequate for comparison with observations but which has been valuable in theoretical studies, the extension of other lunar theories, and the discussion of the orbits of natural satellites of other planets.

In the late 19th century, Hill developed methods with which results of given accuracy could be achieved with much reduced labour. These methods have been effective in the discussion of the general problem of three bodies (see CELESTIAL MECHANICS) as well as in the practical solution of the lunar problem. The two principal features were (1) the introduction of the value of the ratio of the mean motions of the Earth and Moon from the outset to avoid convergence difficulties and (2) the use of a rotating rectangular coordinate system. Hill developed a rather elegant method of determining the motion of perigee that was also applied by John Couch Adams to the motion of the node. Hill's methods were used by Brown to develop the lunar theory that has been generally used to compute the coordinates of the Moon since 1923, although the theoretical motions of node and perigee were still unsatisfactory and were replaced by the observed values.

Brown's series contain about 1,500 terms, five times as many as in Hansen's theory. The evaluation of these series is very laborious, and Brown, like some of his predecessors, prepared tables for the simplification of this work. They filled 650 quarto pages, and one man working alone could extract the coordinates just fast enough to keep up with the real Moon. The advent of the digital electronic computer has relieved this burden and also made feasible the direct evaluation of the series. Direct calculations of the coordinates for the years 1952–71 were compiled by the U.S. astronomer Wallace J. Eckert and coworkers, and these have been contained in the national ephemeris volumes since 1960. A correction due to the variability of the Earth's rotation rate was introduced in this computation.

When Brown's solution was completed, the accuracy was more than adequate for comparison with the best observations, but subsequent improvements in observing techniques have surpassed the accuracy of the theory. Eckert and his collaborators determined corrections to Brown's theory, but studies during the late 1960s have indicated that further extension to higher accuracy may be unfruitful. An entirely new solution of the main problem that contains over 6,000 terms, including those as small as 0″.00002 (about four centimetres), has been completed. It had not yet been published at the time of Eckert's death in 1971.

Use of computers

Recent innovations in the usage of electronic computers have brought a resurgence of interest in the purely algebraic approach abandoned after Delaunay. The technique of non-numeric manipulation has been applied to the complete solution of the main problem, particularly by André Deprit, who has subsequently attacked the problems of the second stage in the solution process. Deprit's theory contains not only the coordinates but also the necessary functions for determining or applying corrections to the lunar orbit parameters, an important innovation.

Numerical lunar theory. Although strictly numerical theories have long been constructed for some solar system objects, this was impractical for the lunar motion before the development of the electronic computer. The nearest approach to a special theory was the attempt by the English Astronomer Royal George Biddell Airy in 1883. The attempt was unsuccessful, and the idea was abandoned.

In recent years, the necessities of space exploration have been primarily responsible for the application of numerical integration of the equations of lunar motion, as the available general theories are now inadequate for high-precision usage. The first such widely available ephemeris in practical use was constructed by J. Derral Mulholland and collaborators, who pioneered this method. The procedure has the advantage of arbitrarily high precision but the disadvantage of uncorrectability; changes in parameter values require a completely new integration to be performed.

Interpretation of small irregularities in the motion. Study of the lunar motion not only provided Newton with the key to universal gravitation but has more recently shed light on the dynamics of Earth's rotation; shown up flaws in the theory of the Earth's heliocentric motion; and, in conjunction with new observing methods of unprecedented accuracy, promises in the near future to provide information on the interior structure of Earth and Moon and a possible test between relativity theories.

In the 19th century gravitational theory could not account for all of the observed motion. For some time it was thought that this was due to some oversight in the theory, but it was later shown to be due to fluctuations in the Earth's rotation. The magnitude of the effect on the observed motion of a celestial object is greater for more rapidly moving bodies, so it first appeared in the lunar motion. From 1954, the lunar motion was used to correct the time determined by the rotation of the Earth, although by the 1970s this practice was being replaced by the use of atomic clocks.

Much smaller discrepancies in the lunar theory are now observable than was heretofore possible because of the introduction of the two new observing methods of grazing occultation and laser ranging. A grazing occultation occurs when the Moon occults a star at the edge of its disk, so that the star disappears behind lunar mountains but reappears in the valleys between. This method provides latitude measures the accuracy of which is limited only by the knowledge of the star's position and thus comparable with the accuracy of photoelectric longitude measures from normal occulations. These data are being applied to improve the parameters of the lunar orbit, the location of the equinox, knowledge of the shape of the Moon, and of the effects of tidal friction in the Earth.

The Apollo 11 crew deployed a reflector array on the lunar surface (see also below), which permits direct distance measures from a laser-equipped telescope on Earth to a fixed point on the Moon. Similar arrays were later landed by Luna 17, Apollo 14, and Apollo 15. A short laser light-pulse is emitted through the telescope and reflected off the lunar target back to Earth, where the light travel-time is measured with high precision. Such measurements, first made at Lick Observatory, are now being made regularly from the McDonald Observatory with a resolution of ±30 centimetres (12 inches); higher precision is achievable. Other observatories are expected to establish regular programs. After these data are used to improve the lunar orbit parameters, qualitatively new irregularities of still smaller size will be observable, including the effects of solid-body tides in both Earth and Moon; the rotational irregularities of both bodies (symptomatic of peculiarities in internal structure); and tectonic motions of the Earth's crust. In addition, just as the lunar motion provided a decisive test for Newtonian dynamics, these new observations will provide a test between alternative theories of realistic dynamics, since the Brans-Dicke scalar-tensor theory and the Einstein theory predict lunar motions that differ by amounts of the order of one metre (about three feet). (J.D.Mu.)

II. The mass and gravitational field of the Moon

Several space probes had flown past or made physical contact with the Moon before the Soviet probe Luna 10 in April 1966 became the first artificial lunar satellite. Successive orbital rendezvous with the Moon were then made by several space probes, the orbits of which have yielded information about the Moon's mass and gravitational field.

UNDERLYING THEORY

The Moon is spherical to within a fifth of 1 percent of its radius, and it is a reasonable assumption that it is also essentially spherically symmetrical. The gravitational potential (indicated V) at a distance r from the centre of a spherically symmetrical planet of mass M is equal to that due to a similar mass at the centre of the planet; that is, if G denotes the gravitational constant, $V = -GM/r$. The gravitational force (F) is given by the negative gradient of V, $-\partial V/\partial r$ (rate of change of potential with distance), and which gives $F = GM/r^2$. Gravitational attraction is a force that varies with the inverse square of the separation between two point masses. Now, the gravitational force exerted on a small unit mass by a planet of mass M is measured by multiplying the acceleration (g) due to the gravity of the planet by that unit mass. For points at a distance (r) greater than the radius (a) of the planet—that is, outside the surface—acceleration (g) is equal to the product of the mass M and the gravitational constant divided by the square of distance: $g = GM/r^2$. This formula for g shows that, as an object approaches the Moon, the radial acceleration of the object increases rapidly as r decreases to its final, minimum value equal to the radius of the Moon, and the final value of g, the acceleration that is caused by gravity at the surface of the Moon, is a function of the mass M and radius a of the Moon.

As lunar probes approached the Moon, radio frequency signals from them were received on Earth, and measurements of the final phases of the acceleration of the vehicles, as they fell freely, allowed improved estimates to be made of the gravitational force from the Moon. The assumption that the Moon is spherical is not, however, strictly correct, and the final value of the radius, r, at any particular place of impact may be less than or greater than the mean radius. Observations of the Moon-orbiting vehicles have shown that even the interior departs slightly from spherical symmetry. Thus, a single encounter of the type described cannot be used to determine the Moon's mass to a high degree of accuracy.

The Earth and the Moon describe elliptical paths around their common centre of gravity and the semi-major axes of the Earth's and Moon's ellipses around the centre of gravity of the Earth–Moon system are in the ratio of the respective masses of the Moon and the Earth. The mass of the Moon (7.35×10^{25} grams [about 8×10^{19} tons]) may be determined from knowledge of the distance (384,400 kilometres) of the Moon, the distance (4,670 kilometres) of the joint centre of gravity from the centre of the Earth, and the mass of the Earth (about 6×10^{27} grams). Determinations of the mass of the Moon have been based upon the trajectories of planetary space probes departing from the Earth–Moon system, beginning with the Mariner 2 vehicle in 1962.

Effects of gravitational irregularities on satellite orbits

For less than half its orbit period, while the satellite is above the far side of the Moon, the radio signals from a lunar satellite are, in general, lost. Information on complete orbits may nevertheless be built up from those signals that are monitored. The manner and extent in which observed orbits differ from ellipses are functions of the Moon's particular gravitational irregularities and the gravitational action of the other astronomical bodies (principally, the Earth) on the lunar satellite.

The motion of a satellite around the Moon is made more complicated because the Moon, in addition to presenting more mass along and around the Earth-directed axis than in either of the other, mutually perpendicular axes, has a noncircular equator. More precisely, if the Moon's moment of inertia about the Earth-pointing axis of inertia is A, that about the perpendicular axis, which is approximately in the plane of the orbit, is B, and that about the rotation axis is C, then A is less than B, and B is less than C. Although it is not possible to determine any of these moments of inertia directly, ratios such as $(C - A)/B$ have been derived from studies of physical librations. Studies of the motions of close lunar satellites have, in addition, yielded values of the functions $C/Ma^2 - (A + B)/2Ma^2$ and $(B - A)/Ma^2$, and, by combining the two methods, it is possible to derive the important ratio C/Ma^2. The value of this ratio is 0.4001 and the amount by which this differs from the value 0.4000, which corresponds to a sphere of uniform density, limits the nature and size of the Moon's core. If the Moon had an iron core, for instance, its radius would probably have to be less than one-fifth of the Moon's radius.

DISCOVERY OF THE LUNAR MASCONS

Formulas such as C/Ma^2 describe the bulk properties of a planet but give no information about the detailed distributions of mass in the planet. Methods of spherical harmonic analysis have been applied to investigate how the gravitational potential of a planet differs from that of a spherically symmetrical body. The potential may be expressed as a mathematical series of terms involving the position of an artificial satellite in its orbit and the constants describing the orbit. Low-order harmonics give terms in the series, and these are important in the case of features such as an equatorial bulge; progressively higher order harmonics show, in a close satellite's motion, progressively smaller scale gravitational irregularities. The more gravitationally complex a planet is, the more numerous the observations of satellite positions at given times have to be in order that the harmonic-series equations may be solved.

The points at which a satellite crosses the equatorial plane of the Moon are called the nodes. They rotate around the Moon's polar axis as a consequence of the low-order departures of the Moon from sphericity. A circumlunar satellite in an orbit inclined to the lunar equator so as to pass vertically over a predetermined, maximum latitude passes vertically, at some time during the course of many orbits, over all parts of the Moon on this selected parallel of latitude.

Space vehicles have been placed in orbits highly inclined to the Moon's equator and thus have been in a position to survey a large fraction of the Moon. Irregularities in the motion of one such satellite were noted in 1968: when it passed over certain of the maria, the inwardly directed radial acceleration had shown small differences that appear to be due to the presence of excess mass (gravitational anomalies) in the Moon. It was suggested that the anomalous accelerations were a consequence of concentrations of mass—mascons—below the points in the orbit at which the residual accelerations were found. A big mascon was associated with Mare Imbrium. Analyses of later lunar satellite orbits showed that there are mascons associated with the Maria Orientale, Serenitatis, Crisium, Smythii, Nectaris, and Humorum, and lesser ones with the crater Grimaldi, Sinus Aestuum, Mare Humboldtianum, an area north of Theophilus, and one east of Vendelnius. (Investigation of gravitational anomalies on the far side of the Moon is difficult.) It is also known that in one or two cases, notably that of Mare Orientale, which has a concentric-ring type of structure composed of alternate areas, or annuli, of lunabase (dark, mare-type material) and lunarite (brighter, highland-type material), the lunabase displays positive anomalies, while the highland, lunaritic zones display negative anomalies (weaker gravity).

A positive or negative gravitational anomaly can be produced in a number of ways. Geophysicists attempt to match the observed gravity profiles of parts of the Earth with those computed for particular models of subsurface structure and density contrasts. Independent geological or geophysical evidence of crustal structure can be of great assistance in this procedure. Unique models are even then only rarely determined. In the case of the Moon, even the topographic profile of a mascon site is not well-known. In the absence of much relevant information it is uncertain, for instance, whether mascons derive from very dense, deeply buried, condensed material or from near-surface accumulations of dense but more widely dispersed material.

Extensive basic lava flows in the prominent mascon area Mare Imbrium have been described (see below). All the positive gravity anomalies are associated with areas that are predominantly of mare-type material, the lunabase. (On the other hand, not all areas of lunabase display

anomalies.) It seems reasonable to conclude that lateral transport of mass has occurred in some places on the Moon. Processes that might have been involved are meteoric impact, the differentiation of material, the migration of magma (molten rock), and the eruption and flow of lava. An observer on Earth with normal, unaided eyesight can distinguish the separate, dark maria and the bright highland areas of the Moon. With a magnification of a few times it is possible to locate other features—for example, the bright ray craters Tycho (diameter 85 kilometres [53 miles]) and Copernicus (diameter 92 kilometres [57 miles]) at Full Moon or the Apennine Mountains (up to 4,000 metres [13,000 feet] high and several hundreds of kilometres in lateral extent) at First Quarter, when shadowing throws them into relief.

III. The physical nature of the Moon: selenology

OBSERVATIONS FROM THE EARTH AND FROM SPACE VEHICLES

Limits of resolution. Lunar craters of the order of 100 metres (300 feet) in diameter may be seen from the Earth under the best conditions using moderately sized telescopes, and, because, in practice, resolution depends on the shape of an object and on its contrast with the surroundings, the same instruments can be used to view lunar lineaments (linear features), such as rilles, considerably smaller than 100 metres in width, provided they are many times longer than 100 metres. The best Earth-based photographs of the Moon show some (the sharpest) craters as small as 500 metres in diameter. Photographs from space vehicles have resolved details in some cases as small as a metre. The first camera was put into an orbit passing behind the Moon (October 1959) by the U.S.S.R. Its useful photography was confined to some 70 percent of the Moon's averted hemisphere lit by a high Sun, and both the picture quality and the resolution were poor by the standards of Earth-based telescopic photography. Useful far-side photographs were obtained from another Soviet space vehicle, which completed the lunar far-side coverage in July 1965, and from a third, which provided higher resolution coverage of a limited portion of the far side. A number of ray craters (see below) were readily identifiable, but the most surprising observation was that most of the far side of the Moon contained no maria. Major crater chains, up to 1,000 kilometres (600 miles) long, were found.

The first far-side pictures

Small-scale details. The photography from a series of vehicles designed to crash-land on the Moon and relay television pictures to Earth prior to crashing was not successful until July 1964, when an American probe approached and crashed in a part of Mare Nubium. The area was renamed Mare Cognitum by the International Astronomical Union. For the first time, more and more small craters were seen as the vehicle effectively decreased the lunar-surface resolution to a limit of about 50 centimetres (20 inches). Craters of various shapes, rock blocks, and ridges 100 to 1,000 times smaller than anyone had seen before were recorded during the last minutes of descent. The Moon was found to be cratered more heavily than had been predicted, and the substantial excess of craters was attributed both to the mechanism of secondary cratering—that is, crater formation by objects ejected from a primary meteoric (or volcanic) cratering event—and to endogenic cratering—that is, crater formation by internal processes such as outgassing or collapse. A few dimple-shaped craters, probably the result of drainage of the Moon's veneer of fragmental rock into subsurface fractures, were discovered. Many more were later found in Mare Tranquillitatis and in the 110-kilometre- (70-mile-) diameter structure Alphonsus. None of these craters has the bowl- or saucer-shaped profile characteristic of a fresh, sharply sculptured (eumorphic) explosion crater or of an older, denuded (submorphic) explosion crater. On the other hand, the majority of the sub-kilometre-sized craters could well be explosion craters produced either by high-speed impact or by violent outgassing (explosive volcanism).

Mare Cognitum is traversed by faint rays of lighter hued material. The lunar rays that emanate from many craters are generally thought to be composed of small craters and blocks of rock. A certain proportion of the craters set in rays may be secondary-impact craters. In one cluster of craters, chains, or linear arrays, of both sharply sculptured and older craters trend in a given, narrow range of azimuth. It is not known whether the two types of crater were formed as respective secondaries when hard and soft material landed to produce craters at the same time and at the same distance from some primary crater; or the eumorphic and submorphic craters were formed internally at widely different times, the latter being formed first and then undergoing denudation.

The principal erosive force on the Moon at the present time is probably that of meteoric impact. Primary meteoroids are increasingly the more numerous with diminution in size, so that micrometeoroids are probably the principal factor involved in the attrition of freshly formed surfaces on the Moon. The net direction of movement of material sprayed from a meteoric cratering event is downslope; and this downward migration of rock fragments tends to level the lunar terrain and to generate a ubiquitous veneer of particles that is variable in thickness from one locality to the next. This layer, known as the lunar regolith, is composed of fines, particles less than one centimetre and mostly much less than one millimetre in size, and of larger fragments of rock. The process of impact mixing of the lunar soil and the particles of the regolith is sometimes referred to as "impact gardening."

The lunar regolith

Some of the blocks of rocks photographed on the Moon appear to be partially buried in the regolith, and a first estimate of the strength of the lunar-surface materials was made from considerations of the depth of penetration of these blocks and the distances they had been thrown. A bearing strength of about one kilogram (2.2 pounds) per square centimetre found for this part of the Moon showed that it was strong enough to support a man clad in a spacesuit or any of the planned soft-landing craft. Following the first soft landing of a spacecraft, Luna 9, on the Moon, in February 1966, the Soviets determined the soil in Oceanus Procellarum to have a bearing strength of at least several kilograms per square centimetre.

It appears to have been confirmed that some of the small craters of the Moon are of internal origin, though it is still believed that the vast majority of these craters are of impact origin. Several chains of contiguous craters are generally thought to be secondary craters from the 85-kilometre- (53-mile-) diameter crater Theophilus, even though many of them run nearly parallel to tectonic lineaments such as linear ridges and troughs—a situation indicative, instead, of an endogenic origin of these craters. The blocks of rock, up to 100 metres in diameter, are commonly scattered in and around craters that were probably deep enough to cut through the friable regolith into cohesive bedrock. This bedrock would have been excavated and tossed, as blocks, from a crater during its formation, it was argued.

Many rilles (lunar trenches of lengths up to a few hundred kilometres and widths up to a few kilometres) have been photographed by space vehicles. Those recorded are particularly interesting because of the intimate connection between the rilles in the floor of the crater Alphonsus and dark halo craters that lie in the rilles. Exceptionally dark deposits encircling these craters have tended to obliterate underlying relief and, locally, fill the rilles. The craters themselves are elongated along the axis of a rille and are without raised rims—characteristics that point to internal origins. The associated dark deposits were probably erupted as sprays of rock and lava.

The rilles

In the next phase of lunar exploration, closeup television photographs of the landing feet of a soft-landing vehicle in June 1966 (within a 100-kilometre- [60-mile-] diameter ring structure, known as Flamsteed P, in Oceanus Procellarum) showed that the feet had penetrated a few centimetres into the ground and had disturbed the soil around them, giving it a darker appearance in the process. Stones and blocks of rock were found to vary considerably in appearance. Some appeared to be partly buried, their surfaces smoothed by erosion. Some ap-

peared to have been put on the surface and to have remained in their original, angular configurations. Some had different shapes and numbers of pores (vesicularity), structure (layering and differential denudation), or, possibly, composition (tonal differences and the presence of inclusions).

Similar results were found from other probes soft landed in Mare Tranquillitatis and in Sinus Medii in missions meant to examine possible landing areas and analyze the soil. One mission was sent 42° south of the lunar equator to examine, for the first time, part of the Moon's highland domain. It landed (January 1968) successfully within what was, by comparison with the area of the highlands around Tycho, a fairly smooth area. The television cameras, however, showed the terrain to be strewn with many more blocks of rock per unit area than had been found in the maria. The most revealing discovery concerned the composition of the rocks; the percent weights of elements in rocks were estimated, and the landing area was found to be similar in composition to the mare samples, with the exception that the iron-group elements (mass numbers 47 to 65) were less abundant in the highlands.

Data on rock composition

The elemental composition of the rocks at three sites approached that of some terrestrial basalts, types of dark igneous rock. The tested areas of the Moon were compositionally unlike chondrites—the most common kind of meteorite—and analyses have indicated that the Moon had been differentiated and heated. The theory that the Moon originated as a cold body and has never been heated suffered a serious blow, although the heating might have been the local result of major impacts. Later analyses of rocks returned to Earth provided further information on heating (see below).

Soviet Union and United States spacecraft both first entered circumlunar orbits in 1966. The emphasis in the Soviet series has been on physical measurements in the vicinity of and pertaining to the Moon; the prime purpose of the first U.S. orbiters was to provide photographs of several alternative astronaut-landing sites dispersed along the equatorial belt in the smoother, mare areas.

A camera in one such craft was able to survey in sequence the whole of the lunar surface. Some areas not recorded or not well-photographed by earlier cameras were later examined.

One consequence of all this lunar photography has been good coverage showing fine detail over most of the lunar globe. Topographic and geological maps of parts of the Moon's near side have been prepared with horizontal errors of about one kilometre. Detailed photographic and visual observations of the lunar surface were made in 1969 to evaluate the terrain characteristics during the close approaches of the lunar module.

Results from manned lunar landings. The first manned lunar landing was made in July 1969 in Mare Tranquillitatis. Some 22 kilograms (49 pounds) of lunar rocks were collected and returned to Earth for study.

Later manned landings were in Oceanus Procellarum, in the Fra Mauro highlands, and in Mare Imbrium between the Apennine Mountains and the Hadley Rille, each adding to the store of lunar rocks returned to Earth. Each crew left behind functioning observing instruments, including a seismometer, a device commonly used to measure earthquakes.

Moonquakes

Though the first seismometer failed after only a short operating time, later ones worked simultaneously and, in registering the vibrations generated by impacts and moonquakes, have provided valuable information about meteoroids that encounter the Moon and the slight, relative movements of rocks along lunar faults or other slip surfaces, as well as about the behaviour of lunar material at depth and the probable nature of the rocks at various depths and localities.

Missions have also landed a type of optical reflector on the Moon, capable of returning to Earth extremely accurately timed laser pulses for reception by large telescopes. Timing of these pulses on their journey to the Moon and back reveals very small changes in the Earth–Moon distance and can improve knowledge of the present rate of recession of the Moon from the Earth. The so-called secular acceleration of the Moon provides information about the origin of the Moon, discussed below. The accuracy of measurement of the Moon's distance by laser measurements is a small fraction of a metre (there is approximately one nanosecond in two-way time delay resulting from about 10 minutes of observation; this is equivalent to about 15 centimetres). Given a suitably wide network of laser reflectors on the Moon, additional studies on the Moon's figure and physical liberation can be performed. The precise shape of the Moon and the amount of its free libration also have a bearing on its past history.

Laser measurement of Earth–Moon distance

Other instruments left behind on the various manned landings provide continuing physical and geophysical data. Earlier measurements of the physical properties of the lunar environment using crash-landing and orbiting vehicles were valuably supplemented or completely superseded by the results drawn from the later missions.

Data from closeup photographs. Increasingly valuable photographs of the Moon were taken by the astronauts on succeeding missions, both from orbit and from the surface. Many of the rock samples were photographed in relation to their surroundings prior to removal from the Moon's surface, to assist in later interpretations (the broad setting from which a rock is drawn is important to an investigator in any comprehensive geological study).

The lunar photographs by no means record all possible bits of lunar surface, even down to a size as coarse as a kilometre, but vast amounts of unmeasured, unprocessed data are stored in existing photographs. From them, lunar craters have been measured for size, shape, and depth (from shadow measurements): crater counts have been used in estimates of the terrain roughness and in statistical studies bearing on the origin of the craters. The roughness of lunar rock blocks, estimated from photographs as well as from returned samples, has been related to theories of denudation (erosion) and aging. Lunar-surface processes, such as transport of the material and impact mixing, have been studied using photographs alone. Lava flows covering large areas of the maria have been detected and mapped using Earth-based infrared and ultraviolet photographs in combination with photographs taken from lunar orbit. The brightness contrasts displayed by different rock units have been measured from Full Moon photographs. Slopes have been assessed from shadows under different Sun angles. Topographic maps have been prepared. Texturally distinct geological units have been mapped and used to produce a lunar stratigraphic column of relative ages. Faults and other tectonic lineaments have been recognized on the photographs.

TYPES OF LUNAR FEATURE

Lunar craters. The ubiquitous nearly circular features on the Moon have all come to be known as craters; the traditional terms ring plain and walled plain describe, respectively, round and irregular lunar features with walls encompassing relatively smooth, level floors.

Forms of craters

Lunar craters, in the broad sense, may be classified, first, on the basis of their form. Many intermediate-sized craters, in particular, display mountain blocks or hills near their centres. The inner (and, to a smaller extent, the outer) walls of many craters are terraced, with interterrace benches.

Although in a lunar photograph (see Figure 1) the majority of the lips of crater rims (walls) appear to be approximately circular, it is common to find that, in detail, the actual shapes approximate better some polygon or an ellipse. A number of lunar craters have ellipticities greater than 30 percent. Some of the so-called craters have a spiral inner ringwall lower than the outer; there are craters with eccentrically placed inner craters and plateau craters that seem to have been filled with dark material. The impact or successful lunar landing sites are shown in Figure 1 and described in the Table.

Bowl-shaped craters with parts of the floor darker than the inner slopes are common among lunar craters less than 20 kilometres (12 miles) in diameter. They may be

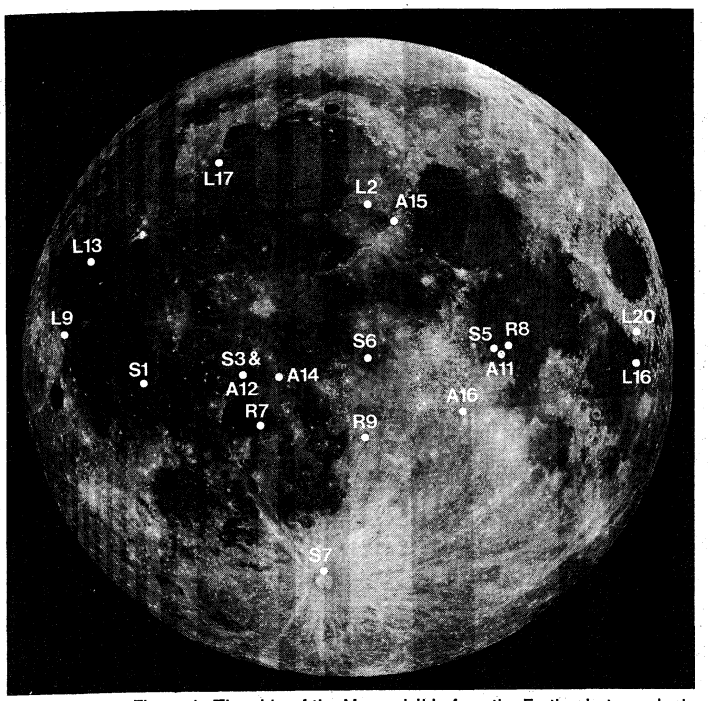


Figure 1: The side of the Moon visible from the Earth, photographed at Full Moon. The approximate points of impact are shown of selected U.S. and U.S.S.R. spacecraft of the Apollo (A), Ranger (R), Surveyor (S), and Luna (L) series. Three dark "seas," left to right, just above centre, are Mare Imbrium (Sea of Rains), Mare Serenitatis (Sea of Security) and Mare Tranquillitatis (Sea of Tranquillity).
Photograph by courtesy of the Lunar and Planetary Laboratory, University of Arizona

as much as three kilometres deep, with inner slopes of up to 50° and outer rims sloping at up to 20° near the lip. The shallowest "saucer" craters have the surfaces of their rims inclined to the horizontal at no more than a few degrees. Crater pits are those without raised rims. Funnel- or dimple-shaped crater pits with convex inner walls near their lips have been discovered in some of the photographs.

Impact or Landing Sites of Successful Lunar Spacecraft

name and launch date	remarks
Luna 2 Sept. 12, 1959	first man-made craft to impact the Moon
Ranger 7 July 28, 1964	first craft to obtain high-resolution (one metre) television pictures
Ranger 8 Feb. 17, 1965	similar mission to Ranger 7
Ranger 9 March 21, 1965	similar mission to Ranger 7, but in highland region
Luna 9 Jan. 31, 1966	first craft to soft-land and relay panoramic and close-up pictures from lunar surface
Surveyor 1 May 30, 1966	first U.S. craft to soft-land and relay panoramic and close-up pictures from lunar surface
Luna 13 Dec. 21, 1966	similar to Luna 9, but also tested lunar soil hardness
Surveyor 3 April 17, 1967	similar to Surveyor 1, also first craft to dig trenches
Surveyor 5 Sept. 8, 1967	similar to Surveyor 1, also first craft to analyze lunar mare soil
Surveyor 6 Nov. 7, 1967	similar to Surveyor 5
Surveyor 7 Jan. 7, 1968	similar to Surveyor 5 and 3, first soft landing in highlands
Apollo 11 July 16, 1969	first manned landing (Mare Tranquillitatis)
Apollo 12 Nov. 14, 1969	first manned landing near an unmanned craft (Surveyor 3)
Luna 16 Sept. 12, 1970	first unmanned craft to return lunar soil to earth
Luna 17 Nov. 10, 1970	first unmanned craft to deploy a roving vehicle (Lunakhod 1)
Apollo 14 Jan. 31, 1971	first manned landing in highland region (Fra Mauro)
Apollo 15 July 26, 1971	first manned landing to use roving vehicle (Hadley Rille)
Luna 20 Feb. 14, 1972	similar to Luna 16, but highlands sample instead of mare
Apollo 16 April 16, 1972	similar to Apollo 15, but in highland region (Descartes)
Apollo 17 Dec, 7, 1972	similar to Apollo 16, but sampled highland and adjacent valley

Ray and dark-halo craters

A crater may be classified also by the neighbouring craters or other features that appear to be associated with it. Thus, craters such as Tycho (Figure 2), Aristarchus, Copernicus (Figure 3), and Kepler are known as ray craters, being major centres of subradial systems of bright rays that are now known to consist of numerous small, bright craters and blocks of light-toned rock. The rays of Tycho are up to 1,900 kilometres (1,100 miles) long. A number of kilometre-sized craters are known as dark-halo craters; these have associated dark-toned rings or halos around them that may be pyroclastic (volcanic) deposits. Other craters are associated with subradial linear formations that are distinct from rays and others again with adjacent positive elements of terrain known as deltoids. Dome-top craters are conspicuous because each occupies the summit of a broad, topographic swelling. Small craters often seen on the summits of hills or mountain peaks may form a class of their own (summit craters). It is common to find an unexpectedly high number of craters on ridges in the maria or centred in the crests of crater walls. Such craters have been referred to as parasitic or ringwall craters. A crater that is one member of a linear progression, or chain, of craters may be described as a chain crater, particularly if other members of the progression are of a common and distinct morphological type. Other mutually similar craters form arcuate chains. Consanguineous craters with strong morphological similarities to one another may be numbered among the many members of a chain, but they are commonly found in close pairs also. The walls of interlocking craters may pass across one another without interruption. Finally, rille craters are those single or chained craters that are set in and form part of the lunar rilles, which are trench-like features. A rille crater may be circular or markedly elliptical and may be the centre of a dark halo.

Height measurements from shadows

Height measurements form an essential element in profile construction. In principle, the height (h) of a lunar mountain peak above the lunar ground where the tip of the mountain's shadow, of some length (ℓ), lies is determined by simple geometry provided the mountain is near the centre of the Moon's disk or measured on a photograph taken vertically. Taking the elevation of the Sun over the mountain as θ degrees, the height h of the mountain is calculated with the formula $h = \ell \tan \theta$; the angle θ (theta) can be obtained from regularly published tables. For mountains not viewed under these conditions, it is necessary to use a more general formula for h.

By courtesy of National Aeronautics and Space Administration

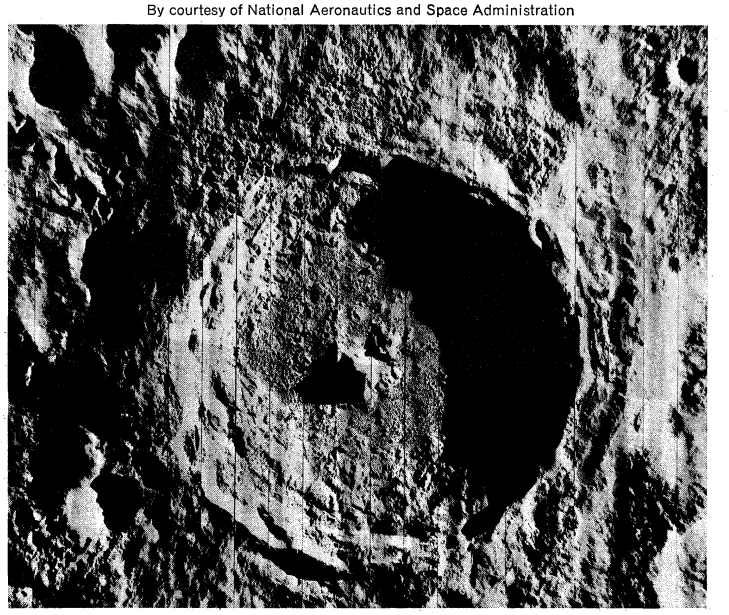


Figure 2: The lunar crater Tycho, in the lunar highlands. Picture obtained by the U.S. Lunar Orbiter 5.

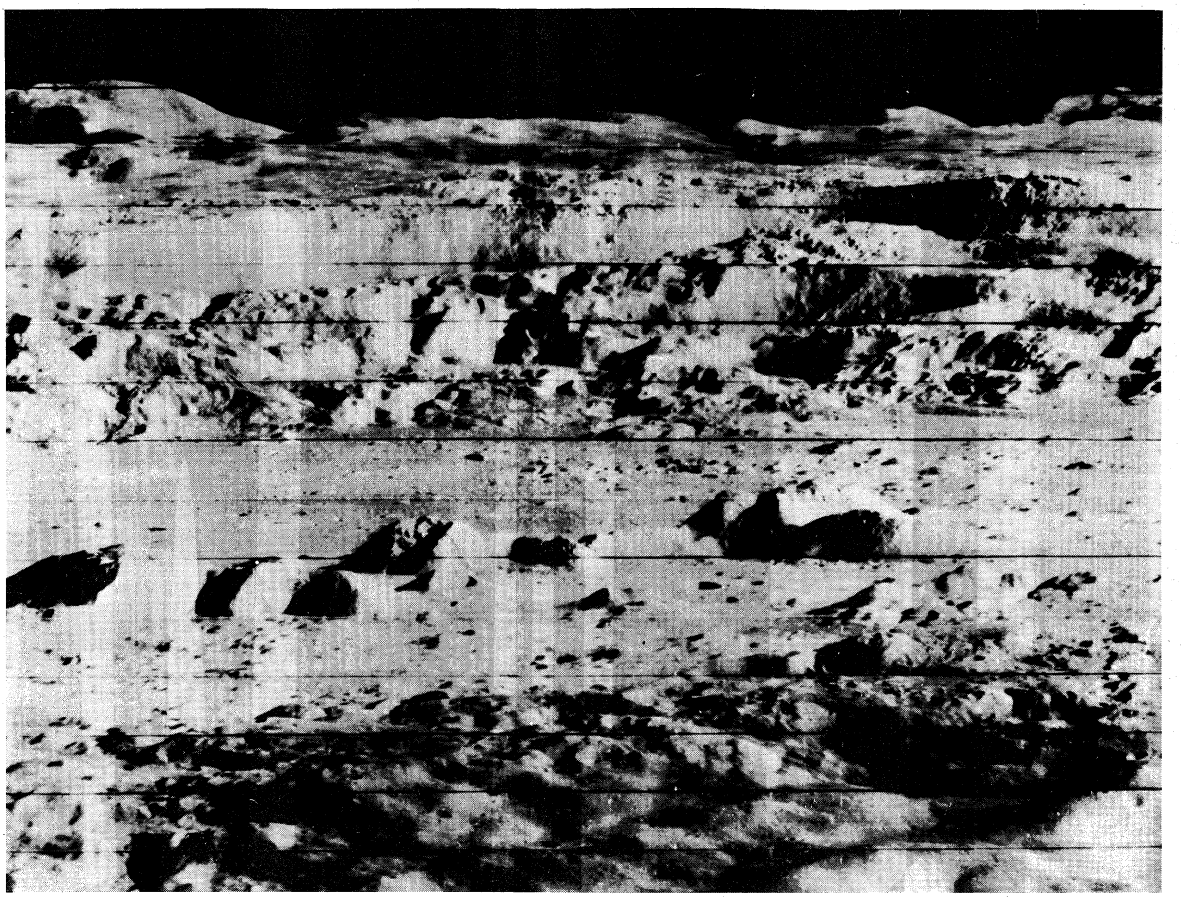


Figure 3: Crater Copernicus as photographed in November 1966 by Lunar Orbiter 2 from a height of 28.4 miles, looking due north from a position above the crater's southern rim. The mountains rising from the flat floor of the crater are about 1,500 feet high; the 3,000-foot mountain Gay-Lussac, on the horizon, is part of the Carpathian range.
By courtesy of National Aeronautics and Space Administration

The near side of the Moon has been surveyed fairly intensively, but, until recently, methods employed to measure the Moon's overall shape, or figure, have been complex and laborious. Unlike the Earth, with its oceans, the Moon has no natural datum from which to measure altitudes. The physical librations (see above) show that the Moon cannot be both spherical and homogeneous. An accurate evaluation of the Moon's figure has become important in modern discussions of the origin of its surface features as well as being relevant in discussions of its internal constitution.

In the limb seen from Earth, the Moon departs from a circle by only about 100 metres (300 feet); but the equivalent length of the lunar axis that points toward the Earth is about 1,000 metres greater than the Moon's polar axis. The maria are generally lower than the lunar highlands by an average of about 1,000 metres. Individual mountains attain heights of up to four or five kilometres (about three miles) above their bases, and this is also the order of depth of the deepest lunar craters commonly measured from the rim crest of the enclosing ramparts, which are in some cases over 200 kilometres (120 miles) in diameter.

Smaller craters are progressively less deep: whereas the deepest and largest craters of the Moon have a depth-to-diameter ratio of only 0.08, lunar craters of ten-kilometre-diameter may have a depth-to-diameter ratio of 0.15. The latter, however, are the eumorphic (sharply defined) craters; and all ten-kilometre-sized craters, taken together, show a wide spread in the depth–diameter ratio. The tendency in the case of eumorphic lunar craters is for the depth-to-diameter ratio to increase as the size of a crater diminishes, a generalization known as Ebert's rule, after Hermann Ebert, a German physicist.

Ebert's rule and Schröter's rule

Schröter's rule, named for a German astronomer, Johann Hieronymous Schröter, states that the volume of crater-rim materials lying above the mean level of the local terrain is approximately equal to the volume of the crater depression below that mean level. Schröter's rule has been repeatedly quoted as evidence supporting the impact origin of the vast majority of lunar craters, yet the data on which it was based were inadequate, and exceptions are known, in which the volume of a crater's rim materials exceeds the volume of the space inside it. Mechanisms that might have contributed to the filling of craters are, for example, isostatic adjustments (movements of matter caused by differences in surface loading); infilling through erosion, transport, and deposition; and lava flooding.

The deepest and largest craters of the lunar near side are concentrated in the central meridional belt of the Moon's southern hemisphere. The deepest craters tend not to be near the maria. The encircling walls of the oldest craters have slumped into the craters or have been cut across by faults or graben; the craters themselves have been overlapped and partially obliterated by smaller craters formed later. In addition, the topography of large craters is often softer than that of many of the sharply sculptured, more recent craters, presumably because of changes resulting from denudation (erosion).

Dense cratering is found in the lunar highlands, and many of the older craters are difficult to observe. About four times as many craters between ten and 50 kilometres (five and 30 miles) in diameter are found per unit area of the nearside highlands as in the same area of the average mare.

Placement of craters

Standard statistical methods applied to lunar-crater data have shown clearly that the lunar craters are not distributed across the lunar surface in the random fashion expected as a result of meteoric impacts.

The maria abut on the highland domains, and mare materials are found in craters with breached walls on the peripheral zones of maria. Geologists believe that the

mare surfaces are younger than the highlands, and it is clear that late phases of mare lava outpourings could blanket craters and render their overall distribution nonrandom. The maria themselves differ in age, as stratigraphic maps and dated rocks returned to Earth show. There are, even in a single mare, distinct geological units (such as individual lava flows), each with its own characteristic population of craters.

An attempt to distinguish by statistical methods between craters of impact (including secondary-impact) origin and endogenic craters in the lunar maria has been made. The ratio of internally formed craters probably varies from one unit to the next, and, of the craters with diameters between 50 metres and hundreds of metres, it is not uncommon for up to half to be arrayed with the trends of known tectonic lineaments (large-scale linear features; see above). Such craters are probably of internal origin.

Lunar lineaments. *Mare ridges.* Lunar lineaments (linear features) are of many different types. Mare ridges (Figure 4) are numerous in the maria and may be much

By courtesy of the Lunar and Planetary Laboratory, University of Arizona

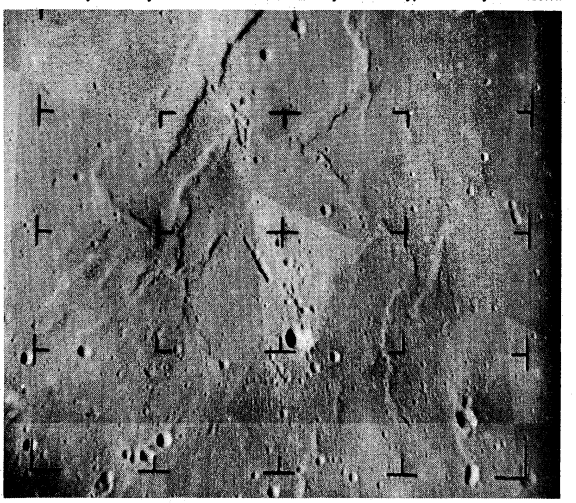


Figure 4: Part of Mare Tranquillitatis, showing dimple craters and mare ridges. Area shown is about 19 kilometres (12 miles) square.

longer than the brighter highland-type or lunaritic ridges. Mare ridges are also found in dark-floored craters and, rarely, may be traced beyond a mare into highland-type material. In transverse profile they consist, typically, of broad, low rises sloping at only a degree or so for ten or more kilometres, capped by slightly steeper, relatively narrow ridges. Typically, the major mare ridges rise to an altitude of about 200 metres (650 feet) above the surrounding country. In plan, they lack continuity and are broken, offset, or arranged *en echelon*. The higher, narrower parts of the ridges show a strong tendency to be in short sections, or segments. While the overall trend of these ridges in a subcircular mare is generally aligned with the curved border of radially more distant mountains, the directions of the mare-ridge lineaments often follow other preferential directions. Like the lunaritic ridges, they may also belong to families of ridges that trend over much of the lunar surface in set ranges of azimuth—a fact that labels many of them as tectonic; *i.e.*, produced by deformation of the crust.

The lunar grid system. Other lineaments, such as linear graben (depressed regions between faults), crater chains, rilles, and faults, together with the ridge network, form the lunar grid system. Near the centre of the Moon's near-side face the principal trends of the grid system are at about 45° to the central meridian. Interference between these dominant directions and other directions of lineaments that are subradial to certain of the maria, notably Mare Imbrium, makes recognition of the Moon-wide topographic pattern difficult. Many of the mare ridges, from which extensive flows of lava have derived, have evidently been built up by extrusion of basic lavas from fractures. Because of their tendency to form an overall grid pattern, many of the fractures must have been produced by uniformly acting crustal stresses of tectonic origin.

Regional lineaments subradially oriented (*i.e.*, somewhat offset from a radial direction) with respect to Mare Imbrium and Mare Orientale are easy to locate; fewer ridge lineaments are associated with Mare Nectaris and Mare Humorum. Such regional fracture or fault systems extend into and are modified by the older terrain around the mare. The circular mare sites on the Moon are at present thought to have been blasted out by gigantic impacts, and the basins so created later filled with lavas from inside the Moon. Much faulting and possibly the extrusion of lavas in the regions surrounding each mare basin would be generated in such major events. Fractures in a previously unstressed homogeneous medium will be radial and tangential at some distance from an explosion site. The systematic deviations of the mare lineament systems from these directions in association with the more general grid system suggest influence from preexisting fractures or weaknesses in the lunar crust.

Similar patterns are found around certain large craters, and these may be detected most readily in mare material. Around the crater Aristillus, bundles of lineaments, rather than a set of precisely radial ridges and troughs, are found. Only the central, longest lineament of each bundle is radial to the crater, while the other members are progressively offset from the radial direction. The central lineament of a bundle tends to be aligned with one of the principal directions of the grid system. The ultimate explanation of these minor sets of subradial lineament is expected to be analogous to that pertaining to the maria.

Where the grid system abuts on the larger craters, northeast- and northwest-trending ridges commonly intersect and sometimes pass inside the northern or southern ringwalls. Well-developed nests of crisscrossing ridges found in such localities have been called lattice patterns. A quite different feature, named the deltoid, also commonly occurs outside the northern or southern walls of certain craters. Deltoids are raised elements of the lunar terrain in the form of a delta, or gothic arch, the base of which abuts on the rim of a crater.

Of the subcircular maria, some tend to be polygonal. As in the case of the polygonal craters, the shape might be due to crustal weaknesses existing before the mare was formed. The peripheries of other areas of lunabase, including some maria, are irregular. In addition to craters and mare ridges, maria also contain features known as rilles (German *Rille*, "trench," "groove," "channel"; Latin *rima*, a term used principally on official maps), which are, however, found also in the lunabase generally, in certain of the larger craters, and in the lunarite.

Normal rilles

Rilles. One class of rille, called the normal rille, appears to be an exclusively lunar phenomenon. This is a linear or broadly arcuate trench without raised rims (except when it cuts through higher ground) and with gently insloping walls. The largest rilles are up to five kilometres (three miles) wide, a few hundreds of kilometres long, and up to several hundreds of metres deep. The floors of normal rilles are relatively level (unless they are so narrow that fallen rocks and dust from the opposing walls meet in the bottom) or are humped or strewn longitudinally with blocks. They slice through the mountainous walls of some craters and traverse the lunarite for great distances. They tend to occur in parallel families, in arcuate paths generally just within the mountain ranges or hills that border maria or the larger craters, and to run in preferential directions suggestive of lines of probable crustal weakness and tectonic control. Some are arranged *en echelon*, a characteristic of strike-slip faults on Earth.

Single or contiguous craters without raised rims define parts of the walls and floors of some normal rilles; these have sometimes been termed crater rilles. The craters appear to be of the collapse type, with flattish floors. Lunaritic ridges traverse some normal rilles (such as the Ariadaeus and Hesiodus rilles) and appear to have sagged or subsided into the rilles. Such ridges seem to be older than the rilles that they cross without a break; and this shows also that the rilles are not partially filled fis-

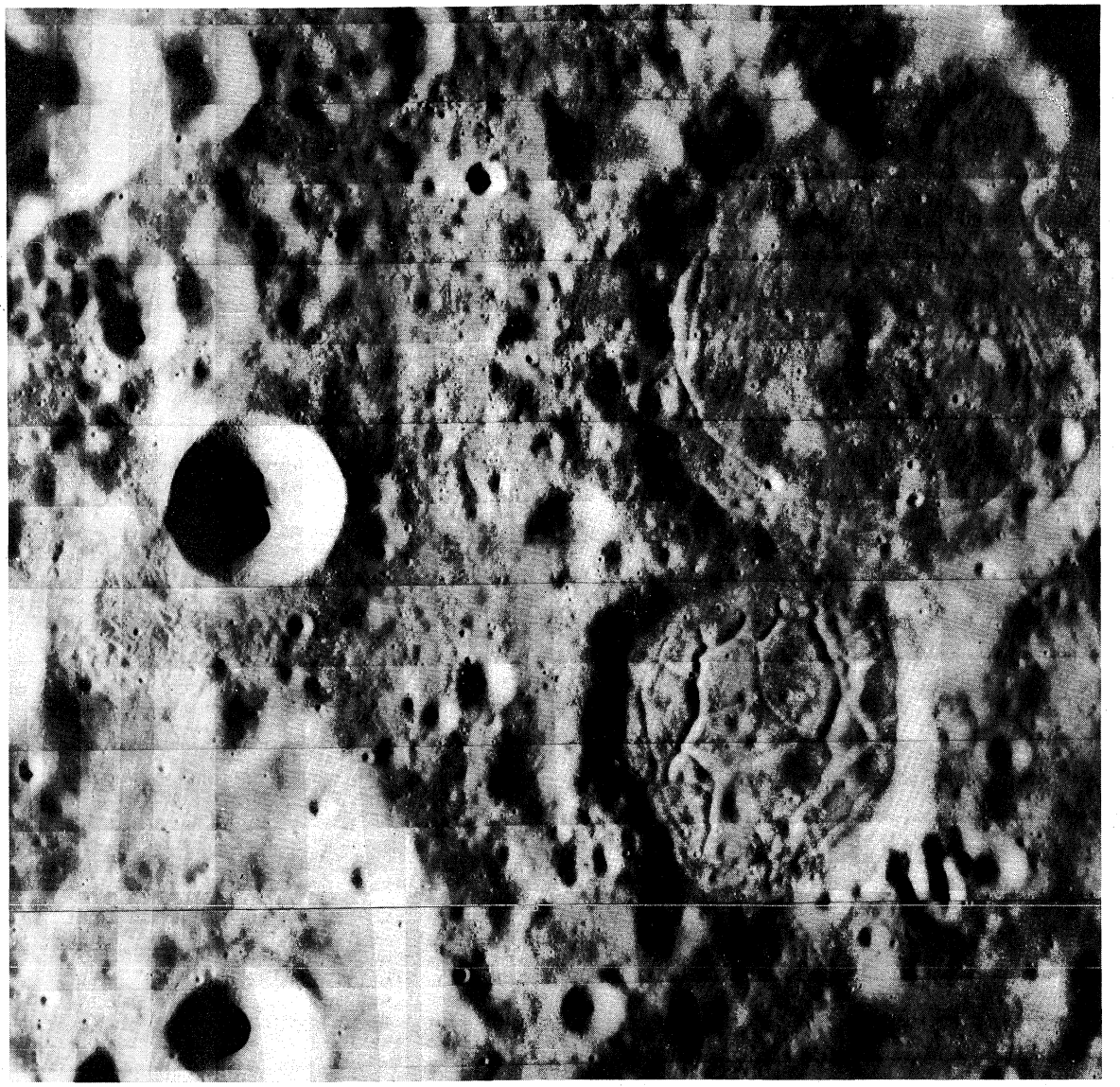


Figure 5: An unusual fractured crater floor on the hidden side of the Moon, photographed by Lunar Orbiter 1. The fractures appear as an irregular pattern similar to a dried mudflat. The heavily fractured crater (bottom right), which may be a source of volcanic materials, is 15 miles wide; the larger crater just above it (to the north) is about 24 miles across at its widest point.

By courtesy of National Aeronautics and Space Administration

sures. These observations suggest that the normal rilles are graben, or collapse features, and are probably controlled by underlying fractures or faults.

In many striking instances, normal rilles run into or connect with other normal rilles at sharp angles and with faults, chains of craters, or mare ridges, suggesting that the rilles in question are controlled by tectonic fractures.

Sinuous rilles

During the 1960s, examination of space photographs stimulated much discussion of another type of lunar rille—the sinuous rille (Figures 2, 5). Some parts of sinuous rilles, like normal rilles, appear to be controlled in direction by tectonic lines of weakness. Other parts have features resembling meanders in a terrestrial river. The superficial resemblance to dry rivercourses—particularly to those cut by rivers draining inland into desert areas—led to a revival of interest in water on the Moon. But later detailed measurements of the geometrical parameters of sinuous rilles lent weight to another hypothesis—that these rilles were shaped by the flow of lavas. Many of the sinuous rilles have one end marked by a probable (elongated) source crater, and many degenerate into a chain of craters that are commonly elliptical and have their longest axes set along the probable courses of underlying lava tubes. In a few cases, smaller sinuous rilles meander along the floors of larger ones, such as Schröter's Valley. Several sinuous rilles have distributaries or re-entrant loops, both of which are common in terrestrial lava channels.

Like normal rilles, sinuous rilles tend to concentrate around the borders of the circular maria or at the junctions of lunabase and lunarite. Exceptionally strong concentrations of sinuous rilles are found in the densely domed region near the crater Marius and also around the crater Aristarchus. Thus, sinuous rilles show a preference to regions that have been assessed from other evidence as showing volcanic characteristics.

Opposing sides of a few stretches in some normal and sinuous rilles are observed to be at different levels. In principle, a rille may form by the collapse of a weakly cohesive regolith into an open fracture. Faulting is initiated when there is relative movement between the two blocks that define the fracture. The displacement of a fault may remain constant, or there may be continuing or intermittent movement. The direction of displacement may be steeply inclined to the horizontal (normal, or dip-slip, fault); steeply inclined to the vertical (thrust fault); or lateral (strike-slip, wrench, or transcurrent fault). Faults on the Earth and Moon may show a combination of these idealized directions of displacement. On the Moon, however, no thrust fault has yet been recog-

Lunar faults

nized. The term wrinkle ridge, a synonym for mare ridge, is suggestive of compression or thrusting, but there is little or no real evidence that any of the mare ridges (Figure 4) are compressional features.

Some rilles are directly associated with probable dip-slip faults. The characteristic *en echelon* pattern of strike-slip faults is reflected in other rilles, in segments of wrinkle ridges, and in lunaritic ridge systems. More than one major strike-slip fault appears to have offset blocks of the lunar Caucasus Mountains; and there are a host of probable minor strike-slip faults on the Moon, one of the most notable being the fault that has displaced the walls of the crater Capella and of another two craters further to the south. Strike-slip faults are perhaps of more importance on the Moon than the better known dip-slip faults, such as the inappropriately named Straight Wall. Major faults, dissecting large areas of lunarite, have been recognized. These "selenofaults" measure hundreds, rather than tens, of kilometres in length.

Bright rays. The bright rays that diverge from craters such as Tycho (Figure 2) are thought by some to be fracture controlled, but the majority of lunar geologists believe the rays to be the result of ejecta of secondary particles from a primary impact site. One of the strongest arguments for the impact origin of rays is that they look like splash marks. The particles that formed the end of the longest ray of Tycho must have been ejected from Tycho at a velocity of at least 1.4 kilometres (0.9 mile) per second if the ejection hypothesis is correct. The velocity of escape of a particle from the lunar surface is 2.4 kilometres (1.5 miles) per second, so potential ray-forming particles would not escape from the Moon. The evidence from closeup photographs is that more than the average number of craters are in unit areas crossed by rays; and evidence gathered by astronauts points to there being a complement of bright-toned rock blocks in a ray-strewn zone.

Mountain height measurements

Lunar mountains. Systematic estimates of the altitudes of lunar mountains were first made in the 18th century. The most comprehensive program of lunar height measurement ever attempted was started in 1959 by the United States Air Force's Aeronautical Chart and Information Center. The results of this concerted program, in which many thousands of lunar photographs were taken at different Earth-based observatories, are incorporated in lunar nearside charts prepared by the Aeronautical Chart and Information Center.

The classical program of heliometer observations of the angular separation of points on the Moon, made in the 19th and early 20th centuries, is of fundamental importance in the present understanding of the lunar physical libration. The time-consuming reductions involved the computation of the distances between Mösting A (a crater near the centre of the Moon's disk) and many control points scattered over the Moon's near side and, finally, solution for constants characterizing the librations.

Another series of classical observations was made to elucidate the motion of the Moon in its orbit. When the Moon passes in front of a star it is said to occult the star. Up to the present time, tens of thousands of occultations have been timed in this continuing study of the Moon's movements and of the shape of the Moon's limb in the librational zone (see also above).

Temporary or transient changes on the Moon. Ever since serious telescopic observation of the Moon began, astronomers have reported mistlike phenomena or temporary obscuration of detail in specific patches of the Moon. In 1966 a significant positive correlation was found between the times of more than 100 observations of transient lunar phenomena made between 1749 and 1963 and the distance of the Moon from the Earth. The majority of transient lunar phenomena occurred close to the time of the Moon's perigee or apogee passages. The indication was that the Earth's tidal forces on the Moon's solid crust had contributed to the cause of the transient lunar phenomena, although the transient lunar phenomena themselves may be manifestations of seismic (moonquake) or seismically triggered volcanic activity or merely caused by dust either blown out from the regolith and dispersed by cold gases escaping from cracks in the Moon or set in motion by seismic activity. Similar periodicities in some types of seismic signal from observing stations planted on the Moon show that lunar seismicity occurs most commonly within a day or so of perigee and of apogee, and this suggests that the majority of the transient lunar phenomena are connected with some kind of seismic activity.

Other kinds of lunar change have been suggested. Perhaps the most frequently discussed case of a possible permanent or long-term change of topography is that of the crater Linné, which may be located on many photographs as a whitish patch in Mare Serenitatis and appears on the best photographs of today as a crater only 1.5 kilometres or so in size, centred in a bright halo. It cannot be concluded from the evidence available that the area has suffered any morphological change.

Observed eruptions

In 1958, a Soviet astronomer, N.A. Kozyrev, claimed to have detected a reddish cloud of volcanic ash, followed by an efflux of gas lasting for at least half an hour, from the vicinity of the central mountain block of the crater Alphonsus. He recorded a lunar emission band having its maximum intensity at 4737 angstroms and gradually fading toward the violet end of the spectrum. This and other bands, he proposed, had their origin in carbon molecules. Some volcanoes on Earth emit carbon dioxide gas. In 1959 Kozyrev again obtained an unusual spectrogram of Alphonsus and maintained that the central block of the crater had given birth to an eruption of lava at a temperature exceeding 1,270° C (2,320° F). Although these observations and interpretations have not been universally accepted, the reports did stimulate interest among other professional astronomers in a Moon long considered to be dead and changed only by impacts.

Notable reddish glows on the crests of ridges in the Aristarchus region were observed in 1963 at the Lowell Observatory in Flagstaff, Arizona. A month later, similar observations of the Aristarchus region were reported at two observatories. At this time, coloured patches persisting for over an hour seem to have been authenticated.

The photometric function. Scientists correctly reasoned, in the early 20th century, that the Moon's photometric function (changes in the intensity of reflected sunlight with different angles of reflection) proved that the lunar surface was far from smooth on a centimetric scale. The respective scattering properties of highland and mare material were deduced in 1956, when it was shown that the lunar surface in both the lunarite and the lunabase was extremely irregular and porous and that both types of terrain displayed essentially similar surface structure. Slight differences were observed in the way light was scattered by lunarite (highland) and lunabase (maria) material, and these differences indicated that, in general, there was a slightly greater roughness (more irregularities, or pores, per unit of surface) in the lunabase. In a more sophisticated analysis, in 1971, it was estimated that the porosity (fractional pore space) of the highland topsoil had an average value of 60 percent and that the mare topsoil averaged 80 percent. The measured values of the Moon's surface porosity at specific sites are not generally in conflict with these overall results.

Porosity of topsoils

The reason for the different structure in the uppermost centimetre or so of the lunarite and lunabase may be found in a study of the differential denudation or of the different histories of outgassing of the two types of terrain. Processes such as mass wastage—that is, downslope migration of material under gravity—may be expected to play a relatively important role in the highlands. Therefore, there may well be a tendency for craters and pores in the highlands to be smoothed over more quickly than corresponding features in the maria. On the other hand, the outgassing of the mare materials is believed to have occurred later than much of that in the highlands; and the initial porosity of the two types of material probably differed as well.

The packing of lunar fines. A special use of photometry is in the determination of the packing of lunar fines (small particles and dust). The acceleration caused by the Earth's gravity is six times that of the Moon at their

respective surfaces, so that fines scooped from the Moon and transported to Earth will undergo an increase in their state of compaction. Agitation, severe during parts of the journey back to Earth, assists the process of compaction. The undisturbed fines and samples returned to Earth will scatter light in different ways, and it is possible, by rearranging the compacted fines into various states on Earth and testing them again, to deduce their approximate surface state before they left the Moon.

Colour contrasts. It is difficult to assess large-scale colour contrasts on the Moon from the Earth by eye. A profitable investigation of lunar colour contrasts may be made from ultraviolet and infrared photographs. The technique is useful for surveying extensive areas of the Moon for differing rock units.

Certain parts of the Moon absorb ultraviolet light strongly. Such an absorption can be produced by traces of the element sulfur, for example. At the other end of the spectrum, lunar rocks display characteristic absorption bands in their infrared reflection spectra. One such prominent band, at a wavelength of about one micron, is due to the presence of iron in certain lunar materials.

Another instructive technique is that of plotting the photometrically determined colours of discrete patches of lunar terrain against their respective albedos and comparing the resulting colour-brightness diagram with others constructed for a wide range of terrestrial rocks prepared in widely different forms. It is found that blocks of terrestrial-type rock, or meteorites, cannot cover more than a small fraction of the lunar surface, although many rocks ground to powders and then irradiated with protons (to simulate the effect of particles from the Sun striking the Moon) have been found to approach the lunar domain of the colour-brightness diagram.

Lunar polarization studies. *The lunar soil.* A much more powerful optical method of studying the nature of the particles of lunar soil is that in which polarization curves are utilized. Polarization is a wave property of light, and different substances have different characteristic polarization curves. At a given lunar phase, the amount by which light from different parts of the Moon is polarized depends on the brightness of those parts: the brightest spots polarize no more than about 5 percent of the incident light, whereas the degree of polarization attains 20 percent or so for the darkest spots. These maxima were recorded at times close to those of the First and Last Quarters of the Moon. Most terrestrial rocks, whether solid or pulverized, polarize much more light than the Moon. The Full Moon polarizes no light, and practically no light is polarized when the Moon offers a very narrow, sunlit crescent. When the light intensity component in the plane Sun–Moon–Earth exceeds the corresponding component at right angles to this plane of diffusion, the maximum polarization in the visible (yellow light) is found to be —1.2 percent, for all parts of the Moon. The independence of the negative part of the polarization curve on the location on the Moon's near side indicates a fairly high degree of textural uniformity of this part of the lunar surface. Extensive laboratory experiments made with visual and photoelectric polarimeters, in attempts to narrow the choice of lunar surface models by careful comparison of laboratory and lunar polarimetric functions, have shown that the best match to the lunar soil at large is in dark, fine-grained basic to intermediate rocks.

Angle of inversion

In changing from positive to negative, the Moon's polarization passes through zero at a phase angle of about 23° called the angle of inversion. The angle of inversion remains constant only when the behaviour of light of a particular wavelength is considered: the angle increases with wavelength (whereas the degree of polarization generally decreases as the light becomes redder). The maximum polarization, at a given wavelength, generally decreases as the particle size decreases. Reducing the grain size of a rock usually has the effect of increasing its albedo, and this, in turn, generally decreases its maximum degree of polarization.

The maxima of the degree of polarization versus wavelength for at least one fines sample returned to Earth from Mare Tranquillitatis are identical to that of Mare Tranquillitatis as a whole, observed telescopically from Earth, suggesting that the polarimetric curves of fines just below the surface are similar to those pertaining to material actually on the surface.

It has been suggested that solar protons not only damage but also darken particles on the lunar surface. It is difficult for an experimenter to simulate all the appropriate aspects of the lunar environment in the laboratory, but it is clear that, if the solar wind were to darken fine particles derived from certain terrestrial rocks, their polarization functions could be made to approach or be rendered identical with those pertaining to the lunar fines.

The lunar rocks. Polarization studies of light-toned rocks brought from the Moon have shown that they either do not occur in large concentrations on the Moon's surface or they are generally covered with the fines responsible for the uniformity of the Moon's nearside polarization.

Infrared (heat) studies. Other evidence of a widely dispersed layer of dustlike material on the Moon is given by measurements of infrared radiation from the Moon. The lag shown by temperature changes of the subsolar point during eclipse shows the very poor conductivity of the lunar surface in the eight- to 14-micron (one micron $= 10^{-6}$ metre) waveband. Highlands and maria behave similarly, so the poorly conducting layer must cover most of the surface.

The poor surface conduction also produces differences in heat curves at different wavelengths. Measured at a microwave wavelength of 1.25 centimetres, the Moon's temperature attains a lower maximum than that determined by the eight- to 14-micron infrared measurements, and this lower maximum temperature occurs about 3.5 days later than the infrared maximum. The eight- to 14-micron radiation is from a surface layer only a millimetre or so thick, whereas the centimetre radiation comes from layers a few tens of centimetres beneath the surface. The Sun's heat generates the highest temperatures on the surface of the Moon, but deeper rocks are also warmed as the thermal wave penetrates and is attenuated in the poorly conducting regolith. Lunar radio emission derives from thermal energy finding its way to different levels in the Moon. Increasingly longer radio wavelengths have been used to observe the Moon at increasingly greater depths; but, although the average depth of the emitted waves increases with wavelength, the radiation at radio frequencies originates over a range of depths. In addition, the limit of angular resolution of a telescope increases with wavelength, so that, whereas the optical heat measurements may refer to a small patch of the Moon, the specific origin of the longer radio waves may be impossible to locate on the disk. At wavelengths above about ten centimetres, no diurnal change in the intensity of received radiation has been observed. In these cases the upper layers of the Moon, which show strong diurnal temperature changes, are contributing a negligible fraction of the total radiation received.

Changes in temperature with depth

In principle, radio observations provide a means of estimating the variation of temperature with depth in the regolith and, thus, of investigating any internal heat source such as one that could be produced by widespread radioactive elements. The thermal conductivity at depth in the Moon, however, is not yet well-known, nor is it in the troublesome regolith, in particular, where the temperature and depth dependence of conductivity is marked and thermal transport by radiation through incompletely specified vacuum pores contributes to the difficulties.

Theoretical models of the Moon's upper layers have been attempted, but it is not possible to derive the conductivity directly: density and specific heat (the amount of heat required to raise the temperature of a particular substance by a given amount) must be determined first. Furthermore, the specific heat of lunar rock depends on its temperature: the specific heats of both a vesicular basalt sample and a sample of regolith from the Moon have been found to change smoothly from 0.06 calorie per gram per °K at 90° K to 0.2 calorie per gram per °K at 350° K (0° K equals —273° C [—460° F]). Such dependence on the prevailing temperature can change the

thermal inertia of lunar surface materials by a factor of up to 2. A representative value of the thermal conductivity of the uppermost part of the regolith is about 5×10^{-6} calorie per centimetre per second per degree.

The so-called hot spots

During optical infrared scans across the eclipsed Moon with a resolution of details of about 20 kilometres, numerous patches of terrain—mostly eumorphic craters—have been found to be significantly warmer than their surroundings. The most easily recognizable thermal anomalies (hot spots) were associated with bright-ray centres such as Tycho, Copernicus, and Aristarchus. These craters remained cooler than their surroundings during the lunar daytime, and the hot spots could be areas of higher than average thermal conductivity permitting heat at depth to flow out during eclipse and be radiated to space more readily than from adjacent areas.

The night spectrum of Tycho, Copernicus, and Aristarchus, observed over the three- to 15-micron waveband, proved that the surface rocks of these craters emitted radiation from areas at widely different temperatures. Parts of the floors of two, Tycho (Figure 2) and Aristarchus, appear to be of ashy or regolith-type material, but other parts are probably volcanic domes apparently not covered by detritus or pyroclastics (substances fragmented by volcanic action). With geologically dissimilar rock units found in and around the craters, including lava flows with contorted surface ridging, such craters must be expected to generate cooling curves that are made up from several different cooling functions.

The hottest craters are invariably the most sharply sculptured, freshest looking features. Large blocks of rock and characteristically rough terrain may be resolved around several of these craters, the hottest 20 of which lie in the diameter range four to 16 kilometres (two to ten miles).

There are relatively few hot spots per unit surface of the highlands compared to the maria; the number density of hot spots changes significantly from one mare to the next. Mare Tranquillitatis, for example, sports an anomalously high number of hot spots per unit of its surface. The reasons are not well understood, though a tentative explanation may be offered in the last case. First, there is a positive correlation between the number of eumorphic craters in the lunabase and the number of hot spots in the same area. Second, a study of the shape parameters of small, concentric craters leads one to believe that at least the southern part of Mare Tranquillitatis has a thinner cover of regolith than most of the maria. It may be imagined that the sharpest eumorphic craters—which are almost certainly explosion craters—would be formed preferentially in areas of thin regolith, where the excavation of coherent bedrock could provide an abundance of ejected rock blocks.

A few cold spots have also been observed. Of the nine cold spots with the highest thermal gradients (those in excess of $\frac{5}{7}°$ C per second of arc), none is associated with a single, eumorphic crater.

Finally, there is the type of thermal anomaly in which a spot is observed to be hot during both eclipse and daytime conditions. Such a spot could be indicative of a source of internal heat.

Analysis of returned radar signals. For better understanding of the nature and origin of lunar thermal anomalies, more widely dispersed and improved estimates of the regolith thickness, layering, and conductivity are required. Analyses of radar returns from the Moon have proved of interest in this connection. The first part of a radar echo from the Moon returned to Earth derives from the central region of the Earth-pointing face. Later parts of the return signal come from rings of progressively increasing radii. Also, the lunar librational motions introduce frequency shifts in the return signals. As these effects are superimposed, it is possible to isolate the radar-reflection characteristics of two small patches on the Moon. Anomalously strong radar reflections were obtained in 1962 from Tycho and the corresponding area in the Moon's northern hemisphere. The latter had no topographic peculiarities, and it was concluded that Tycho was probably responsible for the anomaly: either the terrain in and around Tycho was very much rougher than that in any other equal area of the Moon's near side, or the rocks of Tycho were very much denser than those elsewhere. Since that time, photogeological studies of high-resolution pictures of the Tycho (Figure 2) area have shown that it is indeed rough.

Radar has been used, further, to estimate the porosity of lunar-surface materials. On radar wavelengths of ten to 100 centimetres, the effective dielectric constant (its ability to store electrical potential energy) of the lunar surface lies in the range of 1 to 2.7. To calculate its actual porosity, this can be compared with the assumed dielectric constant of its solid component. Returned fines have been used to estimate that the density of the solid rock is about 3 grams per cubic centimetre and that the corresponding dielectric constant at 450 megahertz is about 6. These figures, with a mean observed lunar dielectric constant of 1.6, show that the average porosity of the lunar surface layers may be over 80 percent. The results compare favourably with those from Earth-based photometry, although the two methods refer to data from different areas and to different effective depths.

Relative reflectivity

Radar studies of the relative reflectivity of small areas of the Moon's surface at 2.2 metres in wavelength involving the circumlunar satellite Explorer 35 show that, on average, the maria reflect approximately 30 percent more power than the highlands; and reflectivity structure was observed in the lunabase, a particularly high reflectivity being associated with a region near to the crater Flamsteed. There appears to be a significant correlation among radar reflectivity, hot spots, and regolith thickness; both the radar reflectivity and the number density of hot spots increase as the thickness of regolith decreases. The effective (mean) dielectric constant of the lunar rocks (lunarite and lunabase together) down to depths of several tens of metres is about 3, slightly more than the mean value of the dielectric constant of layers nearer the surface measured on shorter wavelengths from the Earth. The porosity of the lunar rocks should decrease, on average, with depth as, in a fragmental, or granular, regolith, the weight of the overburden will cause increased packing as the depth increases.

Physical experiments from orbiting spacecraft. Other physical experiments conducted by the Soviets from circumlunar orbits include a gamma-ray-spectroscopy experiment carried aboard the first artificial lunar satellite. Gamma rays are emitted by lunar radioactive elements, particularly potassium, uranium, and thorium, and the character of the emission can indicate whether the Moon has been chemically differentiated. Cosmic rays incident on the Moon induce a very much higher intensity of gamma-ray emission than that deriving from natural radioactive decay in the lunar rocks. Nevertheless, Soviet scientists have deduced that the surface rocks of the Moon were of the basaltic type and that a slightly greater intensity of gamma radiation derived from the maria. The indications were that the Moon was a differentiated body.

In another Soviet study, gamma rays from a cesium source carried by a soft landing craft were backscattered from the Moon as the probe was landing and were received and analyzed to determine the density of the surface rocks. Density was determined using a calibration curve of density against intensity of the scattered radiation and was considered to be 0.8 gram per cubic centimetre in a surface layer about 15 centimetres (six inches) thick. (The form of the calibration curve gave rise to an ambiguity in the estimated density. Of the possible solutions 0.8 and 2.1 grams per cubic centimetre, the former was preferred because of other evidence.)

THEORIES OF ORIGIN OF THE SURFACE FEATURES

In the early 17th century the lunar craters were thought to be endogenic; that is, of internal origin. The first part of the 19th century saw truly under way the argument as to whether most craters were caused by the impacts of external bodies or by volcanism. Theories that could not be classified as either volcanic or meteoric (impact) have also been proposed; but lunar craters are now generally

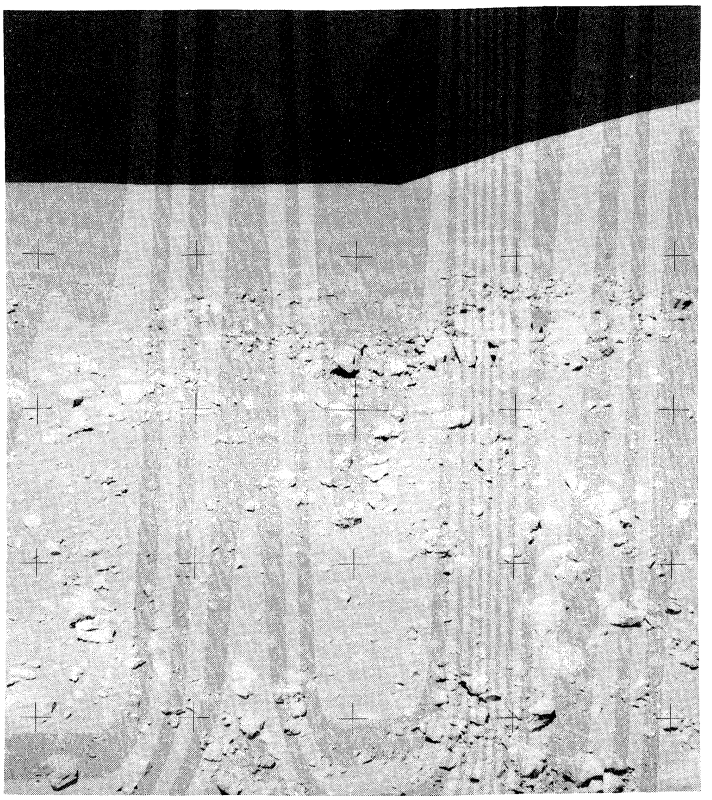


Figure 6: Surface of the moon at the Hadley–Apennine landing site photographed during the third Apollo 15 EVA (extravehicular activity), August 2, 1971.
By courtesy of National Aeronautics and Space Administration

thought to include both impact craters, of all sizes, and endogenic craters, predominantly in the sub-kilometre-size bracket.

Gas-bubble theory

Volcanic theory. One of the earliest hypotheses of the mode of formation of the lunar craters was the so-called gas-bubble theory of the English physicist Robert Hooke, who proposed in 1655 that they were formed by the collapse of huge bubbles of gas; the discharge of gas is a vital part of the volcanic process. On Earth in fluidized mud, as at La Solfatara, Italy, such bubbles commonly measure centimetres or decimetres in diameter, increasing in size with the viscosity of the containing medium and the gas pressure. In Iceland there are hollow domes of lava measuring metres in diameter; on the Moon the low gravity and lack of atmosphere could lead to significantly larger domes. In a later and more attractive theory, as far as the larger lunar craters were concerned, the starting point was a conventional eruptive centre such as might have given rise to a cone volcano of the Vesuvius type. Volcanic materials were ejected from the central orifice in a symmetrical fountain and landed to construct, layer on layer, a roughly circular ringwall. Sometimes, dying activity built up a central eminence; in other cases, the floor of a crater was flooded by dark lavas.

Comparison of lunar features with terrestrial ones

Volcanoes on Earth generally produce cones rather than ringwalls. In Iceland, however, a ring structure about one kilometre in diameter, called Hverfjall, has been formed by steam eruption. Hrossaborg is a similarly formed crater in Iceland; and there are many tuff or lava and tuff rings of kilometre size in terrestrial volcanic fields; for example, in Hawaii, Arizona, and Mexico, where central collapse or engulfment of materials tends to leave landforms resembling many of the lunar craters.

Both the bubble hypothesis and the volcanic-fountain theory were argued without any appreciation of the physical or geological principles involved. More competent arguments for lunar volcanism and caldera collapse are the similar morphologies of particular Hawaiian structures and certain lunar features. In 1944 a U.S. geologist, Josiah Edward Spurr, developed a theory that the larger lunar craters were volcanic collapse structures (calderas); he developed a classification of craters and described in detail the different modes of origin of different classes of crater. More importantly, he was the first to recognize the uniformity and extent of the overall structural pattern in the Moon's crust and to name it "the lunar grid system." Spurr's analyses were closely tied to his meticulous photogeological analyses of the lunar craters. Calderas may have developed during an intensive phase of lunar outgassing, while many of the low-walled ring formations on the Moon marked the positions of ring fractures from which lavas had been erupted in the precaldera phase of lunar crater formation.

Impact theory. The impact hypothesis was formulated comparatively late. Few people had seen a large meteoroid fall on Earth, though many had witnessed terrestrial volcanic eruptions. There was practically no evidence for live, meteoric cratering events on the face of the Moon. Even the U.S. geologist who had been one of the first to argue convincingly (in 1893) for impact craters had disputed the impact origin of the world's best authenticated meteoritic crater, in Arizona. Until recent years, few craters on Earth had been described as of probable impact origin. Many structures on Earth now regarded as impact craters had been branded cryptovolcanic craters by geologists. The predominantly circular shape of lunar craters was thought to be impossible to produce by impact, except from relatively rare, vertical impacts. It is now believed that a nearly circular crater would be blasted out of the lunar rocks by a large meteorite arriving at a high speed even if the angle of approach to the Moon's surface were far from the vertical.

This, essentially, is the basis of the modern impact theory. A vast array of statistical data on both terrestrial and lunar craters has been collected, compared, and analyzed by the U.S. geologist Ralph Baldwin, who developed a thesis on the assumption that the great majority of the lunar craters were of impact origin, argued for the cataclysmic formation of Mare Imbrium following a major collision between a large object and the Moon that resulted in a gigantic explosion. Later, it was reasoned that the nearly circular maria, all the large craters, and the majority of the small ones were impact scars and that later lavas, flooding impact basins such as that of Mare Imbrium, flowed further afield than the confines of the basins and created the irregular maria. Later impact theories differed considerably from Baldwin's in some details.

Impact velocities

The velocities of interplanetary meteoroids in the vicinity of the Moon are commonly of the order of tens of kilometres per second. Because, in such a meteorite-impact process, the energy is transformed and released very rapidly, little of the initial kinetic energy is converted into heat, and the rocks in the floor of a crater are not likely to be melted to any appreciable extent. A longer period of time for the impact is needed to give the Moon rocks time to absorb more of the heat that would otherwise be dissipated in other ways.

A ring 1,000 kilometres (600 miles) in diameter (such as Mare Imbrium; Figure 1) could be formed by impact of a high-velocity meteorite of less than one-tenth of that diameter or with a larger, lower velocity object. Asteroids, for instance, would collide with the Moon at relative velocities around six kilometres (four miles) per second. Objects circling the Earth in a ring at about the Moon's distance could collide with the Moon at velocities as low as the velocity of escape from the Moon's surface, 2.4 kilometres (1.5 miles) per second.

The Moon was nonspherical. It had a nonequilibrium shape (even allowing for the Earth's tidal effects) and a peculiar mass distribution. The most important irregularity in the Moon's distribution of mass was found in the excess mass in or around the axis directed toward the Earth. Support of the excess mass in the Moon's Earthward-pointing bulge implied considerable mechanical strength. If the Moon had undergone any appreciable melting phase in the past, it would still have had insufficient strength in its deep interior to support the bulge. This is regarded by some scientists as an indication that the Moon could never have been melted to any significant

Evidence against large-scale melting

extent. The mare lavas would have been derived from the impacts themselves, and, hence, the impacting bodies must have been large, low-velocity objects. The observed asymmetric pattern of radial grooves and ridges beyond the periphery of Mare Imbrium could be generated by splashlike ejecta following an oblique, low-velocity collision, whereas a high-velocity collision at a similar angle of impact would produce an ejecta pattern essentially symmetrical about the explosion centre. It is to be pointed out, however, that one prediction of this theory was that the Moon had not undergone widespread igneous differentiation, whereas such differentiation is now known to have occurred.

Modern views on the origin of the craters have been much influenced by Baldwin's statistical work. He showed that the ratio of the depth to the diameter of terrestrial shell, bomb, and meteoritic craters and of the craters of the Moon could all be fitted on smooth curves showing logarithmic plots of the ratio against the diameters of the craters. Terrestrial calderas (collapsed volcanic structures) also fit the depth-to-diameter curves of Baldwin. Further studies extended Baldwin's data, relating the ratios of rim widths of lunar craters to their diameters to that for volcanic craters and calderas on Earth; the six calderas studied had much wider rims than lunar craters of the same diameter (1–10 kilometres), though this may be because of underestimations of the lunar rims.

The shape of plots of the logarithm of cumulative number of craters larger than a certain diameter against the logarithm of that diameter has been said to be evidence for impact origin of the craters; probable amounts of meteoroidal debris in the solar system would collide with the Moon to produce craters giving a similar statistical result; that is, the numbers of lunar craters of various diameters seem to be relative to the numbers of meteoroidal bodies of various sizes, in space, available to produce the craters. Other workers, however, found that a high proportion of sub-kilometre-sized craters were not impact craters and that the proportion varied with position.

There is now no real doubt that both impact and volcanic features exist on the Moon. Many details and some major problems remain to be solved. The origin of the circular maria, as has been suggested, may be that basins were created by impacts and were later flooded by lavas from within the Moon; the same argument may apply to many of the larger craters, but, if true in some cases, there are still exceptions. For Mare Orientale, there is much to be said for an explosive origin, since, for one thing, many characteristics of the surrounding terrain may be so explained. Other maria may be predominantly lava-filled calderas of igneous origin. The maria differ in age and were not all formed in one short-lived catastrophe.

IV. The origin and evolution of the Moon

THE EVOLUTION OF THE MOON'S ORBIT

The sea tides raised on the rotating Earth by the gravitational attraction of the Moon and the Sun make an important contribution to the transfer of the Earth's angular momentum to the Moon. Tidal friction is slowly reducing the Earth's rate of rotation: its day is increasing by about 1.8×10^{-3} second of time per century. Some of the energy of the Earth–Moon system is dissipated as heat. Energy is exchanged with the Earth–Moon system by other bodies, the Sun in particular.

Most of the angular momentum lost by the Earth is transferred, by gravitational interaction, to add to the Moon's orbital angular momentum, and this tends to accelerate the Moon. As a consequence, the Moon is lifted into a slightly higher orbit, farther from the Earth, and appears to slow down relative to the fixed stars. This deceleration may be measured by observing the position of the Moon's mean longitude relative to the stellar background. It is common practice to refer to the secular acceleration of the Moon, preceding the numerical value by a negative sign, and estimates range from −40 arcsec/century^2 (40 seconds per century per century, deduced from observations of solar eclipses) to −10 arcsec/century^2. The difference is actually double rather than quadruple because the astronomical value refers to the coefficient T^2 in the equation for mean longitude; the value −10 is thus exactly one-half the larger determination.

Secular acceleration of the Moon

The classical works of the English scientist George Darwin, as recently refined, show that, if the presently observed motion of the Moon is taken as a basis for examining its past behaviour, the Moon was once very much closer to the Earth and was in an orbit more highly inclined to the ecliptic than is its present orbit (which makes an angle of only a little over 5° with the ecliptic). Some 2×10^9 years ago the Moon would have been at a distance of 2.8 Earth radii, revolving around the Earth—which would then have had a five-hour day—in 6.5 hours in an orbit of inclination of about 46° to the ecliptic.

Such calculations, involving the gravitational interaction of more than two bodies, may now be performed effectively using modern computing techniques, but the results relating to large time intervals in the past or future are critically dependent on the assumptions made before the problem is processed. The older observations of the Moon's motion lack high accuracy. A homogeneous set of modern occultation observations is not sufficiently dispersed in time to discover any variability of the secular acceleration. There is still some uncertainty about the respective roles of shallow and deep seas in the tidal-friction problem and about the role of terms from small-scale irregularities in the Earth's gravitation potential regarding the dissipation of the angular momentum. Although, in some calculations of the past history of the Moon's orbit, the Earth's viscosity and moment of inertia (resistance to change in rotation) were assumed constant, the thermal evolution of the Earth, the progressive separation of materials to form the core and the mantle, and the changing positions and areas of the continents and seas must have changed these important parameters. Most scientists now regard the oceans to be a comparatively late acquisition of the Earth.

For reasons such as these, it is clear that it would be unwise to rely too heavily on extrapolations of the Moon's presently observed orbital behaviour in speculating on the origin of the Moon itself. It may be safe, on the other hand, to say that the Moon seems to be receding from the Earth at the present time and that the Earth–Moon distance may have been considerably less in the distant past. It has been suggested that the Moon, not being a fluid, would not have undergone disruption by the Earth's gravity even if the Moon had come within the Roche limit; that is, the distance within which breakup can occur if the body's binding strength is not enough to prevent it. At such a time the inclination of the lunar orbit would have progressed back through 90°, and thus the orbit would have been a retrograde one. Still further backward extrapolation, if correct, leads to the notion that the Moon has been captured by the Earth.

Changing distance of the Moon

A more cautious view is that the Moon might once have been within several Earth radii of the Earth's centre. Suggestions that the Moon separated from the Earth at a very early stage in the history of the solar system were revived in the 1960s. The weakness of such theories is that they are essentially attempts to explain the mechanics of the hypothetical separation process without regard for the geological evidence against them.

EVIDENCE FROM THE COMPOSITION AND PHYSICAL PROPERTIES OF THE MOON

Chemical analysis of lunar samples. The first orbital measurements of the broad composition of the rocks covering the Moon's surface served to confirm the speculations of photogeologists, who had argued for the presence of basic rocks, such as the mare basalts.

Detailed and more accurate measurements of the compositions of lunar samples were made *in situ*, for the first time, through the use of instruments carried by soft-landing vehicles beginning in 1967. One such craft landed in Mare Tranquillitatis, another in Sinus Medii, and another on the edge of a flow unit in the highlands near Tycho. Their findings indicated compositions close to those of some basalts on Earth, but there were differences between

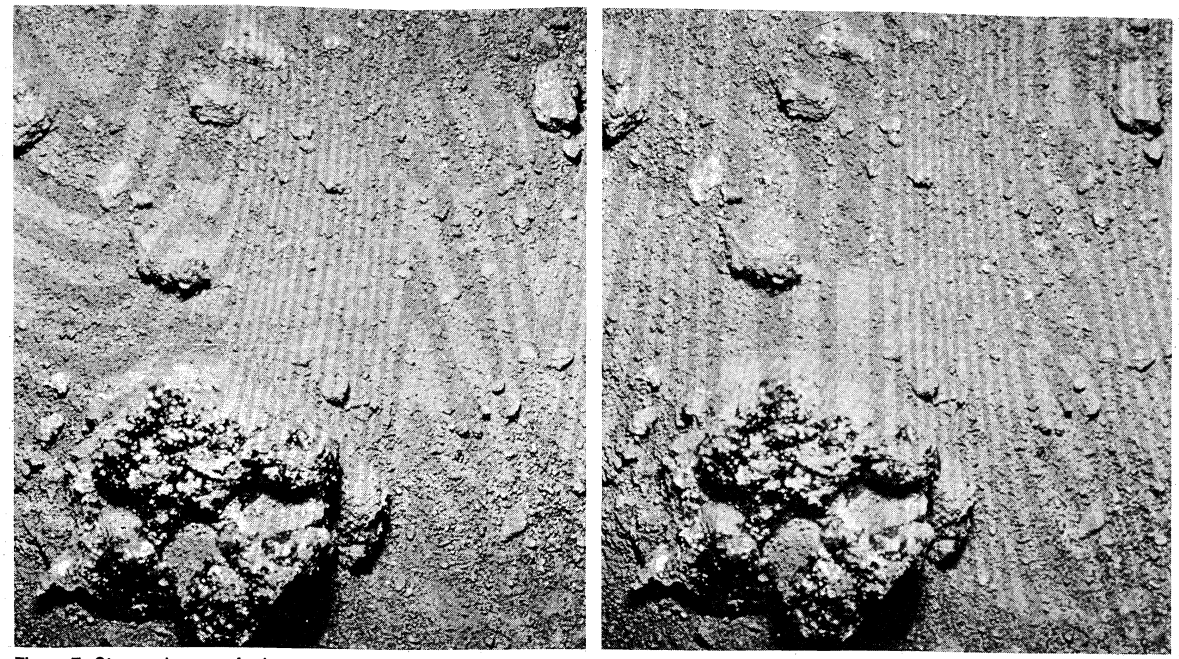


Figure 7: Stereo closeup of a lunar rock taken during the Apollo 12 EVA (extravehicular activity) on the lunar surface. A long handle on the camera permitted the astronaut to position it for surface object photography without stooping. The image scale is half actual size.
By courtesy of National Aeronautics and Space Administration

the highland analyses, on the one hand, and the earlier analyses in two patches of the lunabase. In the highland site, undisturbed lunar soil, disturbed soil, and a small rock block were analyzed separately. The rock block had a chemical composition that closely paralleled that of the undisturbed soil except for the elements magnesium and iron, which were slightly less abundant in the rock, and aluminum, which was slightly more abundant. Taken as a whole, the results indicate that the Tycho flow unit is compositionally similar to anorthosite, a granular igneous rock on Earth.

The results of the soft-landing program led many authors to suggest that the lunar highlands in general were higher in aluminum and calcium and lower in the iron-group elements (those with mass numbers between about 47 and 65) than the maria. More iron in the maria could contribute to their generally darker tones and could also contribute to the possibly higher rock density of the mare units.

Chemical differentiation

By 1968, scientists were beginning to appreciate the fact that the Moon had more in common with the Earth than many of them had been prepared to accept previously. It appeared that the Moon had suffered some degree of chemical differentiation and that its face presented not only physiographic domains but also true geological domains (the highlands and the lunabase) that differed in density. Furthermore, if these areas were to be established as separate geological domains, it would be clear that the Moon had been differentiated and heated. The evidence for the mare–highland rock contrast, however, was slender. First, it was based on only one analysis in the highlands. Second, the rocks analyzed were not necessarily typical highland material, especially in view of the fact that the vehicle had landed on some sort of flow unit. Third, two independent methods were used to determine the iron content of the soil in three cases, and, in the last, the methods gave discordant results. One method seemed to show that iron at Tycho was present in only half the concentration of that at the two mare-type Surveyor sites, while the other indicated there was essentially no difference in iron concentration among the three sites. One possible reason for this discrepancy is that the two methods were used to examine different samples of Moon rock. Magnetic experiments indicated that the lunar samples were of basalt with little (about 1 percent by volume) or no addition of meteoric nickel–iron. This level of iron is less than would be expected if meteoroids were the only agent responsible for breaking up and gardening the lunar rocks.

The first samples of Moon rock (Figure 7) to be returned to Earth for exhaustive study were collected in Mare Tranquillitatis in July 1969. Since that time, rocks from Oceanus Procellarum, Fra Mauro highlands, and Hadley Rille, in Mare Imbrium, have been returned to Earth and analyzed by scientists of several nations; and rocks collected and returned to Earth automatically from Mare Fecunditatis have been analyzed in the Soviet Union.

All the rocks (Figure 7) are volcanic basalts (a type of rock formed by solidification of magma) except for small meteoric (~1 percent), anorthositic (granular igneous rock), and glassy components in the surface debris. The coherent, basaltic rocks subdivide into those of vesicular, crystalline type and breccias (consisting of compacted regolith materials) formed from the regolith materials, although the breccias and regolith contain fractions of meteoric nickel–iron, anorthosite, and glass, and, among themselves, glass particles show a range of composition which includes that of basalt.

Evidence that meteoroids have been an important factor in shaping the Moon is found in the shocked minerals present in the returned rocks. Many authors consider that the compositional similarity between the anorthosites and the rocks of the Tycho highlands points to the anorthosites and some of the shocked materials found in the mare regions having been ejected from the highlands. The glasses that have undergone selective volatilization and that range widely in overall composition have probably derived from the melting of materials at a variety of impact sites. Certain glass globules of more uniform composition may be volcanic in origin.

Density at different depths

The astronomically determined mean density of the Moon as a whole is 3.34 grams per cubic centimetre, and the mean density of the basalts returned from the Moon's surface has a similar value. Yet the Moon cannot be composed of these mare basalts all the way through, since, at the pressures existing at moderate depths in the Moon, such materials would suffer a phase change and transform to eclogite, which has a density of 3.7 grams per cubic centimetre. One hypothesis is that the lunar highlands are composed of anorthosites of density between 2.8 and 2.9 grams per cubic centimetre, and certainly the bulk of the Moon's outer parts must be of rock with lower density than basalt. Yet, by comparison with

the average chondrite (the most common type of meteorite), the returned lunar basalts are, generally speaking, enriched in silicates and oxides and depleted in iron-group elements, and this is consistent with the theory that the Moon has undergone general magmatic differentiation that might have led to the formation of a nickel–iron core. Dynamical considerations limit the mass of iron in any lunar core to 6 percent of the Moon's mass, so that the core, if it exists, must be very small compared to that of the Earth.

Differences in the abundances of the element are observed at different lunar sites. Some differences could be related to crystal fractionation (crystallization of successive minerals with falling temperature) in near-surface melts, and, if the final products of fractionation had neared the lunar surface, the residual liquid rocks would have crystallized as granites. A few such fragments have been reported. In the rocks from Mare Tranquillitatis, titanium (occurring principally as ilmenite, $FeTiO_3$) was found to be much more abundant than in either chondrites or most terrestrial rocks. Evidently, the lavas in Mare Tranquillitatis contain more titanium because they were extruded at a stage in the fractionation process when certain elements, including titanium, were enriched in the melt.

Origin of mare lavas

Formation of the maria and highlands. The geological evidence for magmatic differentiation, partial melting, and crystal fractionation seems to rule against the theory that the mare lavas were generated by impact; rather, it seems that they came from deep within the Moon. Furthermore, the lavas were extruded at widely different times. Standard rock-dating methods show that the crystalline rocks from Mare Tranquillitatis last solidified 3.7 $\times 10^9$ years ago, whereas those from Oceanus Procellarum crystallized between 2.8 and 3.4 $\times 10^9$ years ago. One piece of rock returned in 1971 from the lunar highlands is reported to be 4.1 $\times 10^9$ years old.

Stratigraphers have attributed a greater age to the lunar highlands, relative to the mare units. Consequently, if this age were correct and representative, the lunar highlands would be very much older than any of the Earth's dated rocks—the oldest meteorites. The crater counts in distinct geological units suggest that parts of Tycho and Aristarchus are considerably younger than the surrounding materials, so there may be a dispersion of ages in highland material as well as in mare material. The highlands were probably formed overall, however, in the first few hundred million years of the solar system; whereas in the case of the mare lavas at least some of them were extruded and flowed across the surface after a further interval of hundreds of millions of years.

The thermal history of the Moon. Certain radioactive isotopes of the elements potassium, uranium, and thorium are, together, the most important source of radioactive heating in the Moon, and more of these elements occur in the lunar rocks than in the chondrites, possibly through fractionation. The Moon, if built out of chondrite material, was probably at least partly melted, over an epoch centring on about 2 $\times 10^9$ years after its formation. The extent of the melting would depend on the initial temperature of the Moon, but the period of maximum melting would appear to approach or coincide with the outpourings of the mare lavas.

The observed concentration of radioactive elements in returned lunar basalts, as well as the geological evidence, establishes the fact that the Moon underwent widespread melting. It has been argued that, because this melting occurred early in the Moon's history, the initial temperature of the Moon was high. The gravitational energy of accumulation of the Moon provides one early source of heating. Another source is the transfer of tidal energy to the solid body of the Moon by the Earth when the two planets were closer together, but it is uncertain whether the highest temperatures produced by such transferral of energy would be at the centre of the Moon or in its outer layers.

Possible sources of melting heat

A more complete picture of the Moon's internal composition and structure is needed to clarify details of the Moon's thermal history and origin. Surface faults and the general tectonic pattern on the Moon's surface demonstrate that the Moon has undergone general adjustments at least in its outer parts. Meteoric impacts and at least one deep-focus moonquake have been deduced from records made by instruments left on the Moon. The records are difficult to interpret, because the maria where the first instruments were landed are good scatterers of seismic (earthquake-like) vibrations that, nevertheless, are damped out much more slowly than seismic signals on Earth. Major structural discontinuities, such as are found in the Earth's interior, have not yet been found in interior of the Moon.

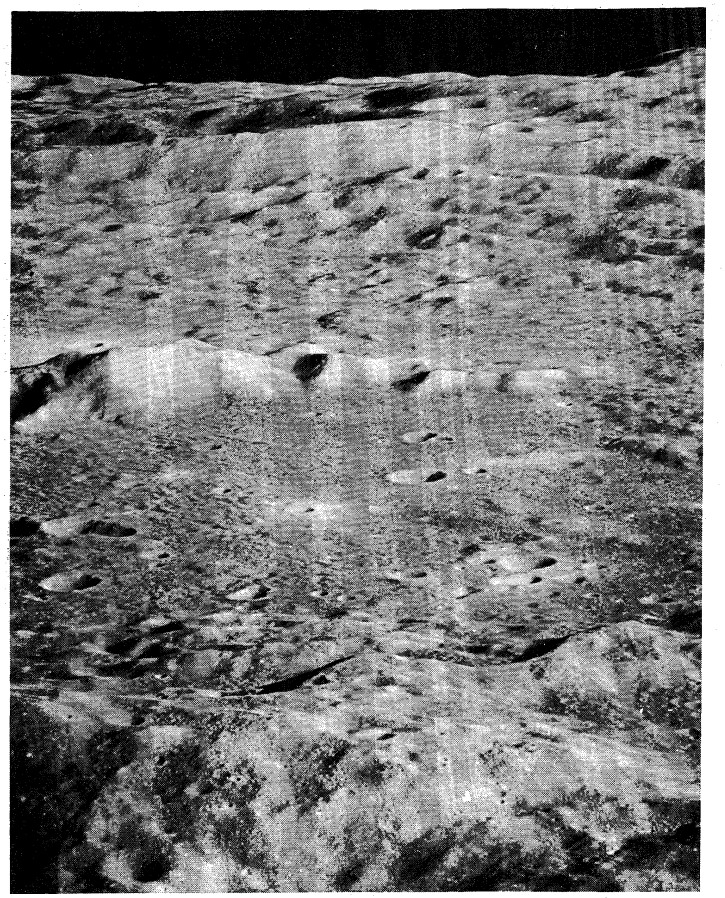


Figure 8: Crater on the Moon's far side with characteristic peaks and terraced walls.
By courtesy of National Aeronautics and Space Administration

Correlation of physical and chemical data. Remanent magnetization has been found in lunar crystalline-rock samples. Further studies of the specimens may throw some light on the origin of the magnetization, which may be associated with the Moon's core. Evidence drawn from surface instruments has been combined with that derived from other sources to deduce the bulk electrical conductivity and temperature profile of the lunar rocks at different depths. Although the electrical conductivity of rocks has a low value near the surface of the Moon, it increases to a depth of the order of 200 kilometres (120 miles). A mantle with the conductivity of basalts and a core with the conductivity of olivine-like materials would be consistent with these observations. The precise temperature of the core is not yet decided; the core's temperature must be related, for example, to the initial temperature of accumulation of the Moon and to the extent and rate of redistribution of radioactive elements in the upper layers of the Moon.

A picture of the Moon's evolution and origin is now beginning to emerge from the mass of extremely varied data. The elemental abundances of the stony meteorites and of the rocks of the Earth seem to be much the same, but they differ from the overall abundances of elements on the Moon. The Moon may be not a captured body but one formed in the immediate vicinity of the Earth during

an era in which planets were forming by the accumulation of particles.

THE "CIRCUMTERRESTRIAL SWARM" THEORY

The Moon's origin from a "circumterrestrial swarm" was proposed in 1950 and later modified. Bits of dust, meteorites, and larger planetesimals are believed to have come together to form the primitive Earth and a surrounding ring of debris. The metallic iron content of the swarm of material was accumulated by the Earth, whereas silicates tended to evaporate in the Earth's extensive atmosphere. The atmosphere dissipated; the silicate phase precipitated and accumulated to form the Moon. In this way the Moon accreted material of lower mean density than that of the Earth. The bulk of the lunar iron may now be found in a small lunar core. The lunar basalts are depleted, relative to chondrites and terrestrial basalts, in water and volatile elements that might have dispersed prior to accretion or as a result of the general outgassing of the Moon's rocks following intensive phases either of early bombardment or of later volcanism.

Nearly all the gases of the lunar atmosphere soon escaped, even perhaps as the Moon was still accumulating, the exact times depending on such factors as the mass of an atom of gas, the temperature of the upper layers of any temporary atmosphere, and the velocity of escape of a particle from the Moon at that level. Any secondary atmospheric gases (those derived from within the Moon) present on the Moon now or any gases remaining from encounters with comets would be composed of the heaviest gases—such as argon. Rare gases are found to be concentrated in the lunar soils and breccias; these gases are believed to have originated from the solar wind or solar cosmic rays. (G.Fi.)

BIBLIOGRAPHY. Basic data may be found in the following works: B.G. MARSDEN and A.G.W. CAMERON (eds.), *The Earth-Moon System* (1966); WILLIAM M. KAULA, *An Introduction to Planetary Physics* (1968); C.W. ALLEN, *Astrophysical Quantities*, 2nd ed. (1955); and ZDENEK KOPAL, *The Moon* (1969).

The actual motion of the Moon: ARTHUR BERRY, *A Short History of Astronomy* (1898, reprinted 1961), a good historical treatment through G.W. Hill; ERNEST W. BROWN, *An Introductory Treatise on the Lunar Theory* (1896, reprinted 1960), the standard text on the subject, containing details of all of the principal gravitational treatments from Newton through Hill; G.W. HILL, *The Collected Mathematical Works*, 4 vol. (1905–07), various aspects of the details of the algebraic treatment of the problem; ISAAC NEWTON, *Philosophiae naturalis principia* (1687; rev. Eng. trans. by FLORIAN CAJORI, *Mathematical Principles of Natural Philosophy*, 2 vol., 1969), details the mathematical analogy between the Moon and terrestrial objects that led to universal gravitation; J.D. MULHOLLAND, *Proceedings of the JPL Seminar on Uncertainties in the Lunar Ephemeris* (1968), and *Observation, Analysis and Space Research Applications of the Lunar Motion* (1969), the only recent summaries.

Data from artificial satellites of the Moon: G.A. CHISNALL and GILBERT FIELDER, *Astronomy and Space Research* (1962).

Types of lunar features: GILBERT FIELDER (ed.), *Geology and Physics of the Moon* (1971), *Lunar Geology* (1965), and *Structure of the Moon's Surface* (1961); JOSIAH E. SPURR, *Geology Applied to Selenology:* vol. 1, *The Imbrium Plain Region of the Moon*, vol. 2, *Features of the Moon*, vol. 3, *Lunar Catastrophic History*, and vol. 4, *The Shrunken Moon* (1944–49); DINSMORE ALTER, *Pictorial Guide to the Moon*, 2nd ed. (1967); ZDENEK KOPAL, JOSEF KLEPESTA, and THOMAS W. RACKHAM, *A New Photographic Atlas of the Moon* (1971); G.P. KUIPER (ed.), *Photographic Lunar Atlas* (1960, with suppl., *Orthographic Atlas of the Moon*); PAUL D. LOWMAN, JR., *Lunar Panorama* (1969); E.A. WHITAKER *et al.*, *Rectified Lunar Atlas* (1963); ERNEST W. BROWN (*op. cit.*); ZDENEK KOPAL and Z.K. MIKHAILOV (eds.), *The Moon* (1962); ZDENEK KOPAL (ed.), *Physics and Astronomy of the Moon*, 2nd ed. (1971); G.P. KUIPER and B.M. MIDDLEHURST (eds.), *Planets and Satellites of the Solar System*, vol. 3, ch. 9 (1960); JOHN W. SALISBURY and PETER E. GLASER (eds.), *The Lunar Surface Layer* (1964); WILMOT N. HESS, DONALD H. MENZEL, and JOHN A. O'KEEFE (eds.), *The Nature of the Lunar Surface* (1966); J.E. BALDWIN, "Thermal Radiation from the Moon and the Heat Flow Through the Lunar Surface," *Mon. Not. R. Astr. Soc.*, 122:513–522 (1961); ROBERT HOOKE, *Micrographia* (1st ed., 1665; 2nd issue, 1667, reprinted 1938); JAMES NASMYTH and JAMES CARPENTER, *The Moon*, 4th ed. (1903); WILLIAM H. PICKERING, "Lunar and Hawaiian Physical Features Compared," *Mem. Am. Acad. Arts Sci.*, 13:151–182 (1908); J.E. SPURR (*op. cit.*); G.K. GILBERT, "The Moon's Face," *Bull. Phil. Soc. Wash.*, 12:241–292 (1893); RALPH B. BALDWIN, *The Measure of the Moon* (1963) and *A Fundamental Study of the Moon* (1965); BRIAN MASON and WILLIAM G. MELSON, *The Lunar Rocks* (1970); G.M. BROWN, "Geochemistry of the Moon," *Endeavour*, 30:147–151 (1971); OTTO SHMIDT, *A Theory of Earth's Origin*, 3rd ed. (1958; orig. pub. in Russian, 3rd ed., 1957); R.W. FAIRBRIDGE (ed.), *The Encyclopedia of Atmospheric Sciences and Astrogeology*, pp. 887–890 (1967).

Chemical analysis of lunar samples: The results of intensive studies of returned lunar rocks have been published in *Science*, vol. 167, no. 3918 (1970); *Geochimica et Cosmochimica acta*, special supplement 1: *Apollo 11 Lunar Science Conference Papers*, 3 vol. (1970); *Apollo 12 Lunar Science Conference, Houston, January, 1970* (1971); and by BRIAN MASON and WILLIAM G. MELSON (*op. cit.*).

(G.Fi./J.D.Mu.)

Moore, Henry

Henry Moore is a British sculptor who represents the continuance of a vital tradition of humanist sculpture into the 20th century. In his early career, before 1940, he seemed to be a revolutionary, introducing new forms and new images into sculpture, deeply involved in the Abstract and Surrealist movements. With World War II, however, the humanism of Moore's sculpture became dominant, and his shelter drawings of Londoners sleeping in the Underground during air raids brought him new popularity and the status of a modern artist who could speak directly and powerfully to ordinary people. In the postwar years Moore was pre-eminent among living sculptors, with an international reputation based on work that seemed to use a new sculptural language to express ideas common to humanity since time immemorial.

© Gisele Freund

Moore in his studio, mid-1960s.

Background and education. Henry Moore was born July 30, 1898, at Castleford, a small coal-mining town near Leeds, Yorkshire, in the north of England. He was the seventh child of Raymond Spencer Moore, a Lincolnshire man of Irish ancestry, and his wife, Mary Baker, who came from Staffordshire, in the English Midlands. Moore's father came from farmworker stock, but for most of his life he was a coal miner. He was a self-educated man and an active Socialist and trade unionist.

A clever child, Moore won a scholarship to the Castleford Grammar School, where he studied from 1909 to 1915. He was helped and much encouraged by the art instructor Alice Gostick, evidently an exceptional teacher who had several gifted pupils. Already ambitious to become a sculptor, the young Moore acceded to his father's wish that he should first train to be a schoolteacher, like his older brother and sisters. For several months he did

practice teaching at his old primary school in Castleford, a necessary preliminary to entering a teachers' training college.

Service in World War I

Because of World War I such plans had to be postponed, and in February 1917 Moore joined the British Army as a private in the 15th London Regiment. He was sent to France in the summer of 1917 and stationed in a quiet sector of the Western Front near Arras. Later in the year he was transferred to the Cambrai section, where the fighting was heavy, and, after an intensive bombardment, Moore suffered from the effects of gas shells. He collapsed and was sent back to England for two months' hospital treatment. After further convalescence he made a complete recovery and became an instructor in physical training until his return to France in November 1918, the month of the Armistice.

Moore left the army in February 1919 and went back to teaching small children. But in September he was given a rehabilitation grant, to which he was entitled as a former serviceman, and he went to the Leeds School of Art, where he studied for two years. In his first year at Leeds, Moore spent most of his time studying drawing. Although he wanted to study sculpture, no teacher was appointed until his second year; and Moore was his first pupil. He was soon joined by a very young student from nearby Wakefield, Barbara Hepworth, who also was to fulfill an early childhood ambition and become a major British sculptor.

As a result of his contact with the flourishing artistic life in Leeds, Moore's intellectual horizons slowly began to broaden. He was excited by the modern paintings that he saw in the private collection of the vice chancellor of the University of Leeds, Sir Michael Sadler, who was responsible for the first English translation (1914) of the classic manifesto of abstract art, *Concerning the Spiritual in Art*, by Wassily Kandinsky, a Russian abstract painter.

At the end of his second year at Leeds School of Art, Moore passed the sculpture examination and was awarded a Royal Exhibition scholarship to study at the Royal College of Art. In September 1921 he moved to London and began three years of advanced study in sculpture; he took his diploma at the Royal College after two years and spent a third doing postgraduate work. Moore found a good friend and lifetime supporter in the director of the Royal College, William Rothenstein, who was not unsympathetic to modern artistic tendencies, although he remained a conservative himself.

Studies at the Royal College of Art

Instruction at the Royal College of Art was less important to Moore than the opportunity of private study in the museums of London—particularly in the British Museum, with its wide-ranging and unrivalled collection of ancient sculpture. There was also close at hand the fine collection of Auguste Rodin sculpture at the Victoria and Albert Museum, but Moore was already reacting against the European sculptural tradition and turning instead to primitive and archaic art. He was discovering for himself the strange power and unusual beauty of Egyptian, Etruscan, and, later, Mexican and African sculpture.

Away from the college, spending vacations in Norfolk (where his parents now lived in retirement), Moore began to carve directly, in wood and especially in stone, making heads and small figures and animals.

Travel and further artistic influences. Upon graduating from the Royal College in 1924, Moore was appointed a part-time instructor in sculpture for a seven-year term. His exceptional gifts and potential stature were already recognized by those who knew him best. He was also awarded a travelling scholarship and spent the first six months of 1925 in France and Italy. He had already made the first of many visits to Paris in 1923, when he was particularly interested in the Paul Cézanne paintings of bathing figures that he saw in the Pellerin Collection. In Italy he stayed in Rome, visiting Florence, Pisa, Siena, Assisi, Ravenna, Padova, and Venice.

Back in England, Moore began work in 1926 on the first of his reclining women. He was also carving in stone: half-length female figures, mother and child groups, masks and heads. Though certain works show his awareness of the Romanian sculptor Constantin Brancusi and the Cubist sculptors, the most important influence on Moore's work at this time was that of ancient Mexican stone carving. In 1925, in the Trocadero Museum in Paris, he had been impressed by a plaster cast of the Chac Mool—a Mayan limestone sculpture of the rain spirit, depicted as a male reclining figure with knees drawn up together, staring head at a right angle to the body, and hands holding on the stomach a flat dish for sacrifices. Moore became fascinated with this sculpture, which seemed to him to have qualities of power, sensitivity, three-dimensional depth, and originality of form that no other stone sculpture possessed. Disdainful of conventional standards of the beautiful, seeking a way of imbuing his own work with such qualities, he changed the Mexican male figure into a female one, the better to express a more human, more earthy, more rhythmic image of his own.

Influence of ancient Mexican stone carving

Guaranteed a small income from part-time teaching, Moore was in no hurry and could absorb influences and work out his own ideas in relative tranquillity. In 1928 he was ready for his first one-man exhibition, at the Warren Gallery in London, and he began his first public commission, a relief carving of the "North Wind" on the new headquarters building for the London Transport Board —part of an ambitious decorative scheme to which Jacob Epstein and Eric Gill also contributed sculptures. In 1929 he married Irina Radetzky, of Russian–Austrian parentage, who was a painting student at the Royal College of Art. The young couple moved into a large studio in Hampstead, one of the northern suburbs of London. He now began a close association with a group of young artists who in 1933 formed Unit One, a deliberate attempt to make the indifferent English public aware of the international modern movement in art and architecture. The driving spirit behind Unit One was the painter Paul Nash, but the leading members were Ben Nicholson and his wife, Barbara Hepworth, who now lived in a neighbouring studio in Hampstead. Another neighbour, close friend, and advocate was the poet and critic Herbert Read, who wrote the first monograph on Moore in 1934.

Formation of Unit One

Achievements in the 1930s. Around this circle of friends the most advanced artistic activity in England in the early 1930s was centred. They were all interested in abstract art at a time when this was considered the ultimate in artistic extremism. In his own work from 1931 onward, Moore moved tentatively away from the human figure to experiment with abstract shapes and, more individually, to combine abstract shapes with references to the figure.

In 1931 Moore had the first of many one-man exhibitions in the Leicester Galleries in London. His work was enthusiastically introduced by Jacob Epstein, but it aroused violent criticism in the press and made Moore a notorious figure. He was urged to resign his position at the Royal College of Art, and, when his contract expired in 1932, he left to start a sculpture department at the Chelsea School of Art. There he taught part-time until the outbreak of war in 1939. Much of his more ambitious work was necessarily done during vacations, when the Moores lived in a country cottage, at Barfreston in Kent from 1931 to 1934, and then at Kingston near Canterbury, where Moore also had a large field that enabled him to work out of doors in the open landscape.

Throughout the 1930s Moore displayed in his work not the slightest inclination to please the public. He was very interested in Pablo Picasso's drawings and paintings of the late 1920s, which have strong sculptural implications, and felt free to distort and break up the forms of the body in a much more radical way than before. Sometimes he seemed to leave the human figure behind altogether. The pages of his sketchbooks of this period show him full of ideas for abstract sculptures that make use of organic and natural forms rather than pure geometrical shapes. He was collecting pebbles, rocks, shells, and bones, making drawings of them, studying them to find what he called "nature's principles of form and rhythm," which he sought to apply to his own sculpture. In particular, this meant opening up the carvings with concavities and even with holes pierced right through the forms—a practice

often found shocking and abhorrent when the sculpture retained a strong suggestion of the human figure.

Surrealist allegiance

The courage to experiment perhaps explains why Moore became a founder member of the Surrealist group in England and shifted his allegiance away from the more dogmatic abstract artists. He was not a polemicist, however, and thus did not engage in the disputes of the 1930s between the rival avant-garde factions of Surrealist and abstract artists. Without contentiousness and with great lucidity, Moore set forth his ideas about sculpture in several articles and reviews written in the 1930s and early 1940s.

Changes wrought by World War II. When the war broke out, Moore's whole pattern of existence changed. The Chelsea School was evacuated from London, and Moore stopped teaching. At first he worked mostly in his cottage in Kent, until its propinquity to the Channel coast, where invasion was hourly expected, forced a return to London. For a time he used one of the Hampstead studios that Barbara Hepworth and Ben Nicholson had evacuated on their move to Cornwall, but in October 1940 this was damaged by bombing, and the Moores took a house at Perry Green, Much Hadham, in Hertfordshire, which became their permanent home. There, in tranquil countryside about 20 miles north of London, he slowly added studios and extra rooms to an ancient farmhouse with wooden beams and low ceilings.

Shortage of materials in the early years of the war forced Moore to concentrate on small sculptures and then exclusively on drawing. Seeing the people of London sheltering at night in the stations of the London Underground in September 1940 led him to begin his shelter drawings. Moore would spend the night observing and making small sketch notes; then, in the next days at the Much Hadham studio, he would work his ideas up into large coloured drawings that expressed in permanent form the resigned but indomitable spirit of Londoners during the bombing of their city. While making the drawings, Moore was appointed an official war artist under a government scheme to employ artists to make a record of the war (not for propaganda purposes but as a historical witness). He also visited the colliery in Castleford, Yorkshire, where his father had worked, and made drawings of the coal miners at work, which have a strength and dignity similar to the shelter drawings.

Moore's wartime experiences led him to reconsider the relationship between the modern artist and his public. Instead of the widespread idea that there is necessarily a conflict, Moore came to believe that the artist should be a leader and that there was a public role for him to play. He agreed to become a trustee of the Tate Gallery in 1941 and has since served on the boards of numerous public institutions, including the Arts Council, the British Council, the National Gallery, the Royal Fine Art Commission, and the National Theatre—all of them state-financed institutions in which decisions are made by independent committees.

Church art

In 1943 Moore accepted a public commission to create a "Madonna and Child" for the church of St. Matthew in Northampton. His willingness to do this, although he was not a Christian himself, is further evidence of his feeling that the modern sculptor must be ready to meet a new audience. Moore has written about his initial uncertainty when asked to carve the "Madonna and Child," but the possibility of reviving the great tradition of religious art appealed to him, and he tried to give his figures for Northampton "an austerity and a nobility, and some touch of grandeur (even hieratic aloofness) which is missing in the everyday 'Mother and Child' idea."

Another commission, for a family group, followed in 1944, and the result was a dramatic change in Moore's style, away from the experimentation of the 1930s and toward a more naturalistic approach that had an immediate popular appeal. Moore had made dozens of studies in clay and terra-cotta when working on the "Madonna and Child" and family-group commissions, and these were cast in bronze and issued in editions of seven or nine copies each. In this way, Moore's work became available to museums and collectors all over the world.

This humanistic work of 1943–46 was the basis of Moore's international reputation, which dates from the large retrospective exhibition held in 1946 at the Museum of Modern Art in New York. On this occasion Moore visited the United States for the first time. American collectors began to buy his work, and henceforward he was at last freed of financial worries and able to work on the scale he felt his sculpture demanded. In Europe, too, Moore's reputation as an outstanding sculptor was confirmed when he won the sculpture prize at the 1948 Venice Biennale; more than 100 retrospective exhibitions followed in various countries.

In Britain, Moore fulfilled several commissions that extended the range and scale of his work: family groups for the new towns of Stevenage, Hertfordshire, in 1948, and Harlow, Essex, in 1954–55; "Three Draped Standing Figures" in stone (1947–48) for Battersea Park, London; a "Madonna" for St. Peter's church in Claydon, Suffolk, in 1949; and a large "Reclining Figure" for the 1951 Festival of Britain. The death of his mother, in 1944, and the birth of his only child, Mary, in 1946, made the theme of the family—particularly the mother and child relationship—a more personal one, expressive of Moore's deepest feelings. Although Moore himself neither encouraged nor endorsed psychological interpretations of his work, it is evident that on its most profound level it is a statement about nature and the qualities of the most basic human relationship—that of mother and child.

Later years. Critics who had begun to think by 1950 that Moore, the revolutionary sculptor, had been tamed, were to be proved wrong by the appearance in that year of the bronze standing figures, with their harsh and angular pierced forms and distinct feeling of menace. When, in the summer of 1953, Moore was ill, his work began to turn inward, showing a willingness to experiment and to follow private concerns. A large marble carving he made in 1957–58 for the headquarters of UNESCO in Paris belongs to a long series of reclining female figures, but the brick sculpture relief made in 1955 for the Bouwcentrum in Rotterdam reintroduces biomorphic forms into his work, and this leads to the series of freestanding totemic upright figures made in 1955–56. Moore also varied his subject matter in the 1950s with such works as the "King and Queen" of 1952–53, and the two warriors, the "Warrior with Shield" of 1953 and the "Fallen Warrior" of 1956–57—rare examples of Moore's use of the male figure. All three works owe something to Moore's visit to Greece in 1951, when he saw Athens, Mycenae, and Delphi for the first time. Most of his sculpture since the war was in bronze, though he had not altogether stopped carving in wood and stone. Furthermore, even when the sculptures were cast in bronze, they were not modelled in clay but built up initially in plaster over a wire and wood armature. Moore always liked to work like a carver, cutting and scraping and chiselling the surfaces with a carver's tools.

Turn toward private concerns

From the time of his 60th birthday in 1958, Moore seemed less concerned with his public role as a modern sculptor and more inclined to pursue his private interests. It is true that he continued to accept commissions, most notably for the Lincoln Center in New York in 1963–65 and for the University of Chicago in 1966; but in both cases, unlike earlier commissions, Moore made no attempt to provide a sculpture peculiarly appropriate for the site: he used the commission to work out on a larger scale than would otherwise have been possible, an idea that had long occupied his imagination. Thus, the Lincoln Center sculpture is the largest of a series of multipart reclining female figures in which Moore makes use of symbolic correspondences between the body and such elements of landscape as cliffs, caves, and hillsides and between the body and organic forms, particularly human and animal bones. Although the University of Chicago's "Atom Piece," with its mushroom-cloud formation at the top, commemorates the splitting of the atom, the sculpture is also closely related to other large abstract sculptures of the 1960s: the "Locking Piece" of 1963–64, the "Three-way Piece No. 1: Points" of 1964, the "Knife-Edge Two-Piece" of 1965, and the "Three-Piece Sculp-

ture No. 3: Vertebrae" of 1968—all of them massive objects that have lost their obvious human connotation as a consequence of their enormous size.

Some of the abstract sculptures from the mid-1960s and after were executed in marble rather than bronze. Beginning in 1965, Moore had a summer cottage at Forte dei Marmi, near the Carrara stone quarries, and, with the assistance of Italian workmen, began to make stone carvings again. At times he used marble of unusual colouring, notably a pinkish stone.

Moore has been much honoured, receiving honorary degrees from numerous universities, including Cambridge, Harvard, and Oxford. He was awarded the Companion of Honour in 1955 and the Order of Merit in 1963—the highest distinction that can be offered to a British subject. Yet he remained a forthright and simple man who never lost touch with his working class origins and continued to be proud of his miner and farm-labourer forebears. He established an unostentatious way of living, his daily routine consisting in working mornings supervising the activities of his small secretarial staff (which arranges sales and exhibitions, supplies information and photographs, and the like) and of the two or three young sculptors who help with the more laborious and time-consuming activities that the making of sculpture entails; in the afternoon, a leisurely lunch and then more supervision in the studios, for every stage is carefully controlled, and Moore's active participation is evident on even the largest sculpture; finally, whenever possible, an hour or more alone, working in a small studio, using pebbles and bones to try out on a minute scale ideas that may lead to nothing or may become the basis of a sculpture of considerable dimension. The only sign of wealth is his art collection, the work of artists whom he most admires—a small "Bathers" composition by Cézanne, figure drawings by Edgar Degas and Pierre Renoir, landscapes by Gustave Courbet, and sculptures by Rodin and the early Italian masters. In later life his taste became less purist than it was when he was a young man, for he came to be more aware of belonging to the sculptural tradition he had persistently reacted against that encompasses Michelangelo and Rodin.

MAJOR WORKS

Sculpture

MARBLE, STONE, WOOD AND LEAD, and WIRE: "Mother and Child" (stone, 1924–25; City Art Gallery, Manchester, England); "North Wind" (stone, 1928–29; St. James's Park, London Transport building, London); "Reclining Figure" (stone, 1929; City Art Gallery, Leeds); "Reclining Women" (stone, 1930; National Gallery of Canada, Ottawa); "Two Forms" (wood, 1934; Museum of Modern Art, New York); "Figure" (marble, 1937; Art Institute of Chicago); "Recumbent Figure" (stone, 1938; Tate Gallery, London); "Mother and Child" (wood, 1938; Museum of Modern Art, New York); "The Bride" (lead and wire, 1939–40; Museum of Modern Art, New York); "Madonna and Child" (stone, 1943; St. Matthew's, Northampton, England); "Memorial Figure" (stone, 1945; Dartington Hall, Devon); "Time-Life Screen" (stone, 1952–53; Time-Life Building, Bond Street, London); "Internal and External Forms" (wood, 1953–54; Albright-Knox Art Gallery, Buffalo, New York).

BRONZE: (most in multiple editions): "Family Group" (1948–49; Tate Gallery); "Helmet Head No. 1" (1950; Tate Gallery); "Reclining Figure" (1951; Musée National d'Art Moderne, Paris); "King and Queen" (1952–53; Tate Gallery); "Reclining Figure External Form" (1953–54; Art Institute of Chicago); "Mother and Child" (1953; Tate Gallery); "Warrior with Shield" (1953–54; Art Gallery of Ontario, Toronto); "Reclining Figure" (1956; Western Australian Art Gallery, Perth); "Falling Warrior" (1956–57; Clare College, Cambridge, Cambridgeshire); "Seated Figure Against Curved Wall" (1956–57; Museum of Fine Arts, Boston); "Reclining Figure" (1957–58; UNESCO, Paris); "Draped Seated Woman" (1957–58; Musées Royaux des Beaux-Arts, Brussels); "Two-Piece Reclining Figure No. 1" (1959; Albright-Knox Art Gallery); "Two-Piece Reclining Figure No. 2" (1960; Tate Gallery); "Standing Figure: Knife Edge" (1961; private collection); "Knife-Edge Two-Piece" (1962; Haags Gemeentemuseum, The Hague); "Locking Piece" (1963–64; Banque Lambert, Brussels); "Atom Piece (1964–66; University of Chicago); "Three-way Piece No. 1: Points" (1964; Columbia University, New York).

BIBLIOGRAPHY. The definitive catalog of Moore's sculptures is *Henry Moore: Sculpture and Drawings*—vol. 1, *Sculpture and Drawings, 1921–1948*, 4th ed. by DAVID SYLVESTER, with an introduction by HERBERT READ (1957); vol. 2, *Sculpture and Drawings, 1944–54*, 2nd ed., with an introduction by HERBERT READ and a catalog of the sculpture by ALAN BOWNESS (1965); vol. 3, *Sculpture, 1955–64*, 1st ed. by ALAN BOWNESS, with an introduction by HERBERT READ (1965). Moore's writings have been collected and published in *Henry Moore on Sculpture*, ed. with an introduction by PHILIP JAMES (1966). Important monographs are: ERICH NEUMANN, *The Archetypal World of Henry Moore* (1959); WILL GROHMANN, *The Art of Henry Moore* (1960); JOHN RUSSELL, *Henry Moore* (1965); and ROBERT MELVILLE, *Henry Moore Sculpture and Drawings: 1921–69* (1970).

(Al.Bo.)

Moravian Church

The Moravian Church, a Protestant communion tracing its origin to the Hussite Unitas Fratrum ("Unity of Brethren") in Bohemia and Moravia, is principally a product of German Pietism. Though numerically insignificant, its impact upon modern Protestantism has been significant in the areas of worship, evangelism, missions, and theology. A membership of 353,228 in 1969 included 21,558 in continental Europe (9,650 Czechs), 4,598 in England, and 65,459 in North America (4,074 Eskimos); while 32,987 are in the West Indies, 83,654 in Latin America, 144,842 in Africa, and 130 in Tibet and north India. Devotional cultivation and practical application of individual and communal "Jesus mysticism" is the distinguishing mark of the movement.

HISTORY

Origins. The Unitas Fratrum originated with the Hussite Reformation in 15th-century Bohemia. Jan Hus's teaching had a twofold emphasis: the first, national and ecclesiastical, seeking a distinctly Bohemian church marked by lay reception of cup as well as bread in the Eucharist—this represented by the Utraquist ("two kinds") party; the second emphasis, religious and ethical, aiming at ecclesiastical purification in accordance with primitive practice—this represented by the Taborites and Horebites until their destruction by the Utraquists in 1434 and 1452. The Unitas Fratrum, reaffirming the religious-ethical emphasis, was founded in the Bohemian village of Kunvald under the leadership of Gregory the Tailor, probably in 1458. Influenced by his uncle, Jan Rokycana, archbishop of Prague, Gregory studied the writings of Petr Chelčický, a puritan-primitivist Hussite. To Chelčický's Wycliffite Biblicism, Donatist conception of the ministry (*i.e.*, moral requirements for ministerial validity), and commitment to political noninvolvement, Gregory added a desire to experiment with a type of primitive Christian community. This emphasis upon common life continues to mark the Moravian Church.

Jan Hus

Composed of clergy and nobles, as well as laymen like Gregory, this congregation was bound together by acceptance of the Bible as the single rule of life. Suspension of their pastor Michael Bradacius, a Utraquist priest of the village, in 1460 prefaced the first persecution of the movement. Efforts to gain a hearing from the authorities led to the first official theological statements of the Unitas, which emphasized the imitation of Christ in gentleness, poverty, patience, and in the love of enemies.

More basic than theological concerns was the experiment taking place in disciplined communal life patterned on primitive Christianity. Supervised by "elders" with Catholic clerical standing, the Unitas at first remained within the Roman Catholic Church and was served by sympathetic Roman priests. Difficulty with this arrangement led to selection of their first priests in 1467. To assure a "pure" priesthood, ordination by Stephen, a Waldensian "bishop," was obtained.

Early popular success confronted this ascetic movement with the necessity to relate with "this world," particularly to review strictures against exercise of political authority. By 1500 the movement in Bohemia and Moravia numbered 100,000 members, including barons and scholars as well as peasants. By the end of the 15th century the formerly withdrawn ascetic community participated fully

in society. Accompanied by heated controversy and temporary schism, a more formal constitution replaced the earlier charismatic structures. Externally this success led from a generalized and rather sporadic persecution to the Edict of St. James (1508), under which the Unitas in Bohemia was to suffer for a century.

The Reformation. Immediately before the impact of the 16th-century Reformation, the Unitas, led by Luke of Prague (1460–1528), enjoyed a literary and theological revival. Yet initial contacts between Luther and the Unitas (1522–23) were marred by mistrust, misunderstanding, and divergent preoccupations leading not to bitterness but to indifference. Only under the leadership of Jan Augusta (1500–72), the "Bohemian Luther," did the 16th-century Reformation reach the Unitas as a religious impulse. The Habsburg goal of eradicating Protestantism and Bohemian nationalism, coupled with the impact of Luther in Bohemia, led directly to a vision shared by Augusta with the Evangelical Party of the Utraquists to create a Bohemian Protestant National Church. After 1532 cordial relations existed between the Unitas and Luther, and the Unitas Confession of 1535, introduced by Luther, was published in Wittenberg in 1538 as a sign of agreement. Under Lutheran influence the Unitas abandoned the observance of seven sacraments, rebaptism of converts, and mandatory clerical celibacy.

Luther

The Bohemian nobility was embroiled in the disastrous collapse of the Schmalkaldic League (coalition of German Protestant states) at Mühlberg in 1547, and, despite only marginal involvement, the Unitas and Augusta were singled out for royal retribution. The resulting persecution led to migration ending in Poland and contacts with Calvinism. In 1575, three years after Augusta's death, his unionist policies bore fruit in the Bohemian Confession, creating a National Czech Evangelical Church. With the Letter of Majesty (1609) this union finally gained toleration.

The period of religious peace in Bohemia was to be short. The Counter-Reformation, particularly that aspect led by Jesuit educators, made impressive gains against the Protestants; and demise of the Unitas in the lands of its origin came quickly with the early events of the Thirty Years' War. Support of Frederick V of the Palatine ("the Winter King") by the Protestant barons, and their common defeat at the Battle of the White Mountain in 1620, sealed the doom of Protestantism in Habsburg lands. In 1624 non-Roman Catholic groups were outlawed in Bohemia and Moravia, and by 1628 the last of the Unitas ministers were exiles. Among these was John Amos Comenius (1592–1670), last great bishop of the movement and grandfather of Daniel Ernst Jablonski (1660–1741), Prussian court preacher who in 1735 consecrated the first bishop of the Revised or Renewed Unitas Fratrum, better known as the Moravian Church.

The Counter-Reformation

The Moravian Church. Revival of the Unitas resulted from the German Pietist awakening. Stirred by Pietism, Kristián David (1690–1751) undertook missionary journeys into his native Moravia to contact families still identified with the Unitas. In 1722 a group of these Moravians under his leadership settled on the estate of Berthelsdorf at a place later called Herrnhut (from which Moravians derived the name "Herrnhuters"). Their encounter with Nikolaus Ludwig, Graf von Zinzendorf (1700–60), was to set the historical course of the renewed Moravians. Zinzendorf, trained under A.H. Francke at Halle, developed a communal piety centring in encounter and identification with the historical Jesus. The various religious groups settled in Herrnhut occasioned disorder until Zinzendorf took personal leadership of the colony, bringing it under a disciplinary structure patterned on the Unitas. A spiritual awakening during the eucharistic celebration of August 13, 1727, marks the birth and unification of the congregation around Zinzendorf's religious genius. Zinzendorf shaped the prosperous colonies of artisans gathered at Herrnhut and numerous other settlements into near-monastic communities with lives completely organized for glorification and service of God through a veritable maze of liturgical functions, communal organizations, and daily vocational activities. No art escaped mobilization by the creative Zinzendorf, but the key to all was the unique application of music and hymnody in worship.

Zinzendorf

Exercise of inner piety was not, however, the raison d'etre of these communities. Although driven at every turn toward a sectarian identity, Zinzendorf's ambitions were ecumenical. Awakening of all denominations in Christendom and extension of Christianity to the ends of the earth were the goals of Herrnhut. By 1732 mission work had begun both at home (so-called Diaspora) and abroad (Greenland and West Indies). Episcopal consecration from the Unitas Fratrum (1735) was obtained to facilitate foreign missions rather than to introduce a new denomination in Germany. Although large mission congregations were developed overseas, the tendency of the Diaspora policy was to hold people in their own churches and avoid development of a large membership in the Moravian settlements.

Ecumenism

Zinzendorf's time was a period of religious, theological, and organizational formation; but it was also an era marked by religious excesses and financial irresponsibility. The sensual aberrations of the Sifting Time (1740s), with its idealization of childlike behaviour and preoccupation with a sensual participation in the dying, burial, and bodily corruption of Jesus, were short lived but badly damaged the public image of the Moravians. Still further difficulties were occasioned by their leader's unrealistic fiscal policies. The second generation, under the leadership of August Gottlieb Spangenberg (1704–92), witnessed consolidation both in structure and public acceptance. The instability of one-man dominance was replaced by a representative constitution. Though still bearing its distinctive stamp, the church was freed of religious excess, both real and reputed. The movement was now generally accepted in both intellectual and ecclesiastical circles.

Influence. The primary influence of the Moravians was religious. John Wesley, the founder of Methodism, owed them much, both in his religious development and in the formation of his movement. Much less noted than this influence on Wesley is the impact of Moravian work in Europe on the 19th-century Protestant revivals. Aside from shock at the outcome of the French Revolution, a major factor in the religious awakenings that spread over Europe from the Baltic States to the Netherlands and from Switzerland to Scandinavia was the Diaspora activities of the Moravians. Reports from the far-flung missionary fields of the Moravians played a major role in influencing, instituting, and spreading Protestant missions.

The intellectual and theological contributions of the Moravians are also significant. Friedrich Schleiermacher (1768–1834), a father of modern Protestant theology, brought the heritage of Pietism to bear in theology. Reared under Moravian influence, he gained an enduring conviction of the "empirical" reality of religion and a highly developed taste and skill in its practice. His later claim to be a "Herrnhuter . . . of a higher order" recognized the biographical basis of his highly experiential theology. The Moravians made other significant contributions to the formation of Romanticism. J. Wolfgang Goethe (1749–1832), though not attached to the movement, was impressed in his Romantic inclinations by it. Also, the classic German Romantic Novalis (1772–1801) had been reared in a Moravian Diaspora congregation and was deeply marked by their approach to reality. Nor does the theological contribution of the Moravians end with Schleiermacher. Led by Spangenberg in response to the victories of rationalism, a theological synthesis of the centuries-old enemies—Pietism and Orthodoxy—took place, which spread through the Pietist revivals as the historic foundation for modern Fundamentalism.

Moravians since the 18th century. In North America the Movarians had difficulty implementing Zinzendorf's inner churchly ideals. Operating within a free-church context, they were pressed toward separate denominational identity. Missionary activity among American Indians was from the start a central and successful concern. After removal from Georgia to Pennsylvania (1740), the entire membership in North America was mobilized for this work. Under the so-called Economy (1748–62) centred in Bethlehem, Pennsylvania, all production was held in com-

mon and directed toward the missionary task. In 1753 a second Moravian centre in North America was established in Winston-Salem, North Carolina.

The Moravian Church at present operates in five continents. Still small in membership, it continues to wield great influence on Protestantism in Europe. The work in East Germany has been greatly impaired as a result of partition; but Herrnhut, located there, remains a chief European centre. Early and continually involved in missionary efforts at home and abroad, as well as in ecumenical aspirations, its role in the modern ecumenical movement is obvious. Beyond active membership in various national confederations of churches and in the World Council of Churches, the Moravians—particularly in Europe—have been promoters and leaders in such ecumenical endeavours as the *Kirchentag* ("church day," on which clerics and laity meet to discuss religious and social issues) and Evangelical Academy movements.

BELIEFS, PRACTICES, AND INSTITUTIONS

Despite some rather unusual theological statements, Zinzendorf's intention was to maintain the Moravian movement's Lutheran identity in respect to Scripture and creed. His "Christocratic" keynote never attained creedal formulation, expressing itself rather in hymnody and liturgy. Influenced by the theological movements associated with Schleiermacher, Albrecht Ritschl (1822–89), and finally the "Theology of Crisis" of Karl Barth (1886–1968) and Emil Brunner (1889–1966), there has been a revival of Zinzendorfian thinking in recent decades.

Since the middle of the 19th century Moravian polity has been extremely loose. Provinces (jurisdictional units) are virtually autonomous, representing the common interests of individual congregations. The Moravian episcopacy is spiritual, not administrative. Bishops elected by periodic synods on the basis of spiritual merit are pastoral, even prophetic, not presidential.

Moravian uniqueness lies in its liturgy. A body of hymns and anthems emphasizing the life and passion of Christ is utilized to illustrate, supplement, and even replace spoken discourse and prayer. This idiom completely dominates eucharistic and Love Feast celebrations, where only a few words are spoken.

BIBLIOGRAPHY

Doctrine, polity, and worship: W.A. ALLEN, *Who Are the Moravians?* (1966); A.L. FRIES, *Customs and Practices of the Moravian Church,* rev. ed. (1962), two popular works.

History: For information on the early history, see E.A. DE SCHWEINITZ, *History of the Church Known as Unitas Fratrum* (1885), old but useful; and J.T. MULLER, *Geschichte der Böhmischen Brüder,* 3 vol. (1922–31), definitive classic.

Books on the modern history include: J.T. and K.G. HAMILTON, *History of the Moravian Church: The Renewed Unitas Fratrum,* 1722–1957, rev. ed. (1967), standard; A.J. LEWIS, *Zinzendorf, Ecumenical Pioneer* (1963), a balanced account; J.I. WEINLICK, *Count Zinzendorf* (1956), authoritative but undocumented, *The Moravian Diaspora* (1959), scholarly; J.E. HUTTON, *A History of the Moravian Church,* 2nd ed. (1909), comprehensive and readable; and G.L. GOLLIN, *Moravians in Two Worlds* (1967), a sociological study.

(J.D.N.)

More, Sir Thomas

Thomas More, the English saint and humanist, prolific writer of anti-heretical works, and statesman, was martyred for refusing to accept the Act of Supremacy establishing the king as supreme head of the Church of England. More, however, is best remembered as the author of *Utopia,* a speculative political essay contrasting an ideal society with the evils of the existing one.

Education

Thomas More was born in London on February 7, 1477, the eldest son of John More, a lawyer who was later knighted and made a judge of the King's Bench. More was educated at the City's best school, St. Anthony's in Threadneedle Street, and in the household of John Morton, archbishop of Canterbury and chancellor of England. The future cardinal, a shrewd judge of character, predicted that the bright and winsome page would prove a "marvellous man." His interest sent the boy to Oxford, where More seems to have spent two years, developing his command of Latin and undergoing a thorough drilling in formal logic.

Sir Thomas More, oil painting by Hans Holbein the Younger, 1527. In the Frick Collection, New York.
Copyright The Frick Collection, New York

About 1494 his father brought More back to London to study the common law. In February 1496 he was admitted to Lincoln's Inn, one of the four legal societies preparing for admission to the bar. Grinding at law probably toughened More's delicate intellectual fibre and fostered the healthy realism that struck the readers of *Utopia.*

Marriage and early career. In 1501 More became an "utter barrister," a full member of the profession. Thanks to his boundless curiosity and a prodigious capacity for work, he managed, along with his study and practice of the law, to keep up his literary pursuits. He read avidly from Holy Scripture, the Church Fathers, and the pagan classics. He tried his hand at all literary genres, including drama and poetry, to acquire ease and fluency of pen and tongue.

Although bowing to his father's decision that he should become a lawyer, More was prepared to be disowned rather than disobey God's will. To test his vocation to the priesthood, he resided for about four years in the Carthusian monastery adjoining Lincoln's Inn and shared as much of the monks' way of life as was practicable. Although attracted especially to the Franciscan order, More decided that he would best serve God and his fellowmen as a lay Christian.

Toward the end of 1504 or in early 1505, More married Jane Colt, the eldest daughter of an Essex gentleman farmer. Quite young and with little formal schooling, she received instruction in music and enough Latin from her husband to make her a competent hostess for non-English visitors, such as the Dutch humanist Erasmus, who was eventually given permanent rooms in the Old Barge on the Thames side in Bucklersbury in the city of London, More's home for the first two decades of his married life. More never discarded the habits of early rising, prolonged prayer, fasting, and wearing the hair shirt. God remained the centre of his life.

Friendship with Erasmus. Erasmus first went to England in the summer of 1499, as the guest of William Blount (later Lord Mountjoy), whom he had tutored in Paris. His first extant letter to More, written in October 1499, implies weeks of acquaintance and shows the Dutch priest as the more eager of the two friends.

During his second stay in England (1505–06), Erasmus was More's houseguest. To increase their proficiency in both Latin and Greek, and to learn the art of ironic

dialogue and of persuasive oratory, the two friends vied in translating several pieces from Lucian, the 2nd-century Greek satirist. The fruit of these labours, published in Paris in 1506, was often reprinted.

More's only extant English book of this decade is a *Life of John Picus*—the Italian humanist Giovanni Pico della Mirandola—translated from the Latin, with a devotional anthology from Pico's *Opera*.

Henry VIII's accession, in April 1509, brought Erasmus back to England for his longest stay. In More's house he wrote his *Praise of Folie*. The very title, *Encomium moriae*, plays on the name of More, who was among the very few to relish the full flavour of the sophisticated satire, rich in Lucianic echoes and even biblical undertones.

In the face of the encroachment of legal and commercial business, More needed to do some tightrope walking to maintain his family life, a large dose of prayer, and his humanistic pursuits. The important negotiations he conducted in 1509 on behalf of a number of London companies with the representative of the Antwerp merchants confirmed his competence in trade matters and his gifts as an interpreter and spokesman. From September 1510 to July 1518, when he resigned to be fully in the king's service, More was one of the two undersheriffs of London, "the pack-horses of the City government." He endeared himself to the Londoners—as an impartial judge, a disinterested consultant, and "the general patron of the poor."

More's domestic idyll came to a brutal end in the summer of 1511 with the death, perhaps in childbirth, of "More's little wife": *uxorcula Thomae Mori*, to quote her epitaph. He was left a widower with four children: Margaret, Elizabeth, Cecily, and John. Within weeks of his first wife's death, he married Alice Middleton, the widow of a London mercer. She was several years his senior and had a son and a daughter of her own; she did not bear More any children. He taught her music, though not Latin. A shrewd housewife, she also proved an affectionate mother to a family that was to include a foster-daughter and several wards. The More home was a Christian academy of "virtue and learning," where tutors taught all kinds of subjects and the younger ones were expected both to learn and to teach.

Second marriage

More's *History of King Richard III*, written in Latin and in English between about 1513 and 1518, is the first masterpiece of English historiography. Though never finished, it influenced succeeding historians. Shakespeare is indebted to More for his portrait of the tyrant.

The "Utopia." In May 1515 More was appointed to a delegation to revise an Anglo-Flemish commercial treaty. The conference was held at Bruges, with long intervals that More used to visit other Belgian cities: his favourite friends were men who, like him, combined cultural pursuits with political responsibilities. Begun in the Low Countries and completed after his return to London, the "golden little book" was published at Louvain in December 1516. It was an immediate success with the audience for which More wrote it: the humanists and an elite group of public officials.

Book 1 is a keen analysis and passionate indictment of the economic and social conditions prevailing in Europe, particularly in England. In book 2 the narrator depicts the best state of society, as he has seen it flourishing on an island off the shore of the New World. The Utopians are immune to Europe's evils, thanks to their basic communism. All citizens participate with full equality in activities related to food, clothing, housing, education, philosophy, government, war, and all else, even religion. There is one exception to this universal sharing: monogamy is safeguarded by stricter laws than in Christendom. A common denominator for full citizenship is belief in a good and just God who rules this world and metes out reward or punishment in an afterlife of immortality.

Utopia is a Greek name of More's coining, from *ou-topos* ("no place"); a pun on *eu-topos* ("good place") is suggested in a prefatory poem. As was inevitable with an ironic dialogue, More's fiction has been subjected to the most diverse interpretation. The German Socialist theoretician Karl Kautsky took its communism seriously, thus causing More to be hailed as a father of the Russian Revolution; historians have read into it a blueprint for British imperialism; literary critics have viewed it merely or mainly as the *jeu d'esprit* of a noncommitted intellectual; Humanists have proclaimed it a manifesto of total reform for the Christian renaissance. The endeavour of recent scholarship has been to decipher the witty travelogue in the light of its author's later and larger writing, to ponder the circumstances of its genesis and publication, and to place it firmly in its historical context, including More's personal misgivings about entrance into his king's service. Among the "sources whence flow all social evils," the chief is pride or ambition, the same "pestilent serpent" embodied by More in the hideous figure of the usurper Richard III.

Interpretations of Utopia

The island of Nowhere is a vision to be contemplated (like Plato's *Republic*), rather than a practical goal to be achieved; the narrator's "idle tales" are, as the French humanist and Greek scholar Budé remarked in 1517, a nursery (*seminarium*) of intellectual seeds to be selected and tested: some of them may be transplantable onto the stubborn soil of Europe. More's own household had a few Utopian traits: no gaming, dicing, idleness, or "idle business"; all did some gardening; education was a lifelong affair for both sexes; and there was something patriarchal about the meals and the evening prayers. But folly had been praised under his roof too eloquently for "nature and reason" to hold an undisputed sway over his life and that of his charges.

While *Utopia* made its impact as soon as it was published, its status as a classic is attested by its having remained "a commentary on the morning paper": penology, eugenics, euthanasia, divorce, the gold standard, women's rights, continued and state-controlled education, religious pluralism, and even ecology, More has a word to say on all these topics. When does he speak his mind? Even in the more serious *Dyaloge Concerning Herecies* (1529), the messenger says: "Ye use to look so sadly when ye mean merrily." No one looks more "sadly"—that is, more in earnest—than the humourless, doctrinaire narrator of the *Utopia*, yet even he perhaps means merrily at times.

Career as king's servant. On May 1, 1517, a mob of Londoners rose against the alien residents: More's role in quenching this Evil-Mayday riot inspired a scene, attributed to Shakespeare, in *The Book of Sir Thomas More*, a composite Elizabethan play. More's success in the thorny negotiations with the French at Calais and Boulogne (September to December 1517) over suits born of the recent war made it harder for him to dodge royal service. While yielding to pressure, he embraced the chance of furthering peace and reform. The lord chancellor, Thomas Wolsey, now looked ready to implement some of the political ideas of the Christian humanists.

Between 1515 and 1520 More campaigned spiritedly for Erasmus' religious and cultural program—Greek studies as the key to a theology renewed by a return to the Bible and the Church Fathers—in poems commending Erasmus' *Novum instrumentum*, a new Latin version of the New Testament, and open letters to his various personalities. More's Latin poems were published in 1518 under one cover between his *Utopia* and Erasmus' *Epigrammata*: they are extremely varied in metre and matter, their main topics being government, women, and death.

Erasmus offered his London friend as a model for the intelligentsia of Europe in letters to the German humanist Ulrich von Hutten, the Paris scholar Germain De Brie (1520)—with whom More had just engaged in a polemic and Guillaume Budé, whom More had met in June 1520 at the Field of the Cloth of Gold, the meeting ground, near Calais, between Henry VIII and Francis I. According to Erasmus, More was of medium stature, his complexion was fair, and his eyes were bluish gray. Simplicity was his mark in food and dress. He shrank from nothing that imparted an innocent pleasure, even of a bodily kind. He had a speaker's voice and a memory that served him well for extempore rejoinders. "Born for friendship," he could extract delight from the dullest people or things. His family affections were warm yet unobtrusive. He gave

Erasmus' description of More

freely and gladly, expecting no thanks. Exotic objects and curious animals graced his house and garden. Amid his intense professional activity, he found hours for prayer and for supervising his domestic school. Most of his charges were girls, to whom he provided the most refined classical and Christian education.

In 1520 and 1521 More took part in talks, at Calais and Bruges, with the emperor Charles V and with the Hansa merchants. In 1521 he was made undertreasurer and knighted. His daughter Margaret married William Roper, a lawyer. For Henry VIII's *Defense of the Seven Sacraments*, he acted as "a sorter out and placer of the principal matters." When Martin Luther hit back, More vindicated the King, under the pseudonym of Gulielmus Rosseus, in a learned, though scurrilous, *Responsio ad Lutherum* (1523). In 1526 he attacked all the major dogmas of Luther in an *Epistola ad Pomeranum*, first printed in 1568.

About 1522 More wrote a meditation that he never finished on *The Four Last Things*—death, doom, pain, and joy. In addition to his routine duties at the Exchequer, More served throughout these years as "Henry's intellectual courtier, his tame humanist." He welcomed foreign envoys, delivered official speeches, drafted treaties, read the dispatches exchanged between the King and Wolsey, and answered in the King's name. Often he rode posthaste between the Cardinal's headquarters at Westminster and Henry's various hunting residences. In April 1523 he was elected speaker of the House of Commons: while loyally striving to secure the government's ends, he made a plea for truer freedom of speech in Parliament. The universities—Oxford in 1524, Cambridge in 1525—made him their high steward.

By 1524 More had moved to Chelsea. The Great House he built there bore the stamp of his philosophy, its gallery, chapel, and library all geared toward studious and prayerful seculsion. In 1525 he was promoted to chancellor of the duchy of Lancaster, which put a large portion of northern England under his judiciary and administrative control. Before the end of 1526, the Dutch painter Hans Holbein arrived at Chelsea from Basel recommended by Erasmus and set to work on portraits of his hosts.

On More's return from an embassy to France with Wolsey in the summer of 1527, Henry VIII "laid the Bible open before him" as proof that his marriage to Catherine of Aragon, who had failed to produce a male heir, was void, even incestuous, because of her previous marriage to Henry's late brother. More tried in vain to share the King's scruples, but "diligent search and long study" confirmed his view that Catherine was the King's true wife. After being commissioned in March 1528 by Bishop Tunstall of London to read all heretical writings in the English language in order to refute them for the sake of the unlearned, he published seven books of polemics between 1529 and 1533—the first and best being *A Dialogue Concerning Herecies.*

Years as chancellor of England. In 1529 Tunstall and More assuaged the rancour provoked by Wolsey's duplicity; thus England was included in the Peace of Cambrai, proclaimed that August, between France and Emperor Charles V. The Cardinal parted with the great seal on October 19, and six days later More became chancellor. On November 3, he opened the Parliament that was to forge the legal instruments for his death. As the King's mouthpiece, More indicted Wolsey in his opening speech and, in 1531, proclaimed the opinions of universities favourable to the divorce; but he did not sign the letter of 1530 in which England's nobles and prelates, including Wolsey, pressured the Pope to declare the first marriage void, and he tried to resign in 1531, when the clergy acknowledged the king as their supreme head, albeit with the clause "as far as the law of Christ allows." More was never Henry VIII's chief adviser; he thus discharged his legal duties with the greater energy: his regularity and promptitude became legendary.

His longest book, *The Cōfutacyon of Tyndals Answere*, in two volumes (1532 and 1533) centres on "what the church is." To the stress of stooping for hours over his manuscript More ascribed the sharp pain in his chest, perhaps angina, which he invoked when begging Henry to free him from the yoke of office. This was on May 16, 1532, the day when the convocation of the archdiocese of Canterbury delivered to the crown the document by which they promised never to legislate or convene without royal assent, thus placing a lay person at the head of the spiritual order.

While enjoying the privacy of his home, although in relative poverty, and resting by doctor's order, More prepared a tomb for himself and his wives. Its Latin epitaph underlines the fact that his resignation had been no dismissal and proclaims that he had "been grievous to thieves, manquellers [murderers], and heretics," a duly thought-out crescendo occurring elsewhere in his works.

He continued his campaign for the old faith, defending England's anti-heresy laws and his own handling of heretics, both as magistrate and as writer, in two books of 1533: the *Apology* and the *Debellacyon*. He also laughs away the accusation of greed levelled by William Tyndale, translator of parts of the first printed English Bible. More's poverty was so notorious that the hierarchy collected £5,000 to recoup his polemical costs, but he refused this grant lest it be construed as a bribe.

Anti-heretical writings

Indictment, trial, and execution. His refusal to attend the coronation of Anne Boleyn, whom the King married after his divorce from Catherine in 1533, marked him out for vengeance. Several charges of accepting bribes recoiled on the heads of his accusers. In February 1534 More was included in a bill of attainder for alleged complicity with Elizabeth Barton, who had uttered prophecies against Henry's divorce, but he produced a letter in which he had warned the nun against meddling in affairs of state. The House of Lords passed the bill only after More's name had been struck off the list. Summoned to appear before royal commissioners on April 13 to assent under oath to the Act of Succession, he said he "would not deny to swear to the succession"—that is, to give priority to Anne's issue over Catherine's—but he could not in conscience swear to the whole document. His main objection was the sentence in the preamble repudiating the authority of any "foreign potentate," which implied rejection of papal supremacy. Led to the Tower on April 17, More welcomed prison life. But for his family responsibilities, he would have chosen for himself "as strait a room and straiter too," as he said to his daughter Margaret, who after some time took the oath and was then allowed to visit him.

At the time of his arrest, More was writing a *History of the Passion*, which ends with the institution of the Eucharist. In prison, More gave up such works, which required accurate references. Instead, he wrote *A dialoge of comfort against tribulacion*, a masterpiece of Christian wisdom and of literature, studded with biblical texts and flavoured with humorous anecdotes, some of them autobiographical.

His trial took place on July 1, 1535. Richard Rich, the solicitor general, a creature of Thomas Cromwell, the unacknowledged head of the government, testified that the prisoner had, in his presence, denied the King's title as supreme head of the Church of England. Despite More's scathing denial of this perjured evidence, the jury's unanimous verdict was "guilty." Before the sentence was pronounced, More spoke "in discharge of his conscience." The unity of the church was the main motive of his martyrdom. His second objection was that "no temporal man may be head of the spirituality." Henry's marriage to Anne Boleyn, to which he also referred as the cause for which they "sought his blood," had been the occasion for the assaults on the church: among his judges were the new queen's father, brother, and uncle.

More's defense

More was sentenced to the traitor's death—"to be drawn, hanged, and quartered"—which the King changed to beheading. During five days of suspense, he prepared his soul to meet "the great spouse" and wrote a beautiful prayer and several letters of farewell. On July 6, he walked to the scaffold on Tower Hill. "See me safe up," he said to the lieutenant, "and for my coming down let me shift for myself." He told the onlookers to witness that he was dying "in the faith and for the faith of the

Catholic Church, the king's good servant and God's first." He altered the ritual by blindfolding himself, playing "a part of his own" even on this awful stage.

The news of More's death shocked Europe. Erasmus mourned the man he had so often praised, "whose soul was more pure than any snow, whose genius was such that England never had and never again will have its like." The official image of More as a traitor did not gain credence even in Protestant lands.

Assessment. Margaret Roper, More's favourite child and his confidante for years, the washer of his hair shirt and rescuer of his unpublished writings, died in 1544. Her husband's recollections of More as "a mirror of virtue in worldly greatness" first appeared in 1626. Editions of some of More's works appeared all over Europe. The folio *English Works* (1557) was dedicated to Queen Mary. His Latin works appeared at Basel (1563), Louvain (1565), and Frankfurt (1689).

Cresacre More, a grandson of John, Thomas More's only son, wrote a *Life of More* (Paris, *c.* 1631), which belongs to hagiography. Nicholas Harpsfield, archdeacon of Canterbury in Queen Mary's reign, wrote a full life of More, first printed in 1932, using Roper's notes. Thomas Stapleton, a theology professor in Douai, published a *Thomae Mori vita* in 1588, as part of a trilogy—*Tres Thomae*, the other Thomases being the apostle and Becket.

The triumph of Anglicanism brought about a certain eclipse of Thomas More. Foxe's *Book of Martyrs* (1563) caricatured him as a cruel foe of the gospel. The publication of the state papers restored a fuller and truer picture of More, preparing public opinion for his beatification (1886). He was canonized by Pius XI in May 1935. His name is given to boys, schools, clubs, family clinics, and parishes, where his panegyric is preached ritually every year.

Though the man is greater than the writer, and though nothing in his life "became him like the leaving of it," his "golden little book" has earned him greater fame than the crown of martyrdom or the million words of his English works. His name is listed among the heroes of the Revolution on a stela in the Red Square in Moscow.

Erasmus' phrase describing More as *omnium horarum homo* was rendered later as "a man for all seasons" and was given currency by Robert Bolt's play *A Man For All Seasons* (1960) and Fred Zinnemann's film (1966). Monuments to him have recently been placed in Westminster Hall and the Tower in London. In the words of the English Catholic apologist G.K. Chesterton, More "may come to be counted the greatest Englishman, or at least the greatest historical character in English History."

BIBLIOGRAPHY. R.W. GIBSON and J.M. PATRICK, *St. Thomas More: A Preliminary Bibliography of His Works and of Moreana to the Year 1750* (1961); FRANK and M.P. SULLIVAN, *Moreana: Materials for the Study of Saint Thomas More*, 5 vol. (1964–71).

Editions: The Yale University edition of the *Complete Works* is in progress with 4 of 18 volumes completed to date: *The History of King Richard III*, ed. by R.S. SYLVESTER (1963); *Utopia*, ed. by E. SURTZ and J.H. HEXTER (1965); *Responsio ad Lutherum*, ed. by JOHN HEADLEY (1969); *Confutation*, ed. by R.J. SCHOECK *et al.* (1973). The first English version of *Utopia*, by RALPH ROBINSON (1551), a classic for centuries, has been superseded as a standard by the Yale edition. See also H.W. DONNER, *Introduction to Utopia* (1945); and the paperbound Yale *Utopia*, ed. by E. SURTZ (1964).

Correspondence and studies: The letters exchanged between More and Erasmus are in P.S. ALLEN (ed.), *Opus epistolarum Des. Erasmi Roterodami*, 12 vol. (1905–58). For the rest, see E.F. ROGERS, *The Correspondence of Sir Thomas More* (1947) and *Selected Letters* (1961). *The Latin Epigrams of Thomas More*, ed. by L. BRADNER and C.A. LYNCH (1953), provide the text and an English rendering. Useful monographs are STANLEY MORISON, *The Likeness of Thomas More: An Iconographical Survey of Three Centuries*, ed. by NICOLAS BARKER (1963); and GERMAIN MARC'HADOUR, *The Bible in the Works of Thomas More*, 5 vol. (1969–72).

Biography: The first standard modern life, R.W. CHAMBERS, *Thomas More* (1935), has come under fire recently, because it tones down the starkness of More's asceticism, the radicalism of his Utopian vision, his hatred of schism and heresy, and the dogmatic motives of his martyrdom. E.E. REYNOLDS, *The Field Is Won: The Life and Death of Saint Thomas More* (1968), is the work of a professional biographer, who had prepared for his task by writing six other books on More and his circle. Less literary, less allusive than Chambers, Reynolds is balanced and up-to-date.

(G.P.Ma.)

Morgan, Thomas Hunt

Thomas Hunt Morgan, geneticist and embryologist, is particularly noted for his experimental research with the fruit fly, *Drosophila melanogaster*, by which he established the chromosome theory of heredity. He showed that genes are linked in a series on chromosomes and are responsible for identifiable, hereditary traits. His research in embryology and genetics helped to move biology away from its former emphasis on descriptive morphology to its present experimental and quantitative character.

By courtesy of the California Institute of Technology, Pasadena

Morgan.

Morgan was born in Lexington, Kentucky, on September 25, 1866, into a prominent Southern family. His father, Charlton Hunt Morgan, was the American consul in Messina, Sicily, in 1860, and his uncle was a Confederate Army colonel. His maternal great-grandfather was Francis Scott Key, author of the national anthem.

Early in life, Morgan showed an interest in natural history during his frequent excursions into the Kentucky and western Maryland countrysides, where he spent many summers. In 1886 he received the B.S. degree from the State College of Kentucky, later the University of Kentucky, in zoology and then immediately entered Johns Hopkins University for graduate work in biology.

Criticism of descriptive biology

At Hopkins, Morgan studied under the morphologist and embryologist William Keith Brooks, who led him to see the important interrelationships among apparently disparate disciplines of biology, such as anatomy, embryology, and cytology, and to see heredity and variation as appropriate problems for investigation. Morgan also learned the value of marine organisms—the oldest and most basic forms of life—for understanding general principles. Brooks's descriptive and sometimes highly speculative methods contributed to Morgan's growing dissatisfaction with the morphological tradition and its nonrigorous, nonexperimental approach. In addition, he felt that morphologists subjugated embryology to evolution, thus neglecting the value of the discipline for its own sake.

After being awarded the Ph.D. in 1890, Morgan accepted a teaching post at Bryn Mawr College, near Philadelphia, where he came under the influence of the German-born physiologist Jacques Loeb, who was an outstanding advocate of the experimental and mechanistic approach to biology.

Experimental embryology

Morgan's growing preference for experimental work, particularly in embryology, was strengthened by his visits to the Zoological Station at Naples, Italy, in 1890, 1894, 1895, and 1900. There he made contact with an interna-

tional group of researchers, many of whom strongly favoured experimental as opposed to descriptive methods. He worked closely in 1894–95 with the German embryologist Hans Driesch, who, as an enthusiastic proponent of the experimental approach to embryology, called *Entwicklungsmechanik*, had performed some highly controversial experiments on the sea urchin egg.

During the period 1893–1910, Morgan applied experimental techniques to fundamental problems of embryology. In order to identify causally related events during development, he analyzed such problems as the formation of embryos from separated blastomeres (early embryonic cells) and fertilization in nucleated and nonnucleated egg fragments. As examples of the effects of physical factors, he analyzed the way in which the spatial orientation of eggs affects their future development and the action of salt concentration on the development of fertilized and unfertilized eggs. In 1904 he married one of his graduate students at Bryn Mawr, Lillian V. Sampson, a cytologist of considerable skill. The same year he accepted an invitation to assume the professorship of experimental zoology at Columbia University, where, during the next 24 years, he conducted most of his important research in heredity.

Morgan favoured the development of marine-research laboratories where biologists from different specialities could work together and exchange ideas. He contributed to the development of this approach at the Marine Biological Laboratory in Woods Hole, Massachusetts, where he worked nearly every summer during his Bryn Mawr and Columbia years and served as a trustee from 1897 until his death.

Objections to Darwin's theory of natural selection

Like most embryologists and many biologists at the turn of the century, Morgan found the Darwinian theory lacking in plausibility. It was difficult to conceive of the development of complex adaptations simply by an accumulation of chance variations. Moreover, Darwin had provided no mechanism of heredity to account for the origin or transmission of variations, except his early and hypothetical theory of pangenesis. Although Morgan believed that evolution itself was a fact, the mechanism of natural selection proposed by Darwin seemed incomplete because it could not be put to an experimental test.

Morgan had quite different objections to the Mendelian and chromosome theories. Both theories attempted to explain biological phenomena by postulating units or material entities in the cell that somehow control developmental events. To Morgan this was too reminiscent of the preformation theory—the idea that the fully formed adult is present in the egg or sperm—that had dominated embryology in the 18th and early 19th centuries. Although Morgan admitted that the chromosomes might have something to do with heredity, he argued in 1909 and 1910 that no single chromosome could carry specific hereditary traits. He also claimed that Mendelian theory was purely hypothetical: although it could account for and even predict breeding results, it could not describe the true processes of heredity. The fact that each pair of chromosomes separates, with the individual chromosomes then going into different sperm or egg cells in exactly the same manner as Mendelian factors, did not seem to be sufficient proof to Morgan for claiming that the two processes had anything to do with each other. In 1909, Morgan delivered a particularly vehement attack on Mendelian theory and the seemingly unrigorous approach of many Mendelians.

The work on *Drosophila*

Morgan apparently began breeding *Drosophila* in 1908. Surprisingly, he was not as interested in heredity itself as in determining whether macromutations, observed in plants, also occurred in animals. In 1909 he observed a small but discrete variation known as white-eye in a single male fly in one of his culture bottles. Aroused by curiosity, he bred the fly with normal (red-eyed) females. All of the offspring (F_1) were red-eyed. Brother–sister matings among the F_1 generation produced a second generation (F_2) with some white-eyed flies, all of which were males. To explain this curious phenomenon, Morgan developed the hypothesis of sex-limited—today called sex-linked—characters, which he postulated were part of the X-chromosome of females. Other genetic variations arose in Morgan's stock, many of which were also found to be sex-linked. Because all of the sex-linked characters were usually inherited together, Morgan became convinced that the X-chromosome carried a number of discrete hereditary units, or factors. He adopted the term gene, which was introduced by the Danish botanist Wilhelm Johannsen in 1909, and concluded that genes were possibly arranged in a linear fashion on chromosomes. Much to his credit, Morgan rejected his skepticism about both the Mendelian and chromosome theories when he saw from two independent lines of evidence—breeding experiments and cytology—that one could be treated in terms of the other.

In collaboration with A.H. Sturtevant, C.B. Bridges, and H.J. Muller, who were undergraduates at Columbia, Morgan quickly developed the *Drosophila* work into a large-scale theory of heredity. Particularly important in this work was the demonstration that each Mendelian gene could be assigned a specific position along a linear chromosome "map." Further cytological work showed that these map positions could be identified with precise chromosome regions, thus providing definitive proof that Mendel's factors had a physical basis in chromosome structure. A summary and presentation of the early phases of this work was published by Morgan, Sturtevant, Bridges, and Muller in 1915 as the influential book *The Mechanism of Mendelian Heredity*. To varying degrees Morgan also accepted the Darwinian theory by 1916.

After the last of the Morgans' four children entered school, Mrs. Morgan returned to the laboratory and added considerably to cytological aspects of the *Drosophila* work. She was responsible for discovering and analyzing the attached-X *Drosophila* strain, which became widely used to maintain sex-linked mutants, in which the females are usually weak or sterile.

In 1928 Morgan was invited to organize the division of biology of the California Institute of Technology, an offer he could not refuse because of the opportunity it provided to establish a model department along modern—that is, experimental and quantitative—lines. He was also instrumental in establishing the Marine Laboratory on Corona del Mar as an integral part of Caltech's biology training program.

In subsequent years, Morgan and co-workers, including a number of postdoctoral and graduate students, continued to elaborate on the many features of the chromosome theory of heredity. Toward the end of his stay at Columbia and more so after moving to California, Morgan himself slipped away from the technical *Drosophila* work and began to return to his earlier interest in experimental embryology. Although aware of the theoretical links between genetics and development, he found it difficult at that time to draw the connection explicitly and to support it with experimental evidence.

Assessment

According to Morgan, biology had been dominated for too long by purely descriptive methods. In both research and writing, he sought to demonstrate the importance of the quantitative and experimental approach—including the methods of physics and chemistry—to the study of biology. Only by experiments, he maintained, could biological processes be identified. Although not a strict mechanist or physico-chemical reductionist, Morgan felt that life sciences could progress only by applying the methods of the exact sciences. This emphasis was well reflected in Morgan's lifelong interest in the Marine Biological Laboratory in Woods Hole, Massachusetts, and in his early visits to the Zoological Station at Naples, Italy.

Morgan's career encompassed studies on a wide range of organisms and a large number of topics. He had an active and imaginative mind, which moved easily from one special field to another, while always maintaining awareness of the connecting links between them. His persistent interest in marine biology emphasized his commitment to biology as a whole, as opposed to narrowly specialized disciplines. Even in his most technical work—for example, on the chromosome theory of heredity—Morgan saw the broad implications for other areas, such as embryology and evolution.

In 1924 Morgan received the Darwin Medal; in 1933 he was awarded the Nobel Prize for Medicine or Physiology for his discovery of "hereditary transmission mechanisms in *Drosophila*"; and in 1939 he was awarded the Copley Medal by the Royal Society of London, of which he was a foreign member. In 1927–31 he served as president of the National Academy of Sciences; in 1930 of the American Association for the Advancement of Science; and in 1932 of the Sixth International Congress of Genetics. He remained on the faculty at Caltech until he died, in Pasadena, on December 4, 1945.

BIBLIOGRAPHY. A.H. STURTEVANT, "Thomas Hunt Morgan, 1866–1945," *Biog. Mem. Natn. Acad. Sci.*, 33:283–325 (1959), is a reasonably complete biographical sketch with complete bibliography. Discussions of Morgan's work, and the work of the "Morgan school" of *Drosophila* geneticists may be found in A.H. STURTEVANT, *A History of Genetics* (1965); L.C. DUNN, *A Short History of Genetics* (1965); and E.A. CARLSON, *The Gene: A Critical History* (1966). Specific aspects of the development of Morgan's scientific work are: G.E. ALLEN, "Thomas Hunt Morgan and the Problem of Sex Determination, 1903–1910," *Proc. Am. Phil. Soc.*, 10:48–57 (1966); "Thomas Hunt Morgan and the Problem of Natural Selection," *J. Hist. Biol.*, 1:113–139 (1968); and "T.H. Morgan and the Emergence of a New American Biology," *Q. Rev. Biol.*, 44:168–188 (1969).

Among Morgan's most important books are those dealing with (1) evolution: *Evolution and Adaptation* (1903), in which he strongly criticizes Darwinian theory; and *A Critique of the Theory of Evolution*, 2nd ed. (1925), a more favourable view of the selection process; (2) heredity: *Heredity and Sex* (1913), his first major exposition of the Mendelian system in relation to *Drosophila;* and with A.H. STURTEVANT, H.J. MULLER, and C.B. BRIDGES, *The Mechanism of Mendelian Heredity* (1915); and *The Theory of the Gene* (1926); the latter two works firmly established the Mendelian theory as it applied to heredity in all multicellular (and many unicellular) organisms; and (3) embryology: *The Development of the Frog's Egg: An Introduction to Experimental Embryology* (1897), a detailed outline of the developmental stages of frogs' eggs; *Regeneration* (1901), a discussion of experimental approaches to regeneration; *Experimental Zoology* (1907), a strong statement on the value of experimentation in biology; *Experimental Embryology* (1927), Morgan's statement on the value of experimentation in embryology; and *Embryology and Genetics* (1934), an attempt to relate the theory of the gene to the problem of embryological differentiation and development.

(G.E.Al.)

Mormonism

Mormonism is a religion founded in the United States in 1830. The church's membership is principally concentrated in Utah and in neighbouring states. This article surveys the origins, history, teachings, social ethic, and church structures of Mormonism.

ORIGIN AND CHARACTERISTICS

Distribution

The Church of Jesus Christ of Latter-day Saints, the principal formal body embracing Mormonism, has about 3,600,000 members, nearly 80 percent of whom are in the United States; the majority of the remainder are in Latin America (especially Mexico), Canada, Europe (especially England), and parts of Oceania (especially New Zealand). From headquarters in Salt Lake City, Mormons are organized through a lay priesthood in local units that regulate religious practice and important welfare activities. The church was founded in New York State by Joseph Smith, Jr. (1805–44), after, according to his account, he had translated by revelation the *Book of Mormon*, which recounted the history of tribes of Israel that migrated to America centuries before Christ and that underwent experiences closely similar to those recounted in the Bible. Smith described American Indians as descendants of these tribes.

Smith instituted a new ecclesiology and theology. Mormonism proclaimed a new dispensation and the restoration of the two orders of priesthood and rituals of the "true church." Deity was pronounced to be an attainment of spiritual evolution for which all Mormons were eligible, and the unregenerate were also promised a measure of eternal glory. Mormonism combined universalism with contemporary millennialism but gave millennialism the goal of building up Zion in the American West under a prophet to whom was revealed not only theological truth but day-to-day practical guidance (see MILLENNIALISM).

New secular institutions were proclaimed, including collective ownership (later changed to a system of tithing) and plural marriage, which was practiced by Smith himself (although the schismatic Reorganized Church of Jesus Christ of Latter Day Saints disputes this point) and by most leading Mormons until, under pressure from the federal government, the practice was given up.

Collectivism, persecution, conflict, successive migrations, and self-identification as a chosen people under the control of an inspired religious authority produced a distinctive Mormon way of life characterized by order, activism, in-group orientation, and vigorous proselytism, all of which were sanctioned by continuing revelation.

HISTORY

In 1827 Joseph Smith, Jr., son of an impoverished New England farmer, living in that part of upstate New York now known, because of recurrent revivalistic enthusiasms, as "the burned-over district," claimed to have received from an angel, Moroni, golden plates from which, with the use of special stones set in silver bows, he translated the *Book of Mormon*. Three associates solemnly testified that they, too, saw both the plates and the angel Moroni, and thereafter they and other converts frequently claimed visions and revelations.

Smith's followers soon moved to Kirtland, Ohio, where a prominent Campbellite preacher, Sidney Rigdon, and his following had embraced Mormonism. Authorized by revelation, and under Rigdon's influence, Smith established at Kirtland and in Missouri (where some Mormons had migrated) the communistic United Order of Enoch. Strife with non-Mormons led to the prophet's being tarred and feathered, to killings, and to the burning of Mormon property. Despite persecution, the apostasy of some early converts, occasional quarrels among Smith's brothers, the failure of specific prophecies, and the collapse of the bank that Smith had founded in Kirtland, Mormonism grew. Armed skirmishes led several thousand Mormons to leave Missouri in 1839, even though Smith had designated it as Zion. In Illinois, Smith built a new city, Nauvoo, of which he became mayor and lieutenant general of militia. Mormon self-assurance, commercial advantage, and political power provoked renewed hostility from the "Gentiles," while Smith's secret teaching and practice of plural marriage and his trade in land caused dissension among his ablest converts. Smith was arrested and murdered by Illinois militiamen in June 1844.

Early growth

The move to Utah. Despite rivalry in the Mormon hierarchy, the majority chose to follow Brigham Young, president of the Council of Twelve Apostles, to Utah in 1846–47. Young dominated his people, decided on their place of settlement, built Salt Lake City, and promoted further missionary work in Britain and Scandinavia. The Mormon migrations from Europe were admirably organized, and Mormons were highly praised for their discipline, orderliness, and moral probity by the captains of the ships that conveyed them to America. The overland journey to Utah was itself a rigorous test of the converts' faith, especially after the disastrous trek in 1856 when the converts were organized to cross from Iowa City to Utah pulling handcarts.

Young became territorial governor in 1851 and instituted polygyny in 1852. His subsequent defiance of federal authority (under Pres. James Buchanan's administration) led to his deposition and the "Utah War," in which a Mormon militia was mobilized. Mormon settlers fled from northern Utah, burning the crops at Ft. Supply and Ft. Bridger and the grass on the route of the advancing United States federal troops. In 1857 occurred the Mountain Meadows massacre, by uncontrolled Mormon enthusiasts, of non-Mormon settlers bound for California. Although the whole military episode became known as "Buchanan's blunder," it proved to be the end of direct Mormon political control of Utah. Polygyny was relinquished by the church in 1890.

Splinter groups. An important minority of Mormons rejected Young and remained in Iowa and Illinois, where, with the adherence of Emma Smith, the prophet's (first) wife, and their son, Joseph Smith III, the Reorganized Church of Jesus Christ of Latter Day Saints was formed (1852–60). That church eventually set up its headquarters at Independence, Missouri, which Smith had designated as Zion in 1831. Splinter groups also arose after Smith's death. Rigdon led one; another went to Texas with Lyman Wight, who had been an apostle. David Whitmer and Martin Harris, two witnesses of the golden plates, eventually set up a church in Kirtland. James Jesse Strang, a recent convert, was joined by John C. Bennett (second in command of the Nauvoo militia until expelled by Smith for licentious conduct) and established a polygynous community of about 3,000, eventually proclaiming himself "King in Zion."

SCRIPTURES

Mormons accept the Bible "as far as it is translated correctly." Smith did not finish his inspired translation of the Bible, which incorporates prophecies of his own coming and of the *Book of Mormon*, which is more central to Mormon belief. The *Book of Mormon* is largely similar in style and themes to the Old Testament. It recounts the history of the virtuous, industrious, fair-skinned Nephites, who were eventually exterminated by the sinful, red-skinned Lamanites. A familiar moral cycle is discernible: the virtue of God's people leads to prosperity; then to pride, iniquity, decadence, and sin; and so to God's chastisement and the people's repentance. Smith's other revealed scriptures were later incorporated into the *Pearl of Great Price*, together with his inspired translation of papyri that he declared to be the *Book of Abraham* and the *Book of Moses*. Facsimiles published by Smith as part of the *Book of Abraham* were later declared by experts to be parts of common Egyptian funerary texts—an opinion reinforced by examination of what were claimed to be the original papyri, rediscovered in the Metropolitan Museum of Art, New York City, in 1966. Smith's translation referred to gods (in the plural), affirmed that the stars were peopled with spirits, and held that persons with black skins were cursed "as pertaining to the priesthood." In consequence, Negroes have not been admitted to office in the Mormon church. *Doctrines and Covenants* is a selection of revelations to Smith and one given to Young and includes the manifesto of Pres. Wilford Woodruff abolishing plural marriage. The version of the Reorganized Church includes fewer revelations from the last decade of Smith's life.

DOCTRINES

Mormon doctrine diverges from the orthodoxy of established Christianity, particularly in its polytheism, in affirming that God has evolved from man and that men might evolve into gods, that the Persons of the Trinity are distinct beings, and that men's souls have pre-existed. Mormons accept that Christ came to earth so that all might be saved and raised from the dead but maintain that a person's future is determined by his or her own actions. Justification is by faith and obedience to the ordinances of the church, repentance, Baptism by immersion, and laying on of hands for the Spirit gifts (including prophecy, revelation, and speaking in tongues).

The *Book of Mormon* describes conventional hellfire for the unregenerate, but in 1833 Smith had a revelation of three post-resurrectional kingdoms: the celestial, the terrestrial, and—for the unclean liars and adulterers—the "telestial" kingdom on another earth. Faithful members of the church would inherit eternal life as gods, and even those who had rejected God's law would live in glory. Mormons believe that the return of Christ to earth will lead to the first resurrection and the millennium, the main activity of which will be "temple work," especially Baptism on behalf of the dead. After the millennium and second resurrection, the earth will become a celestial sphere and men will be assigned to the eternal kingdoms.

Mormons regard the Christian churches as apostate; lacking revelations, miracles, and Spirit gifts; and maintaining corrupt rituals, priesthoods, and teachings. Smith came to restore the institutions of the church and God's law in society. Although calling people to repent, Smith's creed reflected contemporary optimism in emphasizing man's inherent goodness and limitless potential for progress. Smith tempered millennialism by calling the faithful to gather and labour to build Zion as a new social order. He curbed revivalistic excess and controlled the claims to revelation of his followers. Successive church presidents have claimed divine inspiration, and private persons may expect inspiration that is helpful to their church duties.

INSTITUTIONS AND PRACTICES

The restored priesthood eliminates distinctions of priest and layman. At the age of 12, all worthy males become deacons and later teachers in the Aaronic priesthood; before they are 20 they become priests. At that age a man becomes an elder in the Melchizedek priesthood; in later life, he may become a "seventy" (a member of a larger priesthood quorum composed of 70 members) and a high priest. Each rank is organized into a quorum and has its own activities. Thus deacons and teachers accompany elders on the extensive house visiting undertaken among Mormons. Those called abroad as missionaries (at their own or their families' expense) are usually elders.

Adult Baptism, signifying repentance and obedience, has acquired additional importance as a ritual that may be undertaken by a proxy for the salvation of those who died without knowledge of the truth. Mormon interest in genealogy proceeds from concern to save dead ancestors. Baptism for the dead, endowment, and sealing (which may also be undertaken by proxy for the dead) are secret but essential ceremonies that take place in the temple. At endowment, the person is ritually washed, anointed with oil, and dressed in temple garments (underwear that good Mormons continue to wear as protection against evil and as a symbol of faithfulness). Initiates witness a dramatic performance of the story of creation, learn secret passwords and grips, and receive a secret name. The sealing ceremony, which was of special importance in the period when Mormons practiced plural marriage, seals men and women in marriage for time and eternity. Smith, who became a Mason in 1842, borrowed Masonic ritual for Mormon ceremonies, eliminating what he regarded as corruptions of the original Hebrew forms.

Persecution and segregation reinforced Mormon community identity. The patriarchal priesthoods, four decades of polygyny, and Baptism for dead ancestors have consolidated kinship allegiances. Church direction of private and public affairs reveals a persisting sectarian strain, but Mormonism is less opposed to secular culture than most sects and Mormons, although committed to millennialism and Spirit gifts, engage in worldly pursuits, business, and politics. Despite prohibitions (on alcohol, tobacco, tea, and coffee) and a vigorous work ethic, Mormonism is not ascetic. Recreation, sport, and education are positive values. Smith danced, wrestled, and, despite his "Word of Wisdom," sometimes drank. He wrote, "Men are that they might have joy." The positive attitude toward recreation, together with the emphasis on order and moral integrity, has been of great importance.

STRUCTURE OF THE CHURCH

Church polity embraces both hierarchic and egalitarian tendencies. Status is widely diffused, but authority is clearly structured. The church president and two councillors, a Council of Twelve Apostles, the First Council of Seventy (concerned especially with missions), and the presiding bishop and two councillors (who control the Aaronic priesthood) constitute, together with the patriarch (a ceremonial office), the "General Authorities" of the church. They are "sustained in office" by the regular and now ritualized vote of confidence of the General Conference, which is open to all Mormons.

At the local level, Mormons are divided into "stakes" of, on the average, 4,000 to 5,000 members under stake presidents; and wards, each of a few hundred members, under a bishop. The life of the individual Mormon is closely regulated by these local units, through which the

religious, economic, and social activities, tithing, and the operation of the church's elaborate welfare plan are organized. Ancillary organizations, Sunday school, mutual improvement associations for young people, a women's relief society, and the two-year missionary work undertaken by many young men ensure a wide distribution of responsibility and involvement.

The Reorganized Church. The Reorganized Church of Jesus Christ of Latter Day Saints holds firmly to the *Book of Mormon* but rejects the evolutionary conceptions of deity and the polytheism implicit in it, the new covenant of celestial marriage, Baptism on behalf of the dead, plural marriage, and tithing. Secret ceremonies are not performed in the Kirtland temple, to which the Reorganized Church gained legal title in 1890. The *Book of Abraham* is not accepted as of divine origin. The church's presidents, descendants of the prophet, have been appointed by revelation.

Other churches. Some Mormon splinter groups adopted communistic practices. The followers of Granville Hedrick took as their mission the building of Smith's projected temple in Missouri and acquired part of the "Temple Lot." After the Mormon church renounced plural marriage in 1890, some groups in Arizona continued the practice, but they remained secret.

BIBLIOGRAPHY. In addition to the basic scriptures cited in the text, the following commentaries on theology and ecclesiology written by Mormons in good standing are useful: J.E. TALMAGE, *A Study of the Articles of Faith* (1901); J.A. WIDTSOE, *Rational Theology* (1915), *Priesthood and Church Government* (1939), and, with L.D. WIDTSOE, *The Word of Wisdom: A Modern Interpretation*, 4th ed. (1938). For an excellent sociological treatment of Mormonism, see THOMAS O'DEA, *The Mormons* (1957). The well-documented history of Mormonism is best studied in works that deal with specific periods. Two early official histories are JOSEPH SMITH, JR., *History of the Church of Jesus Christ of Latter-day Saints*, ed. by B.H. ROBERTS, 6 vol. (1902–12); and B.H. ROBERTS, *A Comprehensive History of the Church of Jesus Christ of Latter-day Saints*, 6 vol. (1930). A popular history from the Mormon point of view is J.B. ALLEN and G.M. LEONARD, *The Story of the Latter-day Saints* (1976). For the origins and history of Mormonism to the death of Joseph Smith, Jr., see F.M. BRODIE, *No Man Knows My History: The Life of Joseph Smith, the Mormon Prophet*, 2nd ed., rev. and enl. (1971); the author, a Mormon who was excommunicated after the publication of her book, is very thorough on the details of Smith's life. Mormon accounts include PRESTON NIBLEY, *Joseph Smith, the Prophet* (1944); and J.A. WIDTSOE, *Joseph Smith: Seeker after Truth, Prophet of God* (1951). On the period in Illinois, see ROBERT FLANDERS, *Nauvoo: Kingdom on the Mississippi* (1965). On the migration from Nauvoo to Utah, see BERNARD A. DE VOTO, *Year of Decision: 1846* (1943); and W.E. STEGNER, *The Gathering of Zion: The Story of the Mormon Trail* (1964). For Mormon recruitment abroad, emigration to the United States, and the journey to Utah, see L.R. and A.W. HAFEN, *Handcarts to Zion: The Story of a Unique Western Migration, 1856–1860* (1960); WILLIAM MULDER, *Homeward to Zion: The Mormon Migration from Scandinavia* (1957); and P.A.M. TAYLOR, *Expectations Westward* (1965). For the life, views, and times of Brigham Young, see the Mormon account by one of his daughters, SUSA YOUNG GATES, written with L.D. WIDTSOE, *The Life Story of Brigham Young* (n.d.); M.R. WERNER, *Brigham Young* (1925); S.P. HIRSHSON, *The Lion of the Lord: A Biography of Brigham Young* (1969). For the period of the "Utah War", see O.G. HAMMOND (ed.), *The Utah Expedition, 1857–1858: Letters of Capt. Jesse A. Gove* (1928); JUANITA BROOKS, *The Mountain Meadows Massacre* (1950); N.F. FURNISS, *The Mormon Conflict, 1850–1859* (1960). More generally, see K.J. HANSEN, *Quest for Empire: The Political Kingdom of God and the Council of 50 in Mormon History* (1967). Useful essays on early Mormonism and on some Mormon schisms are contained in F.M. MCKIERNAN, A.R. BLAIR, and P.M. EDWARDS (eds.), *The Restoration Movement: Essays in Mormon History* (1973). For life in Salt Lake City in the 19th century, see R.F. BURTON, *The City of the Saints*, new ed. by F.M. BRODIE (1963); and WILLIAM MULDER and A.R. MORTENSEN (eds.), *Among the Mormons: Historical Accounts by Contemporary Observers* (1958). Also of importance for Mormon relations of church and state are NELS ANDERSON, *Desert Saints: The Mormon Frontier in Utah* (1942); and R.J. DWYER, *The Gentile Comes to Utah* (1941). On Mormon economic and social organization, see L.J. ARRINGTON, *Great Basin Kingdom: An Economic History of the Latter-day Saints, 1830–1900* (1958). For a sociological study of Mormon polygyny, see KIMBALL YOUNG, *Isn't One Wife Enough?* (1954). For an interpretation of church history from the perspective of the Reorganized Church, and for its own early development, see JOSEPH SMITH and H.C. SMITH, *History of the Church of Jesus Christ of Latter Day Saints*, 3 vol. (1897–1908). Summary statements of the theological and social perspective of the Reorganized Church may be found in R.A. CHEVILLE, *The Latter Day Saints and Their Changing Relationship to the Social Order* (1942); and E.A. SMITH, *Differences That Persist Between the Reorganized Church of Jesus Christ of Latter Day Saints and the Utah Mormon Church* (1943). (B.R.W.)

Morocco

Morocco, which has been an independent nation only since 1956 and a constitutional monarchy since 1962, is situated in northwestern Africa, where it looks northward across the Strait of Gibraltar to the continent of Europe, westward to the Atlantic Ocean, and northeastward to the Mediterranean Sea. To the east, as well as across a long reach of disputed territory in the southeast, lies Algeria, with which Morocco shares much—existence within the overlapping influences of the Arab and the Islāmic worlds; within the community of the Maghrib; and, in a wider sense, within the community of African and Mediterranean nations. In 1970 Spain ceded its Atlantic enclave Ifni to Morocco, but two more Spanish enclaves on the Mediterranean coast, containing Ceuta and Melilla, still remain under Spanish control. In April 1976 Moroccan territory was further enlarged when Morocco and Mauritania agreed to a boundary dividing the former Spanish Sahara between them; the *de facto* annexations resulting were not recognized in international law. Excluding these territories, Morocco had in 1976 an area of 177,117 square miles (458,730 square kilometres). Rabat is the political capital, but Marrakech, Fès, and Meknès are traditional capitals; Tangier is the summer capital.

Territorial boundaries

In spite of the modern European aspect of large parts of such cities as Casablanca, Morocco must yet be considered an underdeveloped country from a socio-economic point of view. With some 70 percent of Morocco's estimated population of 17,300,000 dependent on traditional agriculture (farming and herding), and nearly 50 percent of it under 15 years of age, the country has a soaring

Morocco, Area and Population*

	area		population	
	sq mi	sq km	1971 census	1975 estimate
Prefectures				
Casablanca	716	1,855	1,719,000	2,011,000
Rabat-Salé	811	2,100	642,000	746,000
Provinces				
Agadir	7,280	18,855	764,000	848,000
al-Hoceima	1,371	3,550	247,000	273,000
Azilal	3,880	10,050	327,000	365,000
Beni Mellal	2,732	7,075	442,000	490,000
Boulemane	5,710	14,790	116,000	130,000
Chaouèn	1,680	4,350	248,000	279,000
el-Jadida	2,317	6,000	591,000	656,000
el-Kelaa des Srarhna	3,888	10,070	461,000	516,000
Essaouira	2,446	6,335	343,000	405,000
Fès	3,917	10,145	1,003,000	1,130,000
Figuig	21,618	55,990	88,000	99,000
Kenitra	3,400	8,805	935,000	1,044,000
Khemisset	3,207	8,305	348,000	388,000
Khenifra	4,292	11,115	246,000	275,000
Khouribga	1,641	4,250	328,000	372,000
Ksar es-Souk	23,006	59,585	333,000	367,000
Marrakech	5,697	14,755	993,000	1,109,000
Meknès	3,286	8,510	604,000	687,000
Nador	2,367	6,130	481,000	531,000
Ouarzazate	22,004	56,990	528,000	581,000
Oujda	7,992	20,700	587,000	670,000
Safi	2,813	7,285	555,000	595,000
Settat	4,419	11,445	671,000	744,000
Tanger	461	1,195	286,000	331,000
Tarfaya	16,765	43,420	72,000	80,000
Taza	5,971	15,465	531,000	588,000
Tétouan	2,326	6,025	541,000	607,000
Tiznit	9,106	23,585	350,000	389,000
Total Morocco	177,117†	458,730	15,379,000‡	17,305,000‡

*Figures exclude area and population of the Moroccan-occupied portion of former Spanish (Western) Sahara, whose status was not finally determined in international law by early 1977. †Converted area figures do not add to total given because of rounding.
‡Figures do not add to total given because of rounding.
Source: Official government figures.

population and an almost stagnant economy—a combination that threatens to pull its standards of living still further behind those of the industrialized nations. The economic development of Morocco is also complicated by its ethnic and linguistic diversity; in addition to Berber and Arab peoples, there are also relatively small but significant European, Jewish, Algerian, and Negro communities. (See CASABLANCA; for associated physical features, see ATLAS MOUNTAINS and SAHARA DESERT; for historical aspects, see NORTH AFRICA, HISTORY OF; for Moroccan arts, see ISLAMIC PEOPLES, ARTS OF.)

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Surface features.* Most of Morocco lies at a high altitude; the average height above sea level is about 2,600 feet. Two chains of mountains, the Rif along the northern coast and the Atlas in the centre, divide eastern Morocco (which physically is much like neighbouring Algeria) from Atlantic Morocco, which the mountains cut off from the rest of northwestern Africa. Eastern Morocco and Atlantic Morocco are connected only by the narrow Taza Pass in the northeast. These systems of mountains were formed during the Tertiary Period—from 2,500,000 to 65,000,000 years ago—by the folding and uplift to sediment that had accumulated in the ancient sea that at that time edged the northern coast of the old continent of Africa.

The Rif, a complex structure, is geologically part of the cordilleras (mountain chains) reaching southward from the Iberian Peninsula of Europe, from which Africa was separated only after the Tertiary Period. Crescent-shaped, almost entirely over 4,900 feet and rising to 8,058 feet (2,458 metres) in Jebel Tidirhine, the limestone peaks of the Rif rise abruptly from the narrow Mediterranean shore.

The Atlas Mountains

The Atlas comprise three quite distinct chains. The Haut Atlas (High Atlas), 460 miles long, begins as small hills at the edge of the Atlantic, rises rapidly to over 6,500 feet, and reaches its highest point in Jebel Toubkal, which is 13,665 feet (4,165 metres) high. The Moyen Atlas (Middle Atlas) slopes away from the Haut Atlas in a northerly direction, rising to 10,794 feet (3,290 metres). The Anti-Atlas extends southwestward from the Haut Atlas to the Atlantic.

To the east of the Atlas, are the flat Hauts Plateaux (High Plateaus) of eastern Morocco, which are between about 3,900 and 4,250 feet in altitude, and which were formed by depressions which became filled with rubble. Atlantic Morocco comprises plains formed of sediment, and plateaus consisting of rubble.

The vast plain of the Rharb, at the foot of the Rif, is an old pass, which has been narrowed by the upheaval of the Rif and has then gradually been silted up by alluvial deposits from the Oued Sebou (a *oued*, or wadi, is a seasonal watercourse). Inland, the plains of Fès, Oujda, and the Angad become progressively smaller. The vast region lying between the Rharb, the Atlantic, and the Atlas, is named the Moroccan *meseta* ("plateau"); it is covered with a thin layer of sediments, thus forming a coastal plateau about 700 feet high running from Rabat to the coastal range of the Haut Atlas.

The coastline is regular, but, before modern ports were built, a strong sandbar made navigation difficult. To the south, the Sous plain separates the Haut Atlas from the Anti-Atlas, south of which begins the great expanse of the Sahara.

The rivers

Drainage. The Saharan area of eastern Morocco is drained by the Dadès, Rheris, Ziz, and Guir rivers, all of which originate in the Haut Atlas and lose themselves in the desert after cutting impressive chasms through the mountains. The northern region is drained by the Moulouya, which begins in the Moyen Atlas and empties into the Mediterranean after flowing for 320 miles. The Rif is drained by small torrential rivers such as the Laou, the Rhis, and the Nekor. Among the rivers flowing into the Atlantic are the Loukkos, which has its source in the western Rif and enters the ocean at Larache; the Sebou, the most strongly flowing river in Morocco, which is 280 miles long, and reaches the ocean at Mehdiya after much meandering; and the Oum er-Rbia, somewhat longer than the Sebou, and the most regular of Morocco's rivers.

All Moroccan rivers are torrential in nature: the Moulouya's flow ranges from 35,000 cubic feet per second to 175 cubic feet per second in the course of the summer.

Climate. In the coastal regions, the Moroccan climate is temperate within a narrow climatic range. Because of its latitude, climate is influenced by the proximity of the Azores anticyclone (a system of winds that rotates in a clockwise direction about a centre of high atmospheric pressure). The west winds that blow for most of the year result in changeable weather and irregular rainfall in the autumn and winter months. During the summer months, this anticyclone moves north, resulting in calmer weather. The cold ocean current, which flows to Morocco from the Canary Islands to the southwest, also helps to produce a milder climate.

In contrast to the coastal pattern, the inland climate assumes an increasingly continental character as the Atlas and the eastern inland plains are approached. Sometimes, but very rarely, the desert wind known as the chergui ("that which blows from the east") can cause the temperature to rise to 95° to 104° F (35° to 40° C) in the mountains and even the coast. Tangier has an average temperature of about 54° F (12° C) in January, and 75° F (24° C) in August; Casablanca's temperature averages 54° F (12° C) and 73° F (23° C) for the same months, while Essaouira has temperatures of 59° F (15° C) and 68° F (20° C). Inland the range is greater: the average at Meknès ranges between 50° and 77° F (10° and 25° C), and at Taza between 52° and 82° F (11° and 28° C). In the mountains the thermometer may plunge to −4° F (−20° C) at the summits or soar to 104° F (40° C) when the chergui is blowing.

Temperatures and rainfall

Rainfall depends on altitude and latitude; the Rif and the Moyen Atlas in the north receive an annual precipitation of over 47 inches; the northern Atlantic plains receive between 16 and 32 inches, which is just enough to cultivate corn without irrigation. To the south, and up to the Atlas, the rainfall average is steady at between eight and ten inches. To the east it ranges from six to 14 inches, whereas in the Sous it is less than eight inches. The rainy season usually lasts from November to January and sometimes continues into March. Rainfall is often irregular and unevenly distributed. Torrential downpours sometimes cause catastrophic floods, particularly in the Rharb, and in the Sous and Tafilalt regions, just north of the Sahara. Snow falls above about 6,500 feet and melts in the spring, by June or July at the latest.

Vegetation and animal life. The vegetation of Morocco is typical of that of the western Mediterranean region and similar to that of the Iberian Peninsula. In the mountain regions are stands of cedars, firs, and junipers. South of Essaouira, the tropical argan tree, which has an olive-like fruit that produces oil, covers 300,000 acres. The commonest trees are the cork oak, which covers 740,000 acres in the Rif and in the Rharb; the evergreen oak, which covers a third of the wooded areas; and the thuja, a pine, which covers 2,000,000 acres in the lower Moulouya Valley and in the high areas of the Sous. The wild olive and the mastic tree are found on high plateaus where there is little rainfall; the dwarf palm grows on the arid inland plains; the fruit-bearing jujube tree is common in the Tadla and the Haouz regions. Tough esparto grass (often used for making cordage and clothing) sometimes mixed in with wormwood covers a vast area. Unless they are grown under artificial conditions, as in the Marrakech area, date palms mark the northernmost limit of the desert, south of the Oued Drâa and east of the Atlas, at Boudenib and Figuig. Such species as the Barbary fig and the eucalyptus are imports; the former was introduced from the Americas via Spain in the 16th century, the latter by the French early in the 20th century.

The animal life is also Mediterranean: camels and horses were introduced at the dawn of history. Lions have disappeared, as the elephants did before them; mouflon (a wild sheep), gazelle, fennec (a kind of fox), and macaco (a kind of monkey) are still common in the Atlas region.

Human imprints on Morocco. *Traditional regions.* Morocco has five great regional units—the Rif, the Rharb, the Haouz, the Sous, and eastern Morocco. The Rif, a mountainous country, is well watered and inhabited by settled farmers, mostly tree cultivators, who have a strong sense of clannish loyalty. The Rharb, in the wider sense of the term, is centred on Fès and Meknès and comprises a unit in which the economy of the Berber-speaking shepherds of the Moyen Atlas, who practice transhumance (the seasonal migration of herds to mountain pastures), mingles with that of the settled Arabic-speaking grain producers of the Atlantic coastal region and that of the seminomadic cattle breeders from the central mountain ranges.

The five main regions

The south–central region, or Haouz, now looks to Casablanca as its urban centre, but its capital was once Marrakech. This region brings together peasants long settled on the soil of the Doukkala and the Abda plains, seminomadic cattle breeders of the central plateaus and the inland Tadla plains, and settled farmers of the Dir area (dirs are conical deposits left by seasonal streams) and the valleys of the Haut Atlas.

The Sous in the historic sense includes the Haut Atlas, the Sous plain, and the Anti-Atlas. This is the land of the Shleuh (Chleuh) who are a Berber people; its high population density has brought about extremely complex community organization. The region is traditionally associated with the production of herbaceous oils, saffron, and the exquisite metalwork traditional to Morocco.

Finally, eastern Morocco covers the region known historically as Tafilalt, in which the desert's influence is strongly felt. Settled inhabitants of the irrigated oases, including a strong proportion of people of mixed negroid descent, or Ḥarāṭīn, live alongside nomadic sheep breeders; they earn their living as agricultural labourers. The great nomads, the camel drivers such as the Regeibat and the Tekna tribes, live only temporarily within the territorial limits of Morocco.

Modes of life. In Morocco, settled mountain people living in villages, with architecture adapted to local conditions, have always lived alongside the nomadic or seminomadic peoples of the plateaus and of the Moyen Atlas, who live in black tents (*khaymah*). Other more or less settled farmers, who are tied to the soil, live together in thatched-roofed dwellings called *noualas*. Nomadism currently is declining rapidly, and the *khaymah* and the *nouala* are being replaced with solid houses, particularly in the irrigated Triffa, Rharb, Tadla, and Haouz plains. The precarious valley economy of the Atlas and the Rif is hard hit by the migration of mountain people to the coast, and the villages are becoming empty of men of working age. As a result, vast areas may become desert as

MAP INDEX

Political subdivisions

Agadir 30·24n 9·36w
al-Hoceima 35·15n 3·56w
Beni Mellal 32·20n 6·21w
Casablanca 33·37n 7·35w
el-Jadida 33·15n 8·30w
Fès 34·02n 4·59w
Kenitra 34·16n 6·36w
Khouribga 32·53n 6·54w
Ksar es-Souk 31·56n 4·26w
Marrakech 31·38n 8·00w
Meknès 33·54n 5·33w
Nador 35·11n 2·56w
Ouarzazate 30·55n 6·55w
Oujda 34·40n 1·54w
Rabat 34·02n 6·50w
Safi 32·18n 9·14w
Tangier 35·48n 5·48w
Tarfaya 27·57n 12·55w
Taza 34·13n 4·01w
Tétouan 35·34n 5·22w

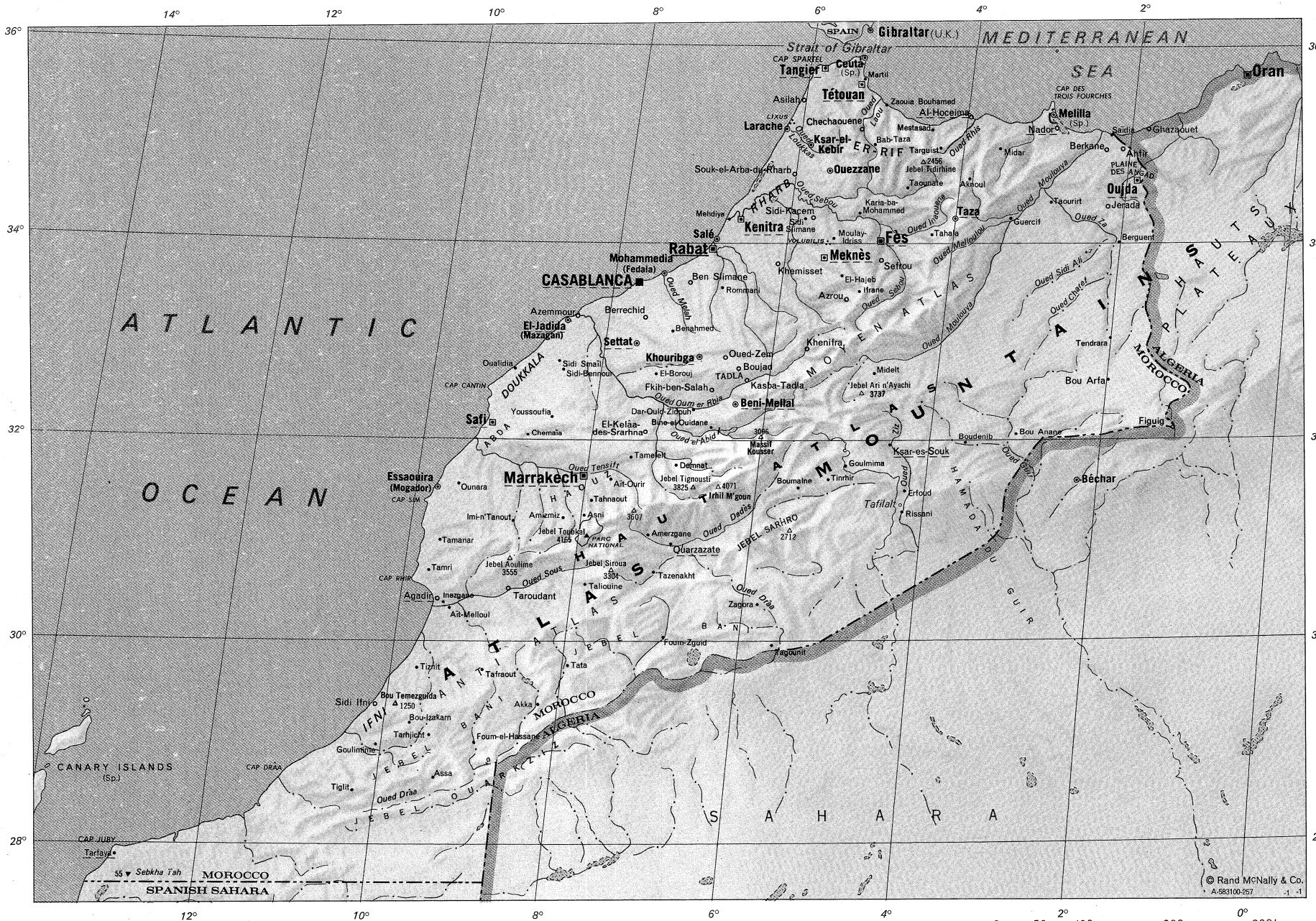
The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Agadir 30·24n 9·36w
Ahfir 34·58n 2·05w
Aït Melloul 30·21n 9·30w
Aït Ourir 31·34n 7·40w
Akka 29·25n 8·15w
Aknoul 34·39n 3·52w
Alcazarquivir, see Ksar el-Kebir
al-Hoceima 35·15n 3·56w
Amerzgane (Alhucemas) 31·03n 7·12w
Amizmiz 31·13n 8·15w
Asilah 35·28n 6·02w
Asni 31·15n 7·59w
Assa 28·37n 9·25w
Azemmour 33·17n 8·21w
Azrou 33·26n 5·13w
Bab Taza 35·04n 5·12w
Benahmed 33·04n 7·15w
Beni Mellal 32·20n 6·21w
Ben Slimane 33·37n 7·07w
Berguent 34·01n 2·01w
Berkane 34·56n 2·20w
Berrechid 33·16n 7·35w
Bine el-Ouidane 32·08n 6·28w
Bou Anane 32·02n 3·03w
Bou Arfa 32·32n 1·57w
Boudenib 31·57n 3·36w
Bou Izakarn 29·10n 9·44w
Boujad 32·46n 6·24w
Boumalne 31·32n 5·27w
Casablanca 33·37n 7·35w
Chechaouene 35·10n 5·16w
Chemaïa 32·05n 8·37w
Dar Ould Zidouh 32·19n 6·54w
Demnat 31·44n 7·00w
el-Borouj 32·30n 7·12w
el-Hajeb 33·42n 5·22w
el-Jadida (Mazagan) 33·15n 8·30w
el-Kelâa des Srarhna 32·03n 7·24w
Erfoud 31·26n 4·14w
Essaouira (Mogador) 31·31n 9·46w
Fedala, see Mohammedia
Fès 34·02n 4·59w
Figuig 32·06n 1·14w
Fkih ben Salah 32·30n 6·42w
Foum el-Hassane 29·02n 8·55w
Foum Zguid 30·05n 6·52w
Goulimime 28·59n 10·04w
Goulmima 31·42n 4·57w
Guercif 34·14n 3·22w
Ifrane 33·32n 5·06w
Imi n'Tanout 31·10n 8·51w
Inezgane 30·21n 9·32w
Jerada 34·19n 2·09w
Karia ba Mohammed 34·22n 5·12w
Kasba Tadla 32·36n 6·16w
Kenitra 34·16n 6·36w
Khemisset 33·49n 6·04w
Khenifra 32·56n 5·40w
Khouribga 32·53n 6·54w
Ksar el-Kebir 35·00n 5·59w
Ksar es-Souk 31·56n 4·26w
Larache 35·12n 6·09w
Marrakech 31·38n 8·00w
Martil 35·37n 5·17w
Mazagan, see el-Jadida
Mehdiya 34·15n 6·41w
Meknès 33·54n 5·33w
Mestasa 35·07n 4·25w
Midar 34·57n 3·32w
Midelt 32·41n 4·45w
Mogador, see Essaoira
Mohammedia (Fedala) 33·42n 7·24w
Moulay Idriss 34·03n 5·31w
Nador 35·11n 2·56w
Oualidia 32·44n 9·02w
Ouarzazate 30·55n 6·55w
Oued Zem 32·52n 6·34w
Ouezzane 34·48n 5·36w
Oujda 34·40n 1·54w
Ounara 31·32n 9·33w
Port Lyautey, see Kenitra
Rabat 34·02n 6·50w
Rissani 31·17n 4·17w
Rommani 33·32n 6·36w
Safi 32·18n 9·14w
Saïdia 35·04n 2·13w
Salé 34·04n 6·48w
Sefrou 33·50n 4·50w
Settat 33·00n 7·37w
Sidi Bennour 32·39n 8·26w
Sidi Ifni 29·23n 10·10w
Sidi Kacem 34·13n 5·32w
Sidi Slimane 34·16n 5·56w
Sidi Smaïl 32·50n 8·31w
Souk el-Arba du Rharb 34·41n 5·59w
Tafraout 29·43n 9·00w
Tagounit 29·58n 5·35w
Tahala 34·03n 4·25w
Tahnaout 31·21n 7·57w
Taliouine 30·32n 7·55w
Tamanar 31·00n 9·40w
Tamelelt 31·48n 7·31w
Tamri 30·43n 9·50w
Tangier (Tanger) 35·48n 5·48w
Taounate 34·33n 4·39w
Taourirt 34·25n 2·54w
Tarfaya 27·58n 12·55w
Targuist 34·57n 4·18w
Tarhjicht 29·04n 9·25w
Taroudant 30·29n 8·52w
Tata 29·45n 7·59w
Taza 34·13n 4·01w
Tazenakht 30·35n 7·12w
Tendrara 33·03n 2·00w
Tétouan 35·34n 5·22w
Tiglit 28·31n 10·15w
Tinrhir 31·31n 5·32w
Tiznit 29·43n 9·43w
Xauen, see Chechaouene
Youssoufia 32·15n 8·32w
Zagora 30·19n 5·50w
Zaouia Bouhamed 35·19n 4·58w

Physical features and points of interest

Abda, *plain* 32·10n 9·05w
Abid, Oued el-, *river* 32·17n 7·03w
Angad, Plaine des, *plain* 34·43n 1·50w
Anti-Atlas, *mountains* 30·00n 8·30w
Aoulime, Jebel *mountain* 30·50n 8·50w
Atlantic Ocean 32·00n 12·00w
Atlas Mountains 32·00n 5·00w
Ayachi, Ari n', *mountain* 32·29n 4·57w
Bani, Jebel, *mountains* 29·30n 8·00w
Bou Ajam, Oued, *watercourse* 33·36n 1·31w
Bou Temezguida, *mountain* 29·11n 9·55w
Cantin, Cap, *cape* 32·34n 9·19w
Charef, Oued, *watercourse* 34·05n 2·03w
Dadès, Oued, *watercourse* 30·55n 6·47w
Doukkala, *historic region* 32·55n 8·50w
Drâa, Cap, *cape* 28·44n 11·05w
Drâa, Oued, *watercourse* 28·43n 11·09w
Gibraltar, Strait of 35·57n 5·36w
Guir, Hamada du, *desert* 31·20n 3·30w
Guir, Oued, *watercourse* 31·40n 2·55w
Haouz, *plain* 31·30n 8·00w
Haut Atlas, *mountains* 31·30n 6·00w
Ifni, *historic region* 29·20n 10·10w
Inaouene, Oued, *river* 34·14n 4·55w
Jebel Toubkal Parc National, *national park* 31·03n 7·55w
Juby, Cap, *cape* 27·57n 12·55w
Kousser, Massif, *mountain* 32·03n 6·02w
Laou, Oued, *watercourse* 35·26n 5·05w
Loukkos, Oued, *watercourse* 35·12n 6·09w
Mediterranean Sea 36·00n 3·00w
Melah, Oued, *river* 33·43n 7·24w
Melloulou, Oued, *river* 34·13n 3·21w
M'goun, Irhil, *mountain* 31·30n 6·27w
Moulouya, Oued, *river* 35·06n 2·20w
Moyen Atlas, *mountains* 33·30n 5·00w
Ouarkziz, Jebel, *mountains* 28·30n 9·30w
Oum er-Rbia, Oued, *river* 33·19n 8·20w
Rharb, *historic region* 34·30n 6·02w
Rhir, Cap, *cape* 30·38n 9·54w
Rhis, Oued, *watercourse* 35·12n 3·53w
Rif, er-, *mountains* 35·00n 4·00w
Sarhro, Jebel, *mountains* 31·00n 6·00w
Sebkha Tah, *dry lake* 27·42n 12·48w
Sebou, Oued, *river* 34·16n 6·41w
Sidi Ali, Oued, *watercourse* 34·07n 2·05w
Sim, Cap, *cape* 31·23n 9·51w
Siroua, Jebel, *mountain* 30·43n 7·37w
Sous, Oued, *watercourse* 30·22n 9·37w
Spartel, Cap, *cape* 35·48n 5·56w
Tadla, *plain* 32·30n 6·20w
Tafilalt, *oases* 31·18n 4·18w
Tensift, Oued, *river* 32·02n 9·21w
Tidirhine, Jebel, *mountain* 34·51n 4·31w
Tignousti, Jebel, *mountain* 31·30n 6·42w
Tigzert, Oued, *watercourse* 28·25n 9·13w
Trois Fourches, Cap des, *cape* 35·26n 2·58w
Volubilis, *ruins* 34·04n 5·33w
Za, Oued, *river* 34·33n 3·02w
Ziz, Oued, *watercourse* 30·39n 4·26w


MOROCCO
Size of symbol indicates relative size of town
Elevations in metres
MEDITERRANEAN SEA
ATLANTIC OCEAN
Strait of Gibraltar
SPAIN
Gibraltar (U.K.)
Tangier
Ceuta (Sp.)
Tétouan
Melilla (Sp.)
Oran
Al-Hoceima
Nador
Berkane
Oujda
Ghazaouet
Saidia
Ahfir
Larache
Ksar-el-Kebir
Chechaouene
Asilah
Souk-el-Arba-du-Rharb
Ouezzane
Taounate
Taza
Fès
Meknès
Sefrou
Ifrane
Azrou
Midelt
Taourirt
Guercif
Berguent
Jerada
Kenitra
Salé
Rabat
Mohammedia (Fedala)
CASABLANCA
Ben Slimane
Rommani
Khemisset
Khenifra
Tendrara
Bou Arfa
Figuig
Azemmour
El-Jadida (Mazagan)
Berrechid
Settat
Khouribga
Oued-Zem
Boujad
TADLA
Kasba-Tadla
Beni-Mellal
Fkih-ben-Salah
El-Borouj
Oualidia
Sidi Smail
Sidi-Bennour
Safi
Youssoufia
Chemaia
Dar-Ould-Zidouh
Bine-el-Ouidane
El-Kelâa-des-Srarhna
Massif Kousser
Boudenib
Bou Anane
Ksar-es-Souk
Goulmima
Tinrhir
Boumalne
Erfoud
Rissani
Tafilalt
Béchar
Essaouira (Mogador)
Ounara
Marrakech
Tamelelt
Demnat
Jebel Tignousti
Irhil M'goun
Tahnaout
Imi-n'Tanout
Amizmiz
Asni
Jebel Toubkal
Amerzgane
Ouarzazate
JEBEL SARHRO
Tamanar
Tamri
Jebel Aoulime
Jebel Siroua
Tazenakht
Taliouine
Taroudant
Agadir
Inezgane
Ait-Melloul
Tiznit
Tafraout
Tata
Foum-Zguid
Tagounit
Sidi Ifni
Bou Temezguida
Akka
Bou-Izakarn
Tarhjicht
Foum-el-Hassane
Goulimine
Assa
Tiglit
CANARY ISLANDS (Sp.)
Tarfaya
Sebkha Tah
SPANISH SAHARA
SAHARA
ALGERIA
MOROCCO
ANTI ATLAS
HIGH ATLAS
MIDDLE ATLAS
TELL ATLAS
HAUTS PLATEAUX
HAMADA DU GUIR
CAP SPARTEL
CAP DES TROIS FOURCHES
CAP TRAFALGAR
CAP SIM
CAP RHIR
CAP DRAA
CAP JUBY
Oued Laou
Oued Rhis
Oued Moulouya
Oued Sebou
Oued Melloulou
Oued Sidi Ali
Oued Charef
Oued Tensift
Oued Oum er Rbia
Oued el Abid
Oued Dadès
Oued Draa
Oued Ziz
Oued Guir
Oued Saoura
Oued Zousfana
Oued Saoura
Oued Guir
Oued Ziz
Oued Rheris
Oued Dadès
Oued Draa
Oued Sous
Oued Tensift
Oued Melloulou
Oued Sebou
Oued Laou
Oued Rhis
Oued Moulouya
Oued Sidi Ali
Oued Charef
Oued Tensift
Oued Oum er Rbia
Oued el Abid

already has happened in eastern Morocco and the Haouz, which now are haunted only by shepherds, the last people to use *khaymahs*.

Towns have traditionally been divided into three groups —port towns, royal towns, and market towns—each with its distinctive style of life. Recent developments have tended to increase the prosperity of port towns throughout the country. Towns in which economic life has declined, such as Ouarzazate and Tiznit in the south, Chechaouene and Tétouan in the north, or Essaouira or Asilah on the coast, maintain the rhythms of the traditional way of life and hide great poverty with dignity.

Rural and urban settlement

Towns experiencing a certain degree of expansion, however, whether they are old like Meknès and Oujda or rebuilt like Casablanca and Safi, have not yet established their new balance and possess a characteristically unfinished quality. These towns are composed of five distinct parts that exist side by side without integrating with one another. The old centre, the *madīnah*, is surrounded by ramparts, which still exist almost everywhere. Overcrowded by immigrants from outlying areas, it usually has fallen into decline, despite efforts by local authorities to prevent it from total decay. The business and entertainment centre, the modern town inhabited mostly by foreigners, is very busy by day but empties rapidly at night. The constantly expanding residential area indicates the rapid rise of the local middle class. Flower-bedecked, lighted, and clean, it consists of a succession of separate houses, each surrounded by a large garden.

The new *madīnah*, a tangle of narrow alleys with tiny houses built on the same model and with little planning or sense of style, is packed with low-ranking officials, small shopkeepers, workers, and pensioners. It unceasingly attempts to check the growth of the shanty towns, or *bidonvilles*, which expand constantly despite all efforts to resettle their inhabitants, who comprise newcomers to the town, the unemployed, and casual labourers who remain on the fringes of urban life. *Bidonvilles* and the newer quarters of the new *madīnah* often partly overlap, depending on the whims of financial speculators. This pattern of separation, common to most Moroccan cities, brings about an atmosphere that lacks the extroversion and tumult usually characteristic of a large city.

THE PEOPLE OF MOROCCO

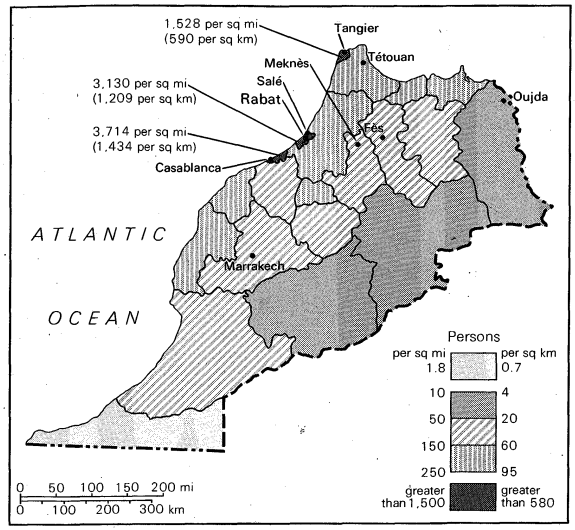
Arabs and Berbers

Linguistic and ethnic groups. The population of Morocco is Arab and Berber in origin, though the difference between the two groups today is more linguistic than racial. The Berber language, although greatly influenced by Arabic, has been preserved in the mountainous regions, which have long been areas of refuge. Berber-speaking inhabitants are divided into three dialect groups: Rifian in the Rif, Amazigh in the Moyen Atlas, and Shleuh in the Haut Atlas and the Sous. With improved communications and the exodus to the towns, Arabic-Berber bilingualism has become common, and the proportion of monolingual Berbers does not exceed 25 percent. The remainder of the Moroccan population speaks Arabic and is composed mainly of Arabized Berbers from the plains and the plateaus, and a small number of Bedouin Arabs who moved to Morocco with the armies of Islām in the 7th century or with the invading Hilāl tribe in the 11th and 12th centuries.

Immigrants from Spain, who travelled to Morocco at various stages of the Christian reconquest of the Iberian Peninsula, which culminated in 1492, were themselves a mixture of Arab, Berber, and Iberian; today they live chiefly in Rabat, Salé, Fès, and Tétouan.

Trade and cultural relations with black Africa since the Middle Ages have resulted in the gradual growth of a significant Negro minority, living chiefly in the southern oases or in the royal towns, particularly Marrakech and Meknès. The Jewish minority—a considerable community until emigration to Israel began in 1948—consists of small Berber-speaking groups whose origins are obscure and of Arabic-speaking groups who arrived from Spain and other Mediterranean countries at various periods of history.

European colonization brought a French and a Spanish minority whose numbers had reached some 500,000 on the eve of independence in 1956. The French and Spanish languages, which spread among the urban population during the preceding protectorate period, are still spoken. With the exception of the rapidly declining gallicized Jewish minority, however, the majority of the Moroccan population has a common culture and religion. Most Moroccans are Sunnī Muslims of the Mālikite rite, who, until recently, had a marked inclination to join religious confraternities of varying degrees of orthodoxy.



Population density of Morocco.

Contemporary demography. The 1971 census put Morocco's Islāmic population at about 15,250,000. The Jewish population, which in 1960 numbered about 160,000, had dropped to 31,000, while foreigners of all nationalities, who previously numbered nearly 400,000, dropped to only 112,000. The Islāmic population had one of the highest rates of increase in the world, estimated at 3.2 percent per year. The birth rate was close to 50 per 1,000 inhabitants per year, whereas the death rate stood at about 16.5 per 1,000; infant mortality reached 15 percent. These figures explain the extreme youth of the population, more than 45 percent of which is under 15 years of age, and more than 55 percent, under 20; as a result, the productive portion of the population amounts to only about 28 percent.

In the early 1970s, some 5,000,000 Moroccans lived in towns whose populations were growing at an annual rate of about 6 percent. Casablanca had a population of 1,500,000 and 9 other towns—Fès, Kenitra, Marrakech, Meknès, Oujda, Rabat, Safi, Tangier, and Tétouan —had populations of more than 100,000. Population density varies outside urban areas: the Doukkala and western Rif regions have an average of 180 inhabitants per square mile; the Anti-Atlas and the Marrakech area, about 50 to 150; and the eastern and southern steppe regions, about 13. The present trend is toward a massive concentration of people in the coastal area between Kenitra and Casablanca.

The population increase

The government has launched a family-planning campaign that the population does not seem to oppose, although the youthful composition of the population suggests that rapid results cannot be expected. Population figures projected for 1985 range from 23,000,000 to 26,000,000, depending on the program's success or failure; it is estimated that by that time about 10,000,000 people will live in the towns, 7,000,000 to 8,000,000 will be of school age; and 8,000,000 jobs will be needed, 4,000,000 of them in the towns. Such projections are alarming, considering the slow rate of economic growth in the 1960s.

Large-scale emigration to Europe remains a continuing phenomenon. After France—the traditional labour market—Belgium, The Netherlands, and West Germany are using increasing amounts of Moroccan manpower. Training centres, jointly financed by the state and by European

employers, are open to those wishing to emigrate. In the early 1970s, 240,000 Moroccans were working in Europe, most of whom were originally from the overpopulated Atlas, Rif, and Tafilalt regions.

THE NATION'S ECONOMY

Economists were optimistic after World War II about Morocco's future. The subsequent population explosion and indecisiveness concerning economic policies in the decade that followed independence, have made them more cautious. From the point of view of area, population, resources, and geographic location, Morocco is a medium-sized country that might be expected to play a significant role in a regional framework. Stagnation of the national product, small individual incomes, and poor development indices threaten Morocco with increasing poverty as the years go by.

Components of the economy. In spite of such gloomy economic prognostications, Morocco's mineral, agricultural, and hydroelectric resources are far from negligible. It has about 40 percent of the world's known phosphate reserves and is the world's second-largest producer. Mining is under the direction of the Office Chérifien des Phosphates, a public body founded in 1921, and exports are nearly all in the form of unrefined ores. The discovery of more easily exploited deposits in the Spanish Sahara, however, is providing increasing competition for the Moroccan mines. Other mineral products include iron, lead, manganese, zinc, cobalt, and copper. A pocket of natural gas was discovered near Essaouira in 1961.

Mining, agriculture, and industry

Petroleum resources in the Rharb are being rapidly depleted, necessitating the importation of most of the crude oil processed by Morocco's two refineries. Domestic coal is a major source of fuel for the nation's power plants; in the early 1970s only about one-half of the hydroelectric potentiality of the rivers was being utilized.

Some 38,300,000 acres of land are under cultivation, but irregular harvests and poor yields force Morocco to import a large percentage of its foodstuffs. To help offset this trade imbalance, the government is pinning its hopes on increasing production of edible beans, cereals, cotton, and sugar from cane and beet. Vines are cultivated, mainly by Frenchmen, for the export wine trade. Citrus fruits, tomatoes, vegetables, and canned fish comprise the largest portion of exports. Other export products include olive oil, cork, oak timber, esparto grass, and sardine and tuna from the rich coastal fishing grounds. Income from livestock—sheep, cattle, goats, and camels—represents one-third of all agricultural income.

About 600,000 Moroccans either work in industry, mines, or handicrafts. Automobile assembly, textile spinning and weaving, and food processing are among the more modern industrial processes; textiles and processed foods provide surpluses for export. The ancient Moroccan handicraft products, which include carpets, hand-tooled leather, and intricate metalwork, are also produced in sufficient quantity for export.

In the early 1970s Morocco continued to have a balance of trade deficit, but this was offset by aid from various Western nations and from international organizations. Most of the nation's trade—about 50 percent of imports, and 60 percent of exports—was with members of the European Economic Community, or the Common Market.

Economic management. The private sector of the Moroccan economy is dominated by the European minority, mainly French, which controls production of a large proportion of the nation's wine, paper, metals, and processed foods. The new class of Moroccan entrepreneurs controls only the textile and hotel trades. In 1969 the state introduced a policy of accelerating the Moroccanization of banks and insurance agencies, and service agencies—by encouraging the conclusion of agreements between Moroccan nationals and foreigners, rather than by nationalization. Private interests are protected by a French chamber of commerce, a French–Moroccan young businessmen's association, and an American–Moroccan chamber of commerce.

While state activity in the economy continues to expand, its role is primarily to stimulate the private sector. In order to encourage and guarantee investment of foreign currency in the economy, liberal benefits, mainly to foreign private investors, have been available since 1958; Moroccan ministerial committees meet regularly with their counterparts from West Germany, France, and Spain to study aid proposals within the framework of current development plans.

Among such plans, one for rapid industrialization in the early 1960s was abandoned in order to concentrate on developing irrigation for agriculture, so as to stimulate exports; on promoting tourism as a national industry; and on the training of administrators and technicians to decrease the degree of extensive foreign assistance still needed in these areas. Morocco's alliance with the Common Market has resulted in more secure markets, as well as in more secure jobs for its surplus manpower; it has also helped to attract tourists and capital.

Within the framework of current policies, annual growth rates in the gross national product of from 3.5 to 5 percent are forecast for the 1970s. Given this rate, however, unemployment would probably remain at more than 1,000,000 persons, and the surplus agrarian population, at about 40 percent, while education would be available to only 30 percent of the school-age population. Meanwhile, a heavy burden on the economy is the employment of about 350,000 persons—five times as many as before independence—as government functionaries of various kinds.

Wages are negotiated directly between management and the trade unions. In keeping with its guarantees of monetary stability, the government has in effect kept wages frozen since 1960; this policy is made easier by the great surplus of labour. The major trade union is the Union Marocaine du Travail (Casablanca); its membership, estimated at more than 350,000, is mainly in the mining, textile, and food-processing industries.

Transportation. The transportation network radiates outward from the regional capitals—Oujda, Fès, Marrakech, Agadir, and Tangier—to focus on Casablanca, Morocco's economic and commercial capital. Casablanca handles four-fifths of Moroccan imports, but the free port of Tangier is the main landing place for tourists.

The ports

Mineral products from eastern Morocco are shipped through the Algerian port of Ghazaouet (formerly Nemours); products from the Rharb region, stockpiled at Fès and Meknès, are exported through Casablanca and Kenitra; the phosphate is refined through Safi; the Sous region ships its products through the port of Agadir; and the port of Tangier serves the Rif region. Because the two ports of Ceuta and Melilla are still in Spanish hands, the Moroccan government is planning to build a new Mediterranean port at Nador. Mohammedia is the principal port handling petroleum imports. The national shipping company, Comanav, owns seven ships.

The railway system, which is nationalized, has about 1,632 miles of track, of which about 472 miles—which includes the Marrakech–Casablanca–Fès section—are electrified. The system includes part of the main line from Algiers to Marrakech, via Oujda, Taza, Fès, Meknès, Rabat, and Casablanca; there are extensions northward to Tangier and southward to the phosphate mines of Khouribga, as well as an extension to Safi. The eastern Moroccan line serves mining areas in that region.

Air traffic is being rapidly expanded; services provided by the national airline are supplemented by those of a second company, which provides flights on inland routes. There are international airports at Nouasser near Casablanca, and at Rabat-Salé; those at Agadir and Tangier are being enlarged to accommodate all categories of commercial aircraft. Smaller but well-equipped airports are at Marrakech, Fès, Oujda, and al-Hoceima.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. *Constitutional framework.* Morocco has been a constitutional monarchy since 1962. It experimented with a parliamentary type of government between 1963 and 1965, but after a period of five years during which full power devolved upon the king, the

constitution of 1970 (a new one was approved by a referendum in 1972) replaced the previous bicameral legislature with a single chamber having limited powers. This chamber comprises members elected by universal suffrage (men and women citizens 21 years of age or older); two-thirds of its members are elected by the municipal councils and committees, and one-third by organizations of professional people and wage earners. The chamber shares with the king the right to initiate legislation; it can, however, be dissolved at a moment's notice. The king, whose person is sacred and inviolable, has the right to submit laws directly to a popular referendum and can send back any law that has been adopted after a second reading in the chamber.

The centralization of local government

Administration in Morocco is highly centralized, and strictly speaking no local government exists. *Qā'īd*s (caids), the heads of local administrative units in rural areas, and pashas, administrators in the towns, are appointed by the Minister of the Interior, Town Planning and Housing; governors of the 19 provinces and of the prefectures of Casablanca and Rabat-Salé are appointed by royal decree. These representatives of the central government exercise tight control over the municipal and communal councils, and over the 21 prefectoral and provincial assemblies, and over the seven regional assemblies that have a consultative role in economic and social matters.

Political life. This centralization and the influence traditionally wielded by government on public life in an underdeveloped country that has widespread illiteracy and no long history of democratic rule has tended to alienate many factions within Moroccan society, including the intellectual and political elite, the trade unions, and youth. These groups feel that, in spite of legal guarantees, severe limitations exist on the citizens' freedom of choice. Supporters of the system of "guided democracy" and "electoral apprenticeship" suggest that the sociocultural backwardness of the country would make any freedom of choice illusory. There was widespread withdrawal from participation in the political process in the early 1970s, and unsuccessful coups d'etat in 1971 underlined instabilities in the nation's political life.

Of the four political parties, the oldest is the nationalist Istiqlāl Party, founded in 1944; it has a bent toward bold social reform in spite of its middle-class membership. The Union Nationale des Forces Populaires was formed in 1959 by left-wing secessionists; it has been the object of severe governmental repression. Two other parties are much in the minority. Campuses in such university cities as Rabat and Fès have tended to become increasingly politicized and antigovernment.

Justice. Although in theory the Qur'ān is the source of law, and the king is the spiritual head of the community, in effect the Qur'ānic jurisdiction, exercised by the *qāḍī*s or cadis (Muslim judges who interpret and administer the religious law of Islām), is limited to matters relating to the personal status of Muslims; whereas rabbinical justice exists for Jews. All other matters, whether they concern Muslims, Jews, or others, are in the hands of the "modern" courts, which apply a code of laws inspired by the French. The various regional and local courts are beneath the three courts of appeal and the supreme court, which sits in Casablanca. All are supervised by the Ministry of Justice. A Franco-Moroccan judicial convention allows Morocco to use the services of 100 French magistrates. French and Spanish lawyers may in principle practice at the Moroccan bar. The Ministry of Justice has conducted its affairs in Arabic since 1965.

Military and police. The Moroccan Army is made up of contingents that previously served under the French or Spanish flags or that formed part of the guerrilla forces that fought against the French before independence was obtained in 1956. The 55,000-man army, divided into five brigades, is equipped with French and Russian tanks. Some 20,000 supplementary soldiers under the Ministry of the Interior, Town Planning and Housing form a militia, which is responsible for maintaining order in the interior. The air force is equipped with French, American, and Soviet fighters, and the navy has several light warships. The royal gendarmerie is made up of 3,500 well-trained men; and the police force, which receives technical assistance from France, numbers 12,000.

Standards of living

The social milieu. Morocco's low standard of living is characterized by great inequalities in the distribution of wealth. One-half of the national product is derived from the modern sector of the economy, which involves only 10 percent of the population. At the same time, 10,000,000 peasants contributed only 20 percent of the product. The average annual income of the great majority is about $80, although workers in private industry earn considerably more. Rising prices of dietary staples threaten to lower standards of living further. The hill peoples fare less well in most areas of social services.

Education. Although about one-fifth of Morocco's budget is spent on education, barely 50 percent of the nation's school-age population receives an education. The system of primary and secondary schools, in which boys greatly outnumber girls, is supplemented by traditional Islāmic Arabic schools, as well as by private schools, both Moroccan and foreign.

The French Cultural Mission, the most important foreign institution in the country, provides schooling to 10,000 Moroccan pupils at the primary and the secondary levels. French and Arab bilingualism at all levels constitutes a problem resulting in extra expense and waste of energy. Only about 40 percent of the teachers in the secondary schools and barely 15 percent in the scientific disciplines at the higher level are Moroccan. Programs promoting unification, Moroccanization, and Arabization of the educational system, long accepted in principle by the government, had not yet been put into effect in the early 1970s.

Health and welfare. Social services are the responsibility of the ministries of Public Health, Work and Social Affairs, and Youth and Sports. These services largely benefit the urban population, especially the civil servants and other middle-class groups. Neither craftsmen, members of the liberal professions, nor agricultural workers benefit from medical insurance or family allowances.

Public health receives about 6 percent of the Moroccan budget, but the entire country has only about 1,150 doctors and fewer than 1,000 nurses. Health centres and dispensaries are located throughout the country to supplement the work of 76 general and specialized hospitals. The faculty of medicine at Rabat (a second one is planned for Casablanca) produces only a small number of doctors in a country in which some 2,000,000 children suffer from the disease of malnutrition, which takes the form of rickets (a childhood disease characterized by faulty ossification of the bones).

Housing. The Ministry of the Interior, Town Planning and Housing is responsible for housing, often acting as trustee for communities whose large real-estate holdings are owned in common. The number of persons inadequately housed—in shantytowns or in the *madīnah*s—is estimated at 2,000,000; governmental policy is to leave the problem of urban housing to the private sector. The state parcels out public land and arranges easy terms of payment; in the early 1970s it was concentrating on improving rural housing in an effort to stop the exodus to the towns. Within the framework of the 1968 to 1972 plan, and with help from the United Nations, 30,000 dwellings were improved in the south and 60,000 built in irrigated areas. It was considered, however, that this relative indifference to urban housing problems on the part of the government would only be temporary.

CULTURAL LIFE AND INSTITUTIONS

Illiteracy and biculturalism are the major facts of Morocco's cultural life. More than three-quarters of young people between 15 and 19 years of age are illiterate. Divisions between town and country and between linguistic and dialectal areas persist. Arabic is the national and official language of Morocco, but in many areas, French is the working language. A heritage from the colonial past, biculturalism is reinforced by the presence of a strong and dominant foreign minority in the modern sector of the economy.

Duality of literary traditions

The arts. Two literatures—Arabic and French—exist side by side in Morocco. The future of literary works in French cannot be predicted with any degree of certainty, though many young Moroccans are following in the footsteps of older writers who used French. A modern Arabic literature is, however, also developing in spite of publishing and distribution difficulties. These works either have to be published abroad, generally in Beirut or Lebanon, or at the author's expense; editions rarely exceed 3,000 copies.

Painting and sculpture have experienced a veritable blossoming since independence.

Both primitive and abstract artists have found encouragement in foreign cultural centres and are exploiting an almost untouched market in the growing Moroccan middle class. The École Municipale des Beaux Arts (Municipal School of Fine Art) in Casablanca is striving to save Morocco's artistic tradition by reviving it. At the same time, the government, in an effort to safeguard a folklore that is rich and varied, organizes every spring a national folklore festival at Marrakech which attracts thousands of tourists. The theatre is subsidized, either by the local authorities as in Casablanca and el-Jadida, or by the Ministry of Youth, Sports, and Social Affairs. Conservatories of music are located in Casablanca, Rabat, and Tétouan; their dual purposes are both to preserve the Andalusian musical tradition and to spread a taste for modern Western music, as well as a knowledge of its techniques.

Cultural organization. The Department of Culture is responsible for libraries and archives; it classifies and restores historic monuments, sponsors art exhibitions, and awards the Prix du Maroc for literature, the social sciences, and the exact sciences. The association of Moroccan writers, which has branches in the main cities, attempts to disseminate the new Moroccan literature in collaboration with similar associations in the Maghrib and in other Arab countries.

Communications media. The bilingual Moroccan press has a small circulation: between 15,000 and 45,000 copies for the largest dailies. Among Arabic newspapers are *al-ʿAlam*, the Istiqlāl organ; *al-Anbāʾ*, the official government paper; and *ash-Shaʿb* (*al-Chaʿn*), which is government inspired. Among the French-language newspapers are *L'Opinion* (Istiqlāl); *La Dépêche* (independent; suppressed in 1971); and *Le Petit Marocain* and *La Vigie Marocaine*, which date from the protectorate period and were Moroccanized under new names, *Le Matin* and *Maroc-Soir*. Morocco has an independent press agency, the Maghreb Arabe Presse (MAP).

As in all countries with widespread illiteracy, the radio plays an important role. According to official figures, there is one receiver for every 16 Moroccans, but since the advent of transistorized receivers, it is probable that nearly every family owns a radio. A powerful station, Voice of America, broadcasts from Tangier, a medium-powered one from Rabat. Moroccan radio broadcasts for 12 hours daily in Arabic, Berber, French, Spanish, and English. The state television network in Rabat, inaugurated in 1962, broadcasts for four and a half hours daily in Arabic and in French; entertainment outweighs education in its programming.

PROSPECTS

Ongoing economic and social instabilities

The 1970s will be decisive for Morocco's future. If the nation contrives to take advantage of its geographical situation, its climate, its countryside, and its cultural patrimony to attract the capital it lacks, if it provides its young people with the adequate education and the work that they require, and if it is able to process its mineral and agricultural resources instead of simply exporting them as raw materials, then observers believe that Morocco might yet achieve a major economic and social turnabout. If it does not, illiteracy, unemployment, and underemployment could lead Morocco into poverty and decline.

These problems have not yet reached catastrophic proportions; between 1960 and 1970 the royal government tried to resolve them within a traditionalist, relatively liberal, and moderate framework. In this way it has been able to win the confidence of international organizations and to receive aid from the industrialized countries of the West; however, because of the concentration of power, administrative centralization, and the growing part played by the army in public affairs, those areas in which the political and intellectual elite could participate shrank just at a time when the problem required nationwide cooperation by all elements of Moroccan society.

Social stability is also called into question by deficiencies in economic growth. Should the birth-control campaign fail, which it is likely to do, the situation would then become aggravated. Most impartial observers felt that the government would then be forced either to stress its authoritarian character or to open the way to new forces.

BIBLIOGRAPHY. No complete, up-to-date monograph on all aspects of Moroccan life exists, although V. MONTEIL, *Le Maroc*, new ed. (1972) and N. BARBOUR, *Morocco* (1965), provide perspectives. Worthwhile sociological studies are regional rather than national in content; CARLETON S. COON, *Tribes of the Rif* (1927), is a major pioneer study. The point of view of the colonial administration is well documented in R. MONTAGNE, *Les Berbères et le Makhzen dans le Sud du Maroc* (1930). More methodical and liberal in inclination is JACQUES BERQUE, *Structures sociales du Haut Atlas* (1955). E. GELLNER, *Saints of the Atlas* (1969), is a remarkable study of a segmental society. A good general survey of the changes introduced by colonization is found in CHARLES F. STEWART, *The Economy of Morocco, 1912–1962* (1964). The INTERNATIONAL BANK FOR RECONSTRUCTION AND DEVELOPMENT, *The Economic Development of Morocco* (1966), is a report by international experts on the problems and perspectives of the Moroccan economy. The fundamental analysis of the subject, however, is found in SAMIR AMIN, *L'économie du Maghreb*, 2 vol. (1966), summarized in *Le Maghreb moderne* (1970; Eng. trans., *The Maghreb in the Modern World*, 1970). The sociopolitical situation under the protectorate is discussed with great finesse in JACQUES BERQUE, *Le Maghreb entre deux guerres* (1962; Eng. trans., *French North Africa: The Maghrib Between Two World Wars*, 1967). The history of the nationalist movement is traced faithfully by JOHN P. HALSTEAD in *Rebirth of A Nation: The Origins and Rise of the Moroccan Nationalism, 1912–1944* (1967). Two complementary points of view concerning the transformations and the problems following independence are found in DOUGLAS E. ASHFORD, *Political Change in Morocco* (1961); and WILLIAM I. ZARTMAN, *Morocco: Problems of New Power* (1964). The mechanisms of power are analyzed in great detail in CLEMENT H. MOORE, *Politics in North Africa* (1970); and JOHN WATERBURY, *Commander of the Faithful: The Moroccan Elite* (1970).

(A.La.)

Morphology

Morphology is a term used in biology for the study not only of shape and structure in plants, animals, and micro-organisms but also of the size, shape, structure, and relationships of the parts comprising them. The term morphology connotes the general aspects of biological form and arrangement of the parts of a plant or an animal. The term anatomy also refers to the study of biological structure but usually connotes study of the details of either gross or microscopic structure. In practice, however, the two terms are used almost synonymously.

Typically, morphology is contrasted with physiology, which deals with studies of the functions of organisms and their parts; function and structure are so closely interrelated, however, that their separation is somewhat artificial. Morphologists were originally concerned with the bones, muscles, blood vessels, and nerves comprising the bodies of animals and the roots, stems, leaves, and flower parts comprising the bodies of higher plants. The development of the light microscope made possible the examination of some structural details of individual tissues and single cells; the development of the electron microscope and of methods for preparing ultrathin sections of tissues created an entirely new aspect of morphology—that involving the detailed structure of cells. Electron microscopy has gradually revealed the amazing complexity of the many structures comprising the cells of plants and animals. Other physical techniques have per-

mitted biologists to investigate the morphology of complex molecules such as hemoglobin, the gas-carrying protein of blood, and deoxyribonucleic acid (DNA), of which most genes are composed. Thus, morphology encompasses the study of biological structures over a tremendous range of sizes, from the macroscopic to the molecular.

A thorough knowledge of structure (morphology) is of fundamental importance to the physician, to the veterinarian, and to the plant pathologist, all of whom are concerned with the kinds and causes of the structural changes that result from specific diseases.

HISTORICAL BACKGROUND

Significant events. Evidence that prehistoric man appreciated the form and structure of his contemporary animals has survived in the form of paintings on the walls of caves in France, Spain, and elsewhere. During the early civilizations of China, Egypt, and the Near East, as man learned to domesticate certain animals and to cultivate many fruits and grains, he also acquired knowledge about the structures of various plants and animals.

Aristotle was interested in biological form and structure, and his *Historia Animalium* contains excellent descriptions, clearly recognizable in extant species, of the animals of Greece and Asia Minor. He was also interested in developmental morphology and studied the development of chicks before hatching and the breeding methods of sharks and bees. The Greek physician Galen was among the first to dissect animals and to make careful records of his observations of internal structures. His descriptions of the human body, although they remained the unquestioned authority for more than 1,000 years, contained some remarkable errors, for they were based on dissections of pigs and monkeys rather than of humans.

Emergence of modern morphology

Although it is difficult to pinpoint the emergence of modern morphology as a science, one of the early landmarks was the publication in 1543 of *De Humani Corporis Fabrica* by the Belgian Andreas Vesalius, a professor at the University of Padua, in Italy, whose careful dissections of human bodies and accurate drawings of his observations revealed many of the inaccuracies in Galen's earlier descriptions of the human body. Leonardo da Vinci was a morphologist as well as a painter, engineer, and inventor, even though many of his original observations in biology were discovered only after his notebooks were deciphered.

In 1661 an Italian physiologist, Marcello Malpighi, the founder of microscopic anatomy, demonstrated the presence of the small blood vessels called capillaries, which connect arteries and veins. The existence of capillaries had been postulated 30 years earlier by the English physician William Harvey, whose classic experiments on the direction of blood flow in arteries and veins indicated that minute connections must exist between them. Between 1668 and 1680, the Dutch microscopist Antonie van Leeuwenhoek used the recently invented microscope to describe red blood cells, human sperm cells, bacteria, protozoans, and various other structures.

Cellular components—the nucleus and nucleolus of plant cells and the chromosomes within the nucleus—and the complex sequence of nuclear events (mitosis) that occur during cell division were described by various scientists throughout the 19th century. *Organographie der Pflanzen* (1898–1901, "Organography of Plants"), the great work of a German botanist, Karl von Goebel, who was associated with morphology in all its aspects, remains a classic in the field. The Scot John Hunter and the Frenchman Georges Cuvier were early 19th-century pioneers in the study of similar structures in different animals—*i.e.*, comparative morphology. Cuvier in particular was among the first to study the structures of both fossils and living organisms and is credited with founding the science of paleontology. A British biologist, Sir Richard Owen, developed two concepts of basic importance in comparative morphology—homology, which refers to intrinsic structural similarity, and analogy, which refers to superficial functional similarity.

Prospect: direction and trends. One of the major thrusts in contemporary morphology has been the elucidation of the molecular basis of cellular structure. Techniques such as electron microscopy have revealed the complex details of cell structure, provided a basis for relating structural details to the particular functions of the cell, and shown that certain cellular components occur in a variety of tissues. Studies of the smallest components of cells have clarified the structural basis not only for the contraction of muscle cells but also for the motility of the tail of the sperm cell and the hairlike projections (cilia and flagella) found on protozoans and other cells. Studies involving the structural details of plant cells, although begun somewhat later than those concerned with animal cells, have revealed fascinating facts about such important structures as the chloroplasts, which contain chlorophyll that functions in photosynthesis. Attention has also been focussed on the plant tissues composed of cells that retain their power to divide (meristems), particularly at the tips of stems, and their relationship with the new parts to which they give rise. The structural details of bacteria and blue-green algae, which are similar to each other in many respects but markedly different from both higher plants and animals, have been studied in an attempt to determine their origin.

Morphology continues to be of importance in taxonomy because morphological features characteristic of a particular species are used to identify it. As biologists have begun to devote more attention to ecology, the identification of plant and animal species present in an area and perhaps changing in numbers in response to environmental changes has become increasingly significant.

THE SCOPE OF CONTEMPORARY MORPHOLOGY

Fundamental concepts. *Homology and analogy.* Homologous structures develop from similar embryonic substances and thus have similar basic structural and developmental patterns, reflecting common genetic endowments and evolutionary relationships. In marked contrast, analogous structures are superficially similar and serve similar functions but have quite different structural and developmental patterns. The arm of a man, the wing of a bird, and the pectoral fins of a whale are homologous structures in that all have similar patterns of bones, muscles, nerves, blood vessels, and similar embryonic origins; each, however, has a different function. The wings of birds and those of butterflies, in contrast, are analogous structures—*i.e.*, both allow flight but have no developmental processes in common.

Application to molecular structure

The terms homology and analogy are also applied to the molecular structures of cellular constituents. Because the hemoglobin molecules from different vertebrate species contain remarkably similar sequences of amino acids, they may be termed homologous molecules. In contrast, hemoglobin and hemocyanin, the latter of which is present in crab blood, are termed analogous molecules because they have a similar function (oxygen transport) but differ considerably in molecular structure. Corresponding similarities occur in the structures of other proteins from different species—*e.g.*, cytochrome c and other enzymes (biological catalysts) such as the lactic dehydrogenases in birds and mammals.

Body plan and symmetry. The bodies of most animals and plants are organized according to one of three types of symmetry: spherical, radial, or bilateral. A spherically symmetrical body is similar throughout and can be cut in any plane through the centre to yield two equal halves. A few of the simplest plants and animals are spherically symmetrical—*e.g.*, protozoans such as radiolaria and heliozoa. Radially symmetrical bodies, such as those of starfishes and mushrooms, have a distinguishable top and bottom and usually have a cylindrical shape, with the body parts radiating from the central axis. A starfish can be cut into two equal halves by any plane that includes the line, or axis, running through its centre from top to bottom. The anterior, or oral, end usually contains the mouth; a posterior, or aboral, end may have an anus. In the bilaterally symmetrical body of higher animals including man, only a cut from head to foot exactly in the centre divides the body into equivalent halves. An anterior, or head, end and a posterior, or tail, end can be

distinguished; and the dorsal, or back, side can be distinguished from the ventral, or belly, side. But because some internal organs of man are not symmetrical (*e.g.*, the heart), even the right and left halves of the human body are not exactly equivalent. A few organisms—amoebas, slime molds, and certain sponges—with an irregular form, or one that changes as the organism moves, have no plane of symmetry.

Morphological basis of classification. The features that distinguish closely related species of plants and animals are usually superficial differences such as colour, size, and proportion. In contrast, the major divisions, or phyla, of the plant and animal kingdoms are distinguished by characteristics that, although usually not unique to a single division or phylum, occur in unique combinations in each.

One morphological feature useful in classifying animals and in determining their evolutionary relationships is the presence or absence of cellular differentiation—*i.e.*, animals may be either single celled or composed of many kinds of cells specialized to perform particular functions. Some multicellular animals have only two embryonic cell, or germ, layers: an outer ectoderm and an inner endoderm, which lines the digestive tract. Other animals have these, in addition to a mesoderm, which lies between the ectoderm and endoderm. Animals may have one of two types of body cavity. The bodies of the Coelenterata (invertebrates such as the jellyfish) and other primitive many-celled animals consist of a double-walled sac surrounding a single cavity with a mouth. Higher animals have two cavities, and their bodies are constructed on a so-called tube-within-a-tube plan. An inner tube, or digestive tract, is lined with endoderm and opens at each end to form the mouth and the anus. An outer tube, or body wall, is covered with ectoderm. Between the two tubes a second cavity, or coelom, lies within the mesoderm and is lined by it. Another major distinguishing morphological feature of animal phyla is the presence or absence of segmentation. The members of several phyla have bodies characterized by the presence of a row of segments, or body units, of the same fundamental structure. Segmented animals include the vertebrates, the annelids (invertebrates such as the earthworm), and the arthropods (invertebrates such as insects); in some segmented animals such as man and most vertebrates, however, the segmental character of the body is obscured. An evolutionary tendency in many animal phyla has been the progressive differentiation of the anterior end to form a head with conspicuous sense organs and an accumulation of nervous tissues, a brain; the tendency is termed cephalization. Some morphological structures are found only in one phylum; for example, only the Coelenterata have stinging cells (nematocysts); the Echinodermata (invertebrates such as starfishes) have a peculiar water vascular system, and only the Chordates (*e.g.*, reptiles, birds) have a dorsally located, hollow nerve cord.

Segmentation and cephalization

Like animals, plants may be either single celled or composed of many kinds of specialized cells. The bodies of most of the lower plants, such as algae and fungi, are comprised of the least differentiated and least specialized type of plant cells, parenchyma cells. The embryonic tissues of higher plants, unlike those of animals, remain extremely active throughout the life of the plant. In addition, the different types of cells characteristic of the body of higher plants arise from meristems, specific regions in the plant body where cells divide and enlarge. In all but the simplest forms, the plant body is composed of various types of cells associated in more or less definite ways to form systems of units called tissue systems—*e.g.*, the vascular system consisting of conductive tissues. The arrangement of the components of the vascular system is a distinguishing morphological feature of various plant groups. The character and relative extent of the two phases in the life history of a plant—the sexual phase, or gametophyte, and the sporophyte—vary considerably among the plant groups and are useful in distinguishing them.

Anatomy. The best known aspect of morphology, usually called anatomy, is the study of gross structure, or form, of organs and organisms. It should not be inferred however, that even the human body, which has been extensively studied, has been so completely explored that nothing remains to be discovered. It was found only in 1965, for example, that the nerve to the pineal gland, which lies on the upper surface of the brain of mammals, is a branch from the sympathetic nerves; the sympathetic nerves receive nerve impulses from a small branch of the nerves that transmit impulses from the eye to the brain (optic nerves). Thus the pineal gland responds by a very indirect route to quantitative changes in the environmental lighting and secretes appropriate amounts of the substance it forms, the hormone melatonin.

Animal structure. Detailed comparisons of the morphological features of different animals, termed comparative anatomy, provide strong arguments for the evolutionary relationships among different species. In the course of evolution, animals and plants tend to undergo adaptive morphological changes that enable them to survive under certain environmental conditions. As a result, animals only remotely related evolutionarily may come to resemble each other superficially because of common adaptations to similar environments, a phenomenon known as convergent evolution. Structural similarities—streamlined shape, dorsal fins, tail fins, and flipper-like forelimbs and hindlimbs—for example, have evolved in such varied animal groups as the dolphins and porpoises, both of which are mammals; the extinct ichthyosaurs, which were reptiles; and both the bony and cartilaginous fishes. In a like manner, the mole, an insectivore, and the gopher, a rodent, have both evolved shovellike forelimbs, an adaptation for digging.

An opposite phenomenon, divergent evolution, occurs when animals originally closely related adapt to different environments and come to be superficially quite different. Although sea lions and seals, for example, are carnivores and thus closely related to bears, cats, and dogs, their adaptations to an aquatic existence have resulted in morphological characteristics distinct from those of the terrestrial carnivores. In the course of mammalian evolution, many features have changed to permit specific animal groups to adapt to particular environments—*e.g.*, the number and shape of the teeth, the length and number of bones in the limbs, the number and attachment sites of muscles, the thickness and colour of the hair or fur, and the length and shape of the tail.

Relationship of morphological phenomena to evolution

Careful study of adaptive morphological aspects has permitted inferences about the course of the evolutionary history of various animals and of their successive adaptations to changing environments. The present-day Australian tree-climbing kangaroos, for example, are the descendents of a ground-dwelling marsupial, from whom evolved forms that began to live in trees and eventually developed limbs adapted to tree climbing. But the events may have occurred in the reverse sequence; that is, specialized limbs may have evolved before the animal adopted an arboreal mode of life. In any event, some of the tree-dwelling kangaroos subsequently left the trees, became readapted to life on the ground (*i.e.*, their hindlegs became adapted for leaping), and then went back to the trees but with legs so highly specialized for leaping as to be useless in grasping a tree trunk; consequently, present-day tree kangaroos climb by bracing their feet against a tree trunk, as do bears. Careful comparisons of the feet of the many kinds of living Australian marsupials reveal the stages in this complicated process of adaptation and re-adaptation.

Changes in genes (mutations) constantly occur and may cause a decrease in size and function of an organ; on the other hand, a change in the environment or in the mode of life of a species may make an organ unnecessary for survival. As a result, many plants and animals contain organs or parts of organs that are useless, degenerate, undersized, or lacking some essential part when compared to homologous structures in related organisms. The human body, for instance, has more than 100 such organs —*e.g.*, the appendix, the fused tail vertebrae (coccyx), the wisdom teeth, the muscles that wiggle the ears, and the hair on the body.

Plant structure. The parts of a seed plant include roots, stems, leaves, and reproductive organs in the flowers. The evolution of specialized conducting tissues called xylem and phloem has enabled seed plants to survive on land and to attain large sizes. Roots anchor the plant; enable it to maintain an upright position; and absorb water, minerals, and other nutrients from the soil. The roots of plants such as carrots, beets, and yams serve as sites for food storage. The stem links the roots with the leaves, where photosynthesis occurs, and its xylem and phloem are continuous with those of root and leaf. The stem supports leaves, flowers, and fruits. Each year, the stems of woody plants add a layer of xylem and phloem, the annual ring, the width of which varies with climatic conditions. A leaf consists of a petiole (stalk), by which it is attached to the stem, and a blade, typically broad and flat, that contains bundles, or veins, of xylem and phloem on the undersurface. The flower contains pollen-producing anthers and egg-producing ovules. After fertilization the base of the flower, or ovary, enlarges and forms the fruit, which is a mature ovary containing seeds, or mature ovules. The bodies of ferns and mosses also are composed of roots, stems, and leaves, but those of lower plants such as mushrooms and kelps are much more simple and lack true roots, stems, and leaves.

Histology. A major trend in the evolution of both plants and animals has resulted in the specialization of cells and a division of labour among them. The cells comprising a tree or a man are quite different; each is specialized to carry out certain functions. Although specialization may permit a cell to function efficiently, it also increases the interdependence of body parts; an injury to or the destruction of one part, therefore, may result in death of the whole organism. The study of the structure and the arrangement of tissues, which are defined as groups or layers of similarly specialized cells that together perform certain special functions, is termed histology. Each kind of tissue is composed of cells with characteristic features such as size, shape, and relationship to adjacent cells and may also contain noncellular material—connective tissue fibres or a bony material.

Types of tissues and their roles

Animal tissues. Morphologists usually separate animal tissues into six groups: epithelial, connective, muscular, blood, nervous, and reproductive tissues. The cells composing epithelial tissues form a continuous layer or sheet that either covers the surface of the body or lines some cavity within the body, thus protecting the underlying cells from mechanical and chemical injury or from invasion by micro-organisms. Epithelial tissues absorb nutrients and water, secrete a wide variety of substances, and may play a role in the reception of sensory stimuli. The connective tissues—bone, cartilage, ligaments, and fibrous connective tissue—support and hold together the other cells of the body. The cells of the connective tissues secrete large quantities of nonliving material (matrix), the characteristics of which largely determine the nature and the function of the specific types of connective tissue; the matrix secreted by fibrous connective tissue cells, for example, is a thick matted network of microscopic fibres surrounding the connective tissue cells. Connective tissue holds skin to muscle, keeps glands in position, makes up the tough outer walls of the blood vessels, and forms a sheath around nerve fibres and muscle cells. Tendons are flexible, cable-like cords of specialized fibrous connective tissue that join muscles to each other or muscle to bone. Ligaments are somewhat elastic cords of specialized fibrous connective tissue that join one bone to another.

Muscular tissues are composed of elongated, cylindrical, or spindle-shaped cells, each of which contains many small fibres called myofibrils. Muscle cells perform mechanical work by contracting—that is, by becoming shorter and thicker. The three types of vertebrate muscles include cardiac muscle, which is found only in the walls of the heart; smooth muscles, which are found in the walls of the digestive tract and in other internal organs; and skeletal muscles, which make up the bulk of the muscle masses attached to the bones of the body. Skeletal and cardiac muscles have alternating light and dark stripes the relative sizes of which change during the contraction process. Evidence from electron microscopy indicates that two types of filaments occur in muscle; during contraction, one type of filament slides past the other.

Nerve tissue is made of cells, called neurons, which are specialized to conduct nerve impulses. Two or more thin hairlike fibres, called axons and dendrites, extend from the enlarged cell body containing the nucleus. The neurons extending from the spinal cord to the end of an appendage (*e.g.*, arm, leg) may be a metre (about three feet) or more long in man and up to several metres long in an elephant or a whale.

Egg cells in the female and sperm cells in the male are reproductive tissues adapted for the production of offspring. The egg cell is modified by the accumulation of considerable amounts of yolk and other food reserves. The highly specialized spermatozoon contains a tail, the beating of which propels it to the egg. Blood is composed of red cells, which are specialized for the transport of oxygen and carbon dioxide, and white cells, which engulf bacteria and produce antibodies (proteins formed in response to foreign substances called antigens). Blood also contains platelets, small fragments of cells from the bone marrow that play a key role in initiating the clotting of blood.

Types of plant tissues

Plant tissues. The cells of higher plants may be differentiated into meristematic, protective, fundamental, and conductive tissues. Meristematic tissues, which are composed of small, thin-walled cells with few or no vacuoles (cavities), differentiate into the other types of plant tissue and are found in the rapidly growing parts of the plant—*e.g.*, at the tips of roots and stems. Protective tissues are composed of thick-walled cells that protect the underlying thin-walled cells from mechanical abrasion and dehydration; examples of protective tissues include the epidermis of leaves and the cork layers of stems and roots. The fundamental tissues comprising the body of a plant include the soft parts of the leaf, the components of the pith and the cortex of stems, the roots, and the soft parts of flowers and fruits. These tissues function in the production and storage of food. Two types of conductive tissues occur in higher plants: xylem conducts water and dissolved salts, and phloem conducts dissolved organic materials such as sugars. Both types are composed of elongated cells that fuse end to end with other cells to form the sieve tubes through which substances are transported in phloem and xylem vessels.

Cytology. The living material of most organisms is organized into discrete units termed cells; the study of their features is known as cytology. The cellular contents, when viewed through a microscope at low magnification, usually appear to consist of granules or fibrils of dense material, droplets of fatty substances, and fluid-filled vacuoles suspended in a clear, continuous, semifluid substance called cytoplasm. The remarkable structural complexity of the cell is more fully revealed at the higher magnifications attainable with the electron microscope; structural details of various cellular components, or organelles, as revealed by the technique known as X-ray diffraction analysis, have provided information concerning the relationships between the structures of the cellular components and of the molecules comprising them. Although most cells have certain features in common, the kinds and amounts of components vary considerably. Cellular components include structures such as mitochondria, chloroplasts, endoplasmic reticulum, Golgi complex, lysosomes, oil droplets, granules, and fibrils. The membrane around the cell is a three-layered structure called a unit membrane; similar membranes surround many cellular components—*e.g.*, the mitochondria.

Nuclear components. A small spherical or oval organelle, the nucleus, is typically found near the centre of a cell. The genes within the nucleus control the development of the various traits of the cell by controlling the synthesis of specific proteins. The nuclear components are separated from those of the cytoplasm by the nuclear membrane. The structure of the nucleolus, a spherical body within the nucleus, is extremely variable in most cells. Although more than one nucleolus may occur in a nucleus, each cell of an animal or plant species has a fixed

number of nucleoli. The nucleoli apparently play a role in the synthesis of the ribonucleic acid (RNA) constituent of the cellular components called ribosomes, which function in protein synthesis. Adjacent to the nucleus in the cells of animals and certain lower plants are two small, cylindrical bodies, the centrioles, which, during cell division, separate, migrate to opposite sides of the cell, and organize a structure called a spindle between them.

Cytoplasmic components. Within the cytoplasm of both plant and animal cells are components called mitochondria, which may be shaped like spheres, rods, or threads. Each mitochondrion is bounded by a double membrane, the outer layer of which forms the smooth outer boundary of the mitochondrion; the inner layer, folded repeatedly into shelflike folds called cristae, contains enzymes that play an essential role in the conversion of the energy of foodstuffs into the energy used for cellular activities. The cells of most plants contain plastids, small bodies involved in the synthesis and storage of foodstuffs. The most important plastids, the chloroplasts, function in trapping the energy of sunlight during photosynthesis. They are disk-shaped structures with a platelike arrangement of tightly stacked membranes.

The cytoplasmic components important in protein synthesis, the ribosomes, are composed of nucleic acid and protein. Clusters of five or more ribosomes, termed polysomes, appear to be the functional unit in protein synthesis.

Lysosomes are membrane-bound structures containing a variety of enzymes that can break down the large molecular constituents of the cell. The membrane surrounding lysosomes presumably prevents the enzymes from digesting the cell contents before the cell dies.

Embryology. The structures and the relationships among the various parts of a mature plant or animal are usually better understood if the successive developmental stages are studied. Thus, morphologists have traditionally been interested in the study of embryos and their developmental patterns—*i.e.*, the science of embryology.

Animal development. Development typically begins in animals with the cleavage, or division, of the fertilized egg (zygote) to form a hollow ball of cells called the blastula; the blastula then develops into a hollow cuplike body of two layers of cells, the gastrula, from which the embryo ultimately is formed. At one time, the techniques available to embryologists enabled them to study only whole embryos at different developmental stages. The science of experimental embryology began during the first half of the 20th century, when microsurgical techniques became available either for the removal and study of certain structures from tiny embryos or for their transplantation to other regions of the embryo. Advances in understanding the mechanism by which biological information is transferred in DNA and the means by which this information results in the production of specific proteins have led to efforts to describe development in biochemical terms. Although hypotheses regarding the reasons for the appearance of a specific enzyme or some other protein at a specific time during development have been formulated and tested, the biochemical basis of morphogenesis itself—that is, the development of particular structures—has not yet been established.

Development of microsurgical techniques

Plant development. The development of the seed plant is basically different from that of an animal. The egg cell of a seed plant is retained within the enlarged lower part, or ovary, of the seed-bearing organ (pistil) of a flower; two sperm nuclei pass through a structure called a pollen tube to reach the egg. One sperm nucleus unites with the egg nucleus to form the zygote from which the new plant will develop; the second sperm nucleus unites with two nuclei, called polar nuclei, to form a body called a triploid endosperm, the cells of which divide to form a nutritive mass within the seed. The zygote undergoes several cell divisions to form the embryo, which is surrounded by the endosperm. The embryo develops one or two seed leaves, or cotyledons, which may become thick and fleshy with stored foodstuffs. The epicotyl, which consists of a growing point enclosed by a pair of folded miniature leaves, develops above the point of attachment of the seed leaves. Below the seed leaves extends the hypocotyl, the tip, or radicle, of which forms the primary root of the embryonic plant. The factors involved in initiating and controlling morphogenesis in plants have been studied by growing cells, tissues, and organs derived from plants. Indeed, an entire carrot plant has been grown from one cell of a mature carrot; this provides striking evidence that the cell from the adult plant contains all of the genetic information needed to produce an entire plant, including roots, stems, and leaves. The technique of growing plants from isolated plant parts has been useful in studies involving the characteristics of embryonic growth, the correlated growth of plant parts, and the nature of differentiation and regeneration (the replacement of lost parts).

METHODS IN MORPHOLOGY

Chemical techniques. The methods of investigating gross structure depend on careful dissection, or cutting apart, of an organism and on accurate descriptions of the parts. The study of the structure of tissues and cells has been extended by the techniques of autoradiography and histochemistry. In the former, a tissue is supplied with a radioactive substance and allowed to utilize it for an appropriate period of time, after which the tissue is prepared and placed in contact with a special photographic emulsion. Silver grains in the emulsion in contact with radioactive substances darken; thus, the location of the dark spots indicates the position at which the radioactive substance was concentrated in the tissue. Histochemistry involves the differential staining of cells (*i.e.*, using dyes that stain specific structural and molecular components) to reflect the chemical differences of the constituents. By choosing appropriate dyes, the histochemist is able, for example, to determine the acidity or alkalinity of the chemical compounds comprising cell components. In addition, dyes that stain specific molecular constituents such as glycogen, DNA, RNA, and protein also are used. The histochemist is able to locate a specific enzyme in a thin slice of tissue, to provide the specific substance with which the enzyme reacts to form a product, and to add a compound that reacts with the product to form an insoluble coloured compound the location of which is relatively easy to determine. In this way, information has been obtained about the specific location of enzymes within the cell.

Dissection and description

Microscopic techniques. Histologists and cytologists utilize microscopic techniques—light microscopy, phase contrast microscopy, interference microscopy, polarization microscopy, fluorescent microscopy, and electron microscopy—to investigate certain aspects of cell structure. Phase contrast microscopy is widely used to study the structure of living cells because, with such apparatus, internal structures can be observed without killing and staining the cell. In addition, motion pictures of dividing cells or moving cells can be made using phase contrast microscopy.

The interference microscope involves passing two separate beams of light through the specimen. With the appropriate instrument, the mass of material per unit area of the specimen can be determined, and contour mapping of small objects is possible.

Crystalline or fibrous elements, both of which are characterized by an orderly or layered molecular structure, are studied with a polarizing microscope; the polarizing microscope has been particularly useful in studying the detailed structure of bone.

In fluorescence microscopy, the images seen are molecules of fluorescent dyes added to cells that attach to specific cellular components. Appropriate filters are required to insure that only the light of longer wavelength contributes to the image. Fluorescent antibodies have been used to locate specific kinds of proteins and other materials in certain cells of a tissue or in certain regions of a cell. The antibodies are prepared by injecting into a rabbit an antigen (*e.g.*, the protein myosin), which stimulates white blood cells called lymphocytes to synthesize antibodies that react specifically with the antigen. After the antibodies are isolated and purified, the fluorescent dye, fluorescein, becomes attached to them by a chemical

Fluorescence microscopy

reaction. If the fluorescent antibodies are spread over a tissue, they attach specifically to the molecules that stimulated their formation (myosin). The fluorescence microscope reveals the sites containing the antigen–antibody complex as bright luminescent areas in a dark background.

In the scanning electron microscope, a moving spot of electrons (negatively charged particles) is used to scan an object and to produce an image similar to that which appears on a television screen. In this manner, photographs with a three-dimensional appearance can be produced. With the transmission electron microscope, a beam of electrons passes through an object, such as a cell, and is focussed on the other side onto a fluorescent screen or a photographic plate. The beam of electrons in the scanning electron microscope is focussed and then scanned across the specimen. The electrons that leave the specimen, which are not necessarily the same electrons that strike it, are then used to control the beam of a cathode-ray picture tube. Scanning electron microscopes allow photographs to be taken not only of large molecules such as DNA but of very small objects—individual atoms of elements such as uranium or thorium.

MORPHOLOGICAL SOCIETIES AND PUBLICATIONS

The major morphological society in the United States is the American Association of Anatomists, which publishes the *Anatomical Record*. The American Society of Zoologists encompasses a broad group of biologists that includes morphologists, physiologists, and ecologists; similarly, the Botanical Society of America, which publishes the *American Journal of Botany*, includes plant biologists with a wide range of interests in both structure and function. Major biological societies with strong interests in morphology also include the Société de Biologie (France), Linnean Society of London, the Zoological Society of Japan, and Scandinavian Biological Society.

Specific societies are concerned with the study of one major group of animals or plants; *e.g.*, the societies of mammalogists, of ornithologists, of entomologists, of ichthyologists, of parasitologists, of cryptogamic botanists, and of mycologists are concerned with the various groups after which they are named. A primary interest of each society is the structure of the organisms and the range of variation in such structure.

Studies concerning detailed structures of cells are published in various journals throughout the world. These include *The Journal of Cell Biology* (U.S.), *The Journal of Cell Science* (U.K.), and *Experimental Cell Research* (Sweden). Other major morphological journals include *American Journal of Anatomy*, *Journal of Morphology* (U.S.); *Canadian Journal of Zoology; Acta Morphologica* (Hungary); *Acta Anatomica* (Switzerland); *Acta Morphologica Neerlando-Scandinavica* (The Netherlands); *Compte rendu hebdomadaire des séances et mémoires*, Société de Biologie (Paris); *Zoological Journal*, Linnean Society (U.K.); *Zoologicheski Zhurnal* (U.S.S.R.); *Zoologischer Anzeiger* (Germany); *Zoologica Africana* (Union of South Africa); and *Zeitschrift für Anatomie und Entwicklungsgeschichte* ("Journal of Anatomy and Developmental Science").

BIBLIOGRAPHY. B.I. BALINSKY, *An Introduction to Embryology*, 3rd ed. (1970), an excellent general college-level text that includes details of organogenesis in each of the vertebrate types, plus a synthesis of the experimental analysis of cellular differentiation; R.D. BARNES, *Invertebrate Zoology*, 2nd ed. (1968), an excellent modern treatment of invertebrate morphology; E.D.P. DE ROBERTIS, W.W. NOWINSKI, and F.A. SAEZ, *Cell Biology*, 5th ed. (1970), an excellent text describing the cytological features of animal and plant cells; KATHERINE ESAU, *Plant Anatomy*, 2nd ed. (1965), a beautifully illustrated text with an excellent treatment of plant anatomy; DON FAWCETT, *The Cell* (1966), a collection of superb electron micrographs of several kinds of cells and cell organelles, with a brief description of each; LIBBIE H. HYMAN, *The Invertebrates*, 6 vol. (1940–67), the definitive treatise of the invertebrates, covering morphology, development, and evolution; A.S. ROMER, *The Vertebrate Body*, 4th ed. (1969), a standard college-level text of comparative anatomy of the vertebrates; C.A. VILLEE, W.F. WALKER, and R.D. BARNES, *General Zoology*, 4th ed. (1973), a standard college-level text of vertebrate and invertebrate morphology and its relation to function.

(C.A.V.)

Morris, William

Poet, designer, craftsman, and pioneer Socialist, William Morris is now regarded as one of the great men of the 19th century, though he turned away from what he called "the dull squalor of civilization" to romance, myth, and epic. Following his contemporary the art critic John Ruskin, Morris defined beauty in art as the result of man's pleasure in his work and asked, "Unless people care about carrying on their business without making the world hideous, how can they care about Art?" To Morris, art included the whole man-made environment.

The Mansell Collection

William Morris, drawing by C.M. Watts.

In his own time William Morris was most widely known as the author of *The Earthly Paradise*, a series of narrative poems based on classical and medieval sources in the manner of Chaucer's *Canterbury Tales*, and for his designs for wallpapers, textiles, and carpets. Since the mid-20th century it is as a designer and craftsman, rather than as poet or politician, that Morris is valued most, though future generations may esteem him more as a social and moral critic, a pioneer of the society of equality.

Education and early career. Morris was born on March 24, 1834, at Walthamstow (in those days an Essex village on the southern edge of Epping Forest), a member of a large and well-to-do family. From his preparatory school, he went at the age of 13 to Marlborough College. A schoolfellow describes him at this time as "a thick-set, strong-looking boy, with a high colour and black curly hair, good-natured and kind, but with a fearful temper." At Marlborough, Morris said that he learned "next to nothing . . . for indeed next to nothing was taught." As in later life, he learned only what he wanted to learn. In 1853 Morris went to Exeter College at Oxford, where he met Edward Jones (later the painter and designer Burne-Jones), who was to become his lifelong friend.

Friendship with Edward Burne-Jones

Both Morris and Jones became deeply affected by the High Church (Anglo-Catholic) movement of the Church of England, and it was assumed that they would become clergymen. Nevertheless, it was the writings of John Ruskin on the social and moral basis of architecture (particularly the chapter "On the Nature of Gothic" in *The Stones of Venice*) that came to Morris "with the force of a revelation." After taking his degree in 1856, he entered the Oxford office of the Gothic Revivalist architect G.E. Street. In the same year he financed the first 12 monthly issues of *The Oxford and Cambridge Magazine*, where many of those poems appeared that, two years later, were reprinted in his remarkable first published work, *The Defence of Guenevere and Other Poems*. Visits with Street and Burne-Jones to Belgium and northern France, where he first saw the 15th-century paintings of Hans Memling and the Van Eyck brothers and the cathedrals of Amiens, Chartres, and Rouen, confirmed Morris in his

love of medieval art. It was at this time that he came under the powerful influence of the Pre-Raphaelite painter and poet Dante Gabriel Rossetti, who persuaded him to give up architecture for painting and enrolled him among the band of friends who were decorating the walls of the Oxford Union with scenes from Arthurian legend based on the *Morte d'Arthur* by the 15th-century English writer Sir Thomas Malory. Only one oil painting by Morris survives: "Queen Guinevere" (Tate Gallery, London). His model was Jane Burden, the beautiful, enigmatic daughter of an Oxford groom. He married her in 1859, but the marriage was to prove a source of unhappiness to both. Morris appears at this time, in the memoirs of the painter Val Prinsep, as "a short square man with spectacles and a vast mop of dark hair." It was observed "how decisive he was: how accurate, without any effort or formality: what an extraordinary power of observation lay at the base of many of his casual or incidental remarks." From 1856 to 1859 Morris shared a studio with Burne-Jones in Red Lion Square, London, for which he designed, according to Rossetti, "some intensely medieval furniture."

Morris, Marshall, Faulkner & Company

After his marriage, Morris commissioned his friend the architect Philip Webb, whom he had originally met in Street's office, to build the famous Red House at Bexleyheath (so called because it was built of red brick when the fashion was for stucco villas). It was during the furnishing and decorating of this house by Morris and his friends that the idea came to them of founding an association of "fine art workmen," which in April 1861 became the firm of Morris, Marshall, Faulkner & Company with premises in Red Lion Square. The other members of the firm were Ford Madox Brown, D.G. Rossetti, Webb, and Burne-Jones. At the International Exhibition of 1862 at South Kensington they exhibited stained glass, furniture, and embroideries. This led to commissions to decorate the new churches then being built by G.F. Bodley, notably St. Martin's-on-the-Hill at Scarborough. The apogee of the firm's decorative work is the magnificent series of stained-glass windows designed during the next decade by Burne-Jones for Jesus College Chapel, Cambridge, the roof being painted by Morris and Webb. The designs for these windows came to Morris uncoloured, and it was he who chose the colours and put in the lead lines. He also designed many other windows himself, for both domestic and ecclesiastical use.

Two daughters, Jenny and May, were born in 1861 and 1862, and altogether the five years spent at Red House were the happiest of Morris' life. After a serious attack of rheumatic fever, brought on by overwork, he moved in 1865 to Bloomsbury in London. The greater part of his new house was given over to the firm's workshops—an arrangement that, combined with her husband's boisterous manners and Rossetti's infatuation with her, reduced Jane to a state of neurotic invalidism. Morris' first wallpaper designs, "Trellis," "Daisy," and "Fruit," or "Pomegranate," belong to 1862–64; he did not arrive at his mature style until ten years later, with the "Jasmine" and "Marigold" papers.

Success as a poet

Iceland and Socialism. As a poet, he first achieved fame and success with the romantic narrative *The Life and Death of Jason* (1867). In the 20th century, *Jason*, with its lax, easily flowing couplets, appears diffuse to the point of tenuity. Just as in the seemingly endless stories of *The Earthly Paradise* (1868–70), all painful emotion is carefully avoided or smothered in prettiness. The best parts of *The Earthly Paradise* are the introductory poems on the months, in which Morris reveals his personal unhappiness. A sterner spirit informs his principal poetic achievement, the epic *Story of Sigurd the Volsung and the Fall of the Niblungs* (1876), written after a prolonged study of the sagas (medieval prose narratives) read by Morris in the original Old Norse. The exquisitely illuminated *Book of Verse*, telling once more of hopeless love and dedicated to Georgiana Burne-Jones, belongs to 1870.

In 1871 Morris and Rossetti took the Elizabethan manor house of Kelmscott in Oxfordshire. In the same year Morris paid his first visit to Iceland, and the journal he kept of his travels contains some of his most vigorous descriptive writing. He returned to Iceland in 1873. The joint tenancy of Kelmscott, however, was never a success, and, after the final breakdown of his health in 1874, Rossetti left the house for good, to Morris' great relief. At the same time, the firm was reorganized under his sole proprietorship as Morris & Company. In 1875 Morris began his revolutionary experiments with vegetable dyes, which, after the removal in 1881 of the firm to larger premises at Merton Abbey in Surrey, resulted in their finest printed and woven fabrics, carpets, and tapestries.

In 1877 Morris gave his first public lecture, "The Decorative Arts" (later called "The Lesser Arts"), and his first collection of lectures, *Hopes and Fears for Art*, appeared in 1882. In 1877 he also founded the Society for the Protection of Ancient Buildings in an attempt to combat the drastic methods of restoration then being carried out on the cathedrals and parish churches of Great Britain.

The Morris family moved into Kelmscott House (named after their country house in Oxfordshire), at Hammersmith, in 1878. Five years later Morris joined Henry Mayers Hyndman's Democratic (later Social Democratic) Federation and began his tireless tours of industrial areas to spread the gospel of Socialism. But he was considerately treated by the authorities, even when leading a banned demonstration to London's Trafalgar Square on "Bloody Sunday" (November 13, 1887), when the police, supported by troops, cleared the square of demonstrators. On this occasion he marched with the playwright George Bernard Shaw at his side. But by this time Morris had quarrelled with the autocratic Hyndman Federation and formed the Socialist League, with its own publication, *The Commonweal*, in which his two finest romances, *A Dream of John Ball* (1886–87) and *News from Nowhere* (1890), an idyllic vision of a Socialist rural utopia, appeared. Subsequently, he founded the Hammersmith Socialist Society, which held weekly lectures in the coach house next door to Kelmscott House as well as open-air meetings in different parts of London.

The Kelmscott Press

The Kelmscott Press was started in 1891, with the printer and type designer Emery Walker as typographical adviser, and between that year and 1898 produced 53 titles in 66 volumes. Morris designed three type styles for his press: Golden type, modelled on that of Nicolas Jenson, the 15th-century French printer; Troy type, a gothic font on the model of the early German printers of the 15th century; and Chaucer type, a smaller variant of Troy, in which *The Works of Geoffrey Chaucer* was printed during the last years of his life. One of the greatest examples of the art of the printed book, *Chaucer* is the most highly decorated of the Kelmscott publications. Most of the Kelmscott books were plain and simple, for Morris observed that 15th-century books were "always beautiful by force of the mere typography."

A sea voyage to Norway in the summer of 1896 failed to revive Morris' flagging energies, and he died at Hammersmith on October 3, 1896, worn out by the multiplicity of his activities. He was buried at Kelmscott beneath a simple gravestone designed by Philip Webb.

BIBLIOGRAPHY. J.W. MACKAIL, *The Life of William Morris*, 2 vol. (1899; World's Classics edition, with introduction by S.C. COCKERELL, 1950, reprinted 1968), the standard biography; AYMER VALLANCE, *William Morris: His Art, His Writings and His Public Life* (1897); E.P. THOMPSON, *William Morris: Romantic to Revolutionary* (1955), the definitive political study; PHILIP HENDERSON, *William Morris: His Life, Work and Friends* (1967); and (ed.), *The Letters of William Morris to His Family and Friends* (1950); PAUL R. THOMPSON, *The Work of William Morris* (1967), a valuable introduction; RAYMOND WATKINSON, *William Morris As a Designer* (1967); MAY MORRIS, *The Collected Works of William Morris*, 24 vol. (1910–15); *William Morris, Artist, Writer, Socialist*, 2 vol. (1936, reprinted 1966); H.H. SPARLING, *The Kelmscott Press and William Morris, Master-Craftsman* (1924); ASA BRIGGS (ed.), *William Morris, Selected Writings and Designs* (1963); NIKOLAUS PEVSNER, *Pioneers of Modern Design from William Morris to Walter Gropius*, 2nd ed. (1949; rev. ed. in paperback, 1964). See also *The Journal* and other publications of the William Morris Society (1955–).

(P.P.H.)

Morse, Samuel F.B.

Samuel Morse was one of the most respected American painters of his day and also famous as the inventor of the electric telegraph. Born on April 27, 1791, in Charlestown, Massachusetts, Morse was the son of Jedidiah Morse, a distinguished geographer and Calvinist clergyman. From Phillips Academy, Andover, where he had been an unsteady and eccentric student, his parents sent him to Yale College. Although an indifferent scholar, his interest was aroused by lectures on the then-little-understood subject of electricity. To the distress of his austere parents, he also enjoyed painting miniature portraits.

By courtesy of the Addison Gallery of American Art, Phillips Academy, Andover, Mass.

Morse, self-portrait in oils, *c.* 1814.

After graduating from Yale in 1810, Morse became a clerk for a Boston book publisher. But painting continued to be his main interest, and in 1811 his parents helped him to go to England to study painting. During the war of 1812, between Great Britain and the United States, Morse reacted to the English contempt for Americans by becoming passionately pro-American. Yet, like most Americans of his time, he accepted English artistic standards, including the "historical" style of painting—the romantic portrayal of legends and historical events with personalities gracing the foreground in grand poses and brilliant colours.

But when, on his return home in 1815, Morse found that Americans did not appreciate his historical canvases, he reluctantly took up portraiture again to earn a living. He began as an itinerant painter in New England, New York, and South Carolina; after 1825, on settling in New York City, he painted some of the finest portraits ever done by an American artist. He combined technical competence and a bold rendering of his subjects' character with a touch of the Romanticism he had imbibed in England.

Work as a portrait artist

Although often poor during those early years, Morse was sociable and at home with intellectuals, the wealthy, the religiously orthodox, and the politically conservative. In addition, he possessed the gift of friendship. Among his friends in his middle years were the French hero of the American Revolution, the Marquis de Lafayette, whose attempts to promote liberal reform in Europe Morse ardently endorsed, and the novelist James Fenimore Cooper. Morse and Cooper shared several traits: both were ardent American republicans, though both had aristocratic social tastes, and both suffered from American preference for European art.

Morse also had the gift of leadership. As part of a campaign against the licentiousness of the theatre, he helped launch, in 1827, the New York *Journal of Commerce*, which refused theatre advertisements. He also was a founder of the National Academy of Design, organized to increase American respect for painters, and was its first president from 1826 to 1845.

In 1832, while returning by ship from studying art in Europe, Morse conceived the idea of an electric telegraph as the result of hearing a conversation about the newly discovered electromagnet. Although the idea of an electric telegraph had been put forward before 1800, Morse believed that his was the first proposal. He probably made his first working model by 1835. Meanwhile, he was still devoting most of his time to painting, teaching art at the University of the City of New York (later New York University), and to politics. But, by 1837, he turned his full attention to the new invention. A colleague at the university showed him a detailed description of an alternative model proposed in 1831, and a friend offered to provide materials and labour to build models in his family's ironworks. These two became partners in Morse's telegraph rights. By 1838 he had developed the system of dots and dashes that became known throughout the world as the Morse Code. In 1838, while unsuccessfully attempting to interest Congress in building a telegraph line, he acquired a congressman as an additional partner. After failing to organize the construction of a Morse line in Europe, Morse was finally able to obtain, without his partners' cooperation, financial support from Congress for the first telegraph line in the United States, from Baltimore to Washington. The next year he sent the first message, "What hath God wrought!"

Morse was immediately involved in legal claims by his partners and by rival inventors. He fought vigorously in this dispute as he had in controversies about art, religion, politics, and even over the new invention of the daguerreotype, an early photographic process, of which he had been one of the first practitioners in America. A series of legal battles culminating in a United States Supreme Court decision established his patent rights in 1854. As telegraph lines lengthened on both sides of the Atlantic, his wealth and fame increased. By 1847, Morse bought Locust Grove, an estate overlooking the Hudson River near Poughkeepsie, New York, where, early in the 1850s, he built an Italian-villa-style mansion. Morse spent his summers there with his large family of children and grandchildren, returning each winter season to his brownstone home in New York City.

In his old age, a patriarch with a flowing beard, Morse became a philanthropist. He gave generously to Vassar College, of which he was a founder and trustee; to his alma mater, Yale College; and to churches, theological seminaries, Bible societies, mission societies, temperance societies, as well as to poor artists.

Assessment

Even during Morse's own lifetime, the world was much changed by the telegraph. After his death, his fame as the inventor of the telegraph was obscured by the invention of the telephone, radio, and television, while his reputation as an artist has grown. At one time he did not wish to be remembered as a portrait painter, but his powerful and sensitive portraits, among them those of Lafayette, the American writer William Cullen Bryant, and other prominent men, have been represented in special exhibitions at museums in many parts of the country. The number of Morse telegraphic operators has decreased sharply, but his memory is perpetuated by the Morse Telegraph Club, which was organized in 1942 as an American and Canadian association of former telegraphers. His 1837 telegraph instrument is preserved by the United States National Museum of History and Technology in Washington, D.C., while his estate, Locust Grove, now designated as a historic landmark, has been willed to the federal government as a Morse memorial. Morse died on April 2, 1872, in New York City.

BIBLIOGRAPHY. CARLETON MABEE, *The American Leonardo: A Life of Samuel F.B. Morse* (1943, reprinted 1969); OLIVER W. LARKIN, *Samuel F.B. Morse and American Democratic Art* (1954), a brief, useful work, without documentation; MARSHALL B. DAVIDSON, "What Samuel Wrought," *American Heritage*, 12:12–31 (1961), a good reappraisal of Morse; *Samuel F.B. Morse, American Painter* (1932); and *Paintings by Samuel F.B. Morse* (1932), catalogs of special showings of Morse paintings at the Metropolitan Museum of Art, New York, and the Syracuse Museum of Fine Arts, Syracuse, New York.

(C.Ma.)

Mortgages, Law of

As used in modern law, a mortgage is a transfer of a property interest by a debtor (mortgagor) to his creditor (mortgagee) to be held as security for the performance of an obligation. Usually, the obligation is to pay a debt created or evidenced by the debtor's bond or promissory note. The property interest is commonly in land and buildings but may be in tangible goods or intangible assets, such as life insurance policies, corporate stock, government bonds, or other investment securities. Mortgages on land and buildings are known as real estate mortgages; those on movable goods, tangible or intangible, are known as chattel mortgages. So defined, the mortgage differs from the pawn of tangible goods, the pledge of documents representing intangible assets, and the medieval gage (pledge) of land, in that the mortgagee's security interest is not dependent upon his having possession of the land, goods, or documents involved. The mortgage also differs from the long-term contract for sale of land, in which the seller retains title until the price has been paid, and from hire-purchase agreements and conditional sale contracts for sale of movable goods on credit, in that the property is owned by the debtor immediately before the transaction. In the case of the purchase-money mortgage, the debtor's prior ownership is momentary and theoretical. Nevertheless, there are often substantial differences in the relative rights of seller, purchaser, and third parties when the unpaid portion of the purchase price is secured by mortgage rather than retention by the seller of his original ownership.

CIVIL-LAW DEVELOPMENT

Horoi, the inscribed mortgage stones that were removed by Solon when he became the archon of Athens in 594 BC (but soon reappeared), are perhaps the most famous examples from antiquity. Most of them probably represented a type of mortgage known as sale with right of redemption, under which the creditor acquired ownership of the land but left it in the possession of the debtor, who paid interest (12 percent per annum was common) or rent in lieu of interest. If the loan contract specified no due date it would seem that the creditor could take possession on default in payment of interest but the debtor retained a perpetual right of redemption. If a date for payment of the principal was specified it is probable that, on default, the creditor's ownership became absolute. Presumably, the debtor could not transfer his interest or give a second mortgage without the cooperation of the creditor. Some of the later stones may have evidenced the type of mortgage known as *hypotheke*, under which the creditor had only a lien or charge until default, when he could take possession and vest absolute ownership in himself. It is not at all likely that any of these stones would evidence a third type of mortgage known as *antichresis*, under which the creditor took possession of the land and kept the rents and profits in lieu of interest. The detailed effects of these transactions are disputed and may have depended on the terms of particular contracts: some may have permitted the debtor to transfer his interest and give a second, junior, mortgage and may have required the creditor to sell the land on default and account to the debtor for the amount by which the sale price exceeded the debt.

Roman law of mortgage

Roman law produced three forms of mortgage, *fiducia cum creditore*, *pignus*, and *hypotheca*, which developed in that order. *Fiducia cum creditore* involved a transfer of ownership from the debtor to the creditor, effected by a formal ceremony before witnesses or a public officer, accompanied by a *pactum fiduciae* binding the creditor to reconvey on payment. When land was involved the debtor might be allowed to retain possession by way of *precarium* (tenancy at will). The debtor had no property interest in the subject matter and no protection against third parties who dealt with the creditor as owner. *Pignus*, used mainly for movables but sometimes, in the form known as *antichresis*, for land, involved delivery of possession to the creditor and retention of ownership by the debtor. *Hypotheca* gave the creditor only a lien or charge. The debtor retained ownership and, commonly, possession until default and so could give a second or junior mortgage. The terms of these transactions usually empowered the creditor to sell on default and such a power was implied by AD 200. Until forbidden by Constantine, the terms could include the *lex commissoria*, a provision for strict foreclosure giving the creditor absolute ownership on default; after AD 200 it also gave him the right to acquire absolute ownership by a judicial proceeding, which afforded the debtor a year's time for redemption and provided for payment of the amount by which the value of the property exceeded the debt. Sixth-century legislation gave the debtor two years in which to redeem after notice of default and the creditor a clear right to foreclose by sale, accounting to the debtor for any excess of the sale price over the debt. Strict foreclosure by judicial decree was possible if a purchaser could not be found.

Development of German and French law

Medieval German law provided for three types of land gage, all involving possession by the creditor. Ownership could be conveyed to the creditor with a condition allowing for subsequent re-entry or a proviso for reconveyance on payment. Possession could be delivered to the creditor with a conveyance of ownership to become effective upon the condition precedent of default. Possession could be delivered to the creditor, to be retained until payment of the principal, without any transfer of ownership, present or future. As under equivalent English types of gage, the creditor might be bound to apply the rents and profits to reduce the debt or he might be entitled to keep them without accountability. These possessory gages were superseded by the *jüngere Satzung* or execution gage, which resembled the English statute staple. The debtor entered into a recognizance before an official. This was registered in a public office and gave the creditor a lien or charge. The debtor retained ownership and possession until default, when the creditor could secure judicial foreclosure, usually by sale. If the sale proceeds exceeded the debt, the surplus was paid to the debtor. With the reception of Roman law in northern Europe, the *jüngere Satzung* was replaced by the Roman *hypotheca*. This was, in fact, a regressive development because no provision was made for recording. Legislation beginning in Prussia early in the 18th century and culminating in the German Civil Code of 1900 required registration of mortgages in public offices and recognized two types, the *Hypothek*, which entails personal liability of the debtor as well as a lien on the land, and the *Grundschuld* (ground-debt), which is solely a charge on the land. Foreclosure is by judicially ordered sale.

The French development paralleled the German. The possessory gages were superseded by the *obligation*, the equivalent of the *jüngere Satzung*. This in turn was superseded by the *hypothèque*, which had to be acknowledged before a judge or a notary and, after 1771, had to be by notarial deed. The inconvenience caused by the lack of recording was partially overcome by the clearance, a judicial proceeding whereby a purchaser of land could compel all persons claiming to be mortgagees, known or unknown, to present their claims or be barred. The Law of the 11th Brumaire, year VII (as the French Revolution styled November 1, 1798) required every mortgage to describe the land bound by it and to be registered in a public office. The German *Grundschuld*, which is a charge on land but not a personal liability of a debtor, is unknown in France. Chattel mortgages under which the debtor retains possession of the goods are virtually unknown in both countries, except as to ships and aircraft, for which there are registration systems. French law does permit nonpossessory liens on automobiles, farm products, hotel furnishings, shares in the Bank of France, and equipment and stock of manufacturers and oil refiners.

COMMON-LAW DEVELOPMENT

The development of the mortgage in England is somewhat obscured by the papal prohibition of interest on debts, enforced by the secular law until 1545. The Christian lender tended to conceal what was essentially a loan at interest in the guise of some other transaction, commonly a purchase for a lump sum of a beneficial lease of land for a term of years.

From a time before the Norman Conquest of 1066 until

1925, English law permitted a type of gage of land as security for a debt that corresponded in form, although not at all times in effect, to the real-estate mortgage now common in the United States. This was an outright conveyance of the fee simple (perpetual ownership) by the debtor to the creditor accompanied by a proviso that, if the debt was paid by the due date, the debtor might terminate the creditor's ownership by re-entry. After 1500 the proviso usually required the creditor to reconvey the land to the debtor upon timely payment of the debt. Both types of proviso were interpreted as conditions subsequent on the creditor's fee simple; that is, the creditor would obtain the complete legal title only if the debtor did not perform. Before the Conquest the conveyance was made by delivery of the landbook; from the Conquest until 1536 it was usually by "feoffment with livery of seisin," a formal transfer of possession of the land in the presence of witnesses, accompanied by oral words of conveyance that might be accompanied by a written charter. After 1535 the conveyance was usually by delivery of a deed operating under the so-called Statute of Uses of that year, which gave legal titles and the right to possession to those who before were entitled only to the use of the land under an equitable title. Under the gage in fee simple, both the creditor and the debtor had property interests in the land enforceable against third parties. That of the debtor was known as a right of entry. It descended to his heirs at death but could not be transferred *inter vivos* (between living persons) or by will before the 17th century.

By the 12th century another form of gage was also being used. This was an informal delivery of possession of the land to the creditor, accompanied by an understanding that he could retain possession until the debt was paid. The understanding might include a provision that, if the debt was not paid by a stipulated date, the creditor could keep the land forever. In the absence of such a provision and of any provision showing that the creditor's interest was limited to rents and profits, the creditor could bring suit to obtain a decree that if the debt was not paid by a specified date the latter would own the land in fee simple. This form of gage, which left ownership of the land with the debtor until default, proved to be unsatisfactory for the creditor because the courts would not protect his possession against either the debtor or third parties. To remedy this defect, it became common in the 13th century for the debtor to give the creditor a formal lease for a definite term of years with a proviso that, if the debt was not paid by the end of the term, the creditor would have a fee simple. This device gave the creditor an estate for years in the land, which the courts would protect against the debtor and third parties. By the 15th century, however, the device was abandoned because the courts would not enforce the proviso for a fee simple to spring up in the future. This leasehold device was the basis of a type of mortgage prescribed by English statute in 1925.

Vif gage and mortgage

All three types of gage contemplated retention of possession of the land and collection of the rents and profits by the creditor. If the terms of the transaction called for the creditor to credit the rents and profits against the principal of the debt, the transaction was a vif gage (*vivum vadium*). If the creditor was permitted to keep the rents and profits without accountability the transaction was a mortgage (*mortuum vadium*), and in this case the creditor could be suspected of evading the ban on interest. This, coupled with a situation in which unwritten conveyances were possible and there was no system of recording written conveyances in a public office, made it unsafe for the creditor to permit the debtor to occupy the land. Moreover, a lease by the creditor to the debtor, reserving rent that was not to be applied on the principal of the debt, would also have violated the rule forbidding Christians to charge interest.

Irrespective of these prohibitions, it was seen as commonly advantageous to both parties to permit the mortgagor to retain possession of the land until default. This solution had been made feasible for Jewish gagees (those who held the gage, or pledge) in 1194 by an edict establishing public offices for recording conveyances to Jews. As Jews were exempt from the ban on receiving interest, a Jewish gagee could require a Christian gagor in possession to pay interest or rent in lieu of interest. On default the Jewish creditor could secure a court order authorizing him to take possession of the land and sell it after a year and a day or hold it until the debt was paid by application of the rents and profits. These privileges were withdrawn and the Jews expelled from England late in the 13th century, but the convenience of mortgages permitting the mortgagor to retain possession until default had been discovered. Legislation of 1285, amplified in 1353, provided that when a debtor acknowledged a debt by formal recognizance (obligation of record requiring performance of an act) in certain courts and public offices, the creditor would immediately have a security interest in lands then owned by the debtor. In the event of default, the creditor was entitled to take possession of half the lands of the debtor (all in the case of certain mercantile debts) and retain them until the debt was paid by application of the rents and profits. An ordinary creditor holding land under this legislation was called a tenant by elegit; a mercantile creditor was called a tenant by statute merchant or, after 1353, tenant by statute staple. Legislation of 1531 made the statute staple, which bound all lands of the debtor, available to noncommercial creditors.

The common-law mortgage

When a statutory recognizance was used as a substitute for a mortgage, the creditor had, until default, only a lien or charge on the land; the debtor remained the owner, entitled to possession and to convey his interest. On the other hand, the gage by feoffment (the granting of a fee) in fee simple with proviso for reconveyance on payment of the debt, which became the standard form of mortgage in the 15th century and is usually referred to as the common-law mortgage, made the mortgagee the owner, entitled to possession. The mortgagor retained only a right of entry, which was inalienable and gave him no right to possession until he paid the debt. Legislation in 1535 eliminated, for all practical purposes, oral conveyances of land, and a statute of 1545 abolished the ban on interest. These changes made it safe for the mortgagee of land to permit the mortgagor to retain possession until default. The mortgagee could protect himself against third parties by retaining the title deeds, and a provision requiring the mortgagor to pay interest, or rent in lieu of interest, was no longer invalid. By the 17th century most mortgagors of land were permitted, by express proviso or tacit understanding, to retain possession of the mortgaged land until default. In the absence of other agreement, such a mortgagor was not bound to pay rent or to account to the mortgagee for the rents and profits. It remained unsafe for a mortgagee of movable goods to leave them in the possession of the mortgagor until the 19th century, when legislation was enacted providing for the recording of chattel mortgages in public offices and protecting the mortgagee who had made such a recording against purchasers from and creditors of the mortgagor in possession.

The common-law mortgage protected both the mortgagor and the mortgagee against third parties but was unsatisfactory to the mortgagor because his interest in the land was not transferable and was forfeited if full payment of the debt was not made on the due date stated, even if the amount due was only a small fraction of the value of the land. Nontransferability meant that the mortgagor could not give a second or junior mortgage. The harshness of the forfeiture was aggravated in the 15th century as a result of a relaxation in the papal ban on interest that permitted a creditor to charge interest on a debt while it remained unpaid after the due date. To take advantage of this, lenders drew up mortgages with nominal due dates, which were long before payment was actually anticipated. The common-law courts of King's Bench, Common Pleas, and Exchequer would give no relief against such forfeitures, but the Court of Chancery, which could apply equitable principles, began to do so in the 16th century. The mortgagor who had defaulted in payment and so had lost all his rights in the land at common law was then permitted to sue in equity for redemption. Upon payment of the debt, with interest to the time of payment, the mortgagee would be compelled to reconvey the land to the mortgagor. The Court of Chancery did not, however, interfere

with the mortgagee's right to possession pending full redemption. If the mortgage entitled him to possession from its date, he retained that right. If it entitled him to take possession on default, he could still do so, and could keep possession until redemption. The "equity of redemption" created by the Court of Chancery was only a right of the mortgagor to pay the debt and recover the land after the due date of the mortgage.

The Court of Chancery and equity redemption

Although at common law, the interest of the mortgagor was inalienable and could not be transferred, by the 17th century the Court of Chancery had converted his equity of redemption into an equitable estate in land, equivalent to the legal estate in fee simple that he held before giving the mortgage. It was as freely transferable *inter vivos* and devisable by will as like legal estates. Hence the mortgagor could give a second or junior mortgage. The Court of Chancery would not give effect to any provision of a mortgage that would do away with or hinder the equity of redemption. The mortgagor, moreover, could assert his equity of redemption for at least 20 years after default, so the mortgagee, despite the fact that he was the absolute owner at common law, did not have a secure title during this period. He could, however, clear the title at any time after default by a suit in equity for foreclosure of the equity of redemption; that is, for a decree requiring the mortgagor to pay by a date specified in the decree or lose his equity of redemption. Such a decree of "strict foreclosure" gave the mortgagee an indefeasible title on the date specified. A statute of 1881 empowered the court, as an alternative to strict foreclosure, to order foreclosure by sale of the land to satisfy the debt. If the proceeds of the foreclosure sale exceeded the debt and the interest due on it, the excess was paid to the mortgagor. For purposes of redemption and foreclosure, the Court of Chancery treated deposit of the title deeds with a creditor as a mortgage. A suit for foreclosure, however, was likely to be slow and expensive. The net effect of the development of the equity of redemption by the Court of Chancery was that, whereas at common law the mortgagee was the owner of the land and the mortgagor had only a right of entry, in equity the mortgagor was the owner and the mortgagee had only a lien or charge on the land.

Once it was established that, despite the provisions of the mortgage, the mortgagee's title would be encumbered by the equity of redemption for at least 20 years after default unless he secured a judicial decree of foreclosure, efforts were made to put terms in the mortgage that would give the mortgagee a power to override the equity of redemption by sale. The first case involving this device, decided in 1738, was interpreted as holding that a purchaser from the mortgagee was also subject to the equity of redemption. As a result of this decision, use of the common-law mortgage was abandoned by some conveyancers in favour of a substitute known as the deed of trust. The deed of trust is a conveyance by the debtor to a third party trustee who is to allow the debtor to occupy the land until default. If the debt is paid, the trustee is to reconvey the land to the debtor. If it is not paid, the trustee, after the default, is to sell the land and, from the proceeds, to pay the amount due to the creditor and account for the balance to the debtor. Both debtor and creditor are beneficiaries of the trust so each has an interest in the land good against third parties and freely alienable and devisable. The debtor can mortgage his equitable interest by a second or junior deed of trust.

Early in the 19th century it was held in England that the terms of a mortgage could confer on the mortgagee power to sell the land after default, free of the equity of redemption. English and some American conveyancers then returned to the use of the common-law mortgage, but the deed of trust is the usual form of security transaction in land in a number of U.S. jurisdictions, notably California, Colorado, the District of Columbia, Delaware, Illinois, Missouri, Mississippi, North Carolina, Tennessee, Texas, Virginia, and West Virginia. Its effects and enforcement in part tend to be regulated by statute in the jurisdictions where it is commonly used. The debtor is ordinarily entitled to some notice of the trustee's intent to sell and an opportunity to redeem for at least a short time after default.

Modern law of mortgage in England and the United States

English legislation of 1925 permits two types of mortgage of land. The first is a lease from the debtor to the creditor for a term of years, usually 3,000 years, with a provision that the term shall cease on timely payment of the debt. The other is by deed from the debtor to the creditor imposing a "charge by way of legal mortgage." The first type resembles the form of gage used in the 13th century; the second resembles the statute staple of the 14th century. In either case the mortgagor retains the legal fee simple, which is transferable *inter vivos* and devisable by will and may be subjected to a second or junior mortgage. A mortgagor in possession has statutory power to grant leases for terms of up to 50 years, which will bind the mortgagee if an adequate rent is reserved. After default the mortgagee has statutory power to sell the fee simple after three months written notice to the mortgagor. He is accountable to the mortgagor for proceeds of the sale in excess of the amount of the debt. Alternatively, the mortgagee may sue for a decree of foreclosure providing that, if the debt is not paid within six months, the mortgagee will own the fee simple. The court may, in lieu of the strict foreclosure sought, direct foreclosure by sale. In any of these cases the proceedings remove the mortgagor's equity of redemption.

In most of the United States the common-law mortgage, in form a conveyance in fee simple on condition subsequent made by the mortgagor to the mortgagee, is still the usual type of land security. In many, however, it does not have the precise effects that it had in England before 1925. It is usually held that, at law as well as in equity, the mortgagor, at least until default, remains the owner and the mortgagee, despite the form of the transaction, has only a lien or charge on the land. This is called the lien theory of mortgage, which stands in opposition to the older common-law title theory. Where the lien theory has been adopted, the mortgagor is entitled to possession, at least until default, and the mortgagee is not. In some states an express provision of the mortgage giving the mortgagee the right to possession is void, and in most states strict foreclosure by judicial decree vesting indefeasible title in the mortgagee is no longer permitted. It has been replaced by foreclosure through judicially ordered sale, as under the English statute of 1881. A provision in a mortgage empowering the mortgagee to foreclose by sale, without judicial proceedings, is valid in most states, but the mortgagee who sells under such a power is accountable to the mortgagor for any amount by which the proceeds of sale exceed the amount of the debt. Statutes in some states give the mortgagor a period of redemption after a sale under judicial decree or under a power conferred by the mortgage. In states where the mortgagor is entitled to possession of the land until the expiration of the period of redemption he may be able to retain possession for more than a year after the institution of foreclosure proceedings.

As in the case of real estate mortgages, a number of U.S. states apply the lien theory to chattel mortgages. Even in them, however, a chattel mortgagor has fewer of the attributes of ownership than a real estate mortgagor. He may be entitled to possession of the mortgaged goods until default, but statutes tend to make it a criminal offense for a chattel mortgagor to sell the goods or remove them from the county in which the chattel mortgage is recorded without the permission of the mortgagee. Moreover, his equity of redemption is commonly more readily and quickly extinguished, usually by sale under a power conferred by the chattel mortgage or under a summary statutory procedure.

SOCIAL IMPORTANCE OF MORTGAGES

Given the fact that the chief finite resources of any society are men and land, it is of maximum importance that each parcel of land be devoted to the use for which it is most fit. The socialist countries attempt to achieve these ends by government planning and regulation. A society that relies upon private ownership as the incentive for the best development and fullest use of its resources must provide a

system for the ready transferability of land and a means by which men without available cash can finance the purchase and improvement of land. The real estate mortgage is what performs this function.

A typical mortgage problem

By way of simple illustration, suppose the following rural problem: An elderly farmer who is anxious to retire owns a quantity of arable land in need of ditching, agricultural buildings, and other improvements, which would double its productivity. The local bank has money available for investment, while a vigorous young man is eager to buy and improve the land but has no money. If the land is not transferable because of legal restrictions, the owner can neither finance improvements by mortgaging to the bank nor sell to the younger man. At best, he could lease the land for a limited term, in which case the potential buyer in this example would have neither the means nor the motive to make the needed improvements. Given a situation in which, however, the land is freely transferable and a type of mortgage that is fair to both debtor and creditor exists, the buyer can borrow enough from the bank on first mortgage to pay for the improvements and make a down payment to the seller. In addition, he can give the seller a second mortgage to secure payment of the balance of the purchase price in annual installments over an extended period. The down payment here would enable the retiring owner to purchase another home and the installment payments would provide support in his declining years. Neither the bank nor the elderly farmer could safely extend this much credit to the young buyer without effective security interests in the land since there are always risks; he could, for example, be conscripted, lose his health, or die prematurely. Urban home and industrial sites produce similar problems which only mortgages can solve.

PARTICULAR ASPECTS OF MORTGAGES

The mortgage device performs its social function well only if it is fair to both creditor and debtor. The potential lender will demand excessive interest or invest his money in some other way unless the mortgage gives him adequate security, is readily transferable, and can be foreclosed without undue delay and difficulty. A device unfair to the debtor, on the other hand, generates social unrest and discourages persons who, like the young farmer of the illustration, can devote their full talents to productive enterprise only by borrowing. If, for example, the mortgagee can acquire absolute title to the land by strict foreclosure when a small installment payment is delayed in the mail for a day beyond the due date, the mortgagor will lose his land without fault on his part even though it is worth much more than the total amount of the mortgage debt. Similarly, if a mortgagee who has led the mortgagor to believe that he does not want payment on the due date can foreclose without giving the mortgagor notice and a reasonable time in which to redeem, the mortgagor may lose his land although, with notice and time, he could easily have raised the amount due. All modern systems offer mortgagors some protection against such harsh treatment.

Most land is mortgaged; hence transferability of the interests of mortgagors is necessary if land is to be kept in the hands of people able to use it efficiently. The able-bodied farmer of today may be a conscript, an invalid, or dead tomorrow, and the stable and contented homeowner of today may be moved by his employer to a distant city tomorrow. Such transferability may impair the mortgagee's security to some extent. Some U.S. courts do, in fact, uphold provisions in a mortgage giving the mortgagee a right of pre-emption or a right to accelerate the due date if the mortgagor transfers his interest, but the trend of all modern systems to the lien theory of mortgages encourages free transferability. English and French legislation empowering a mortgagor in possession to give leases that override the mortgage permit him to transfer even more than his interest.

The lien theory

The hypothecary or lien theory of mortgage, to which all modern legal systems have moved, is unquestionably superior to the types that involve ownership or possession by the mortgagee. The *hypotheca* of Roman law had, however, a serious weakness in the lack of a recording system. This made it possible for a mortgagor in possession to sell or mortgage to a person who did not know of the outstanding mortgage. The Greek *horoi* and the English deposit of title deeds with the first mortgagee were crude efforts to deal with this problem. Only a compulsory system of recording deeds and mortgages, which gives priority to those recorded first, eliminates the weakness. Such systems have been in effect on the European Continent since the 18th century and in North America and Australasia since the 19th. Compulsory recording was instituted in Yorkshire and Middlesex in the 18th century and an English statute of 1897 empowered the queen in council to extend it to other areas. All areas in England are expected to be subject to it by 1980. Unfortunately, no system of recording chattel mortgages provides adequate protection. The purchaser of a steak in New York, for example, cannot ascertain whether the animal from which it was cut was subject to a chattel mortgage duly recorded in Montana, and so for most items that are readily transported and not easily identified.

If interest rates decline after a mortgage is given, the mortgagor has a motive for terminating the transaction; if they rise, the mortgagee has such an interest. Therefore, in the absence of provision in the mortgage itself, neither party to the standard Anglo-American mortgage may compel termination before the due date; when that arrives the mortgagor may terminate by payment, the mortgagee by foreclosure. Provisions for earlier termination by the mortgagor, often on payment of a penalty to compensate the mortgagee for loss of interest, are common. Termination by mutual agreement is always possible. The rarely used Welsh mortgage, which corresponds to the Graeco-Roman *antichresis* and the early medieval Anglo-German possessory gage without a due date, may be paid off by the mortgagor at any time but may not be foreclosed by the mortgagee. French and Italian law empower a purchaser of mortgaged land who is not personally liable for the mortgage debt to compel the mortgagee to accept payment of the principal although the due date has not arrived. This right of "purge" or clearance has public policy advantages when the mortgage has no due date or has a very remote one, such as 3,000 years, but in the case of a mortgage with a five- or ten-year due date, it may be unfair to the mortgagee to deprive him of the interest that he was promised.

BIBLIOGRAPHY. The leading contemporary treatises on this subject are W.R. FISHER and J.M. LIGHTWOOD, *Law of Mortgage*, 8th ed. by E.L.G. TYLER (1969), for England; and G.E. OSBORNE, *Handbook on the Law of Mortgages*, 2nd ed. (1970), for the United States. The current law of mortgages in civil-law countries is described in M.S. AMOS and F.P. WALTON, *Introduction to French Law*, 3rd ed. by F.H. LAWSON, A.E. ANTON, and L.N. BROWN (1967); and K.W. RYAN, *An Introduction to the Civil Law* (1962). English historical development is discussed in H.D. HAZELTINE, "The Gage of Land in Mediaeval England," *Select Essays in Anglo-American Legal History*, vol. 3 (1909); and R.W. TURNER, *The Equity of Redemption* (1931), with an introduction by H.D. Hazeltine comparing the Roman *fiducia cum creditore* with the English mortgage. Information on the historical development in the Greek and civil-law systems may be found in A.R.W. HARRISON, *The Law of Athens*, vol. 1, *The Family and Property* (1968); W.W. BUCKLAND, *Text-Book of Roman Law from Augustus to Justinian*, 3rd ed. by P. STEIN (1964); J.B. BRISSAUD, *History of French Private Law* (1912; Eng. trans. by R. HOWELL from the 2nd French ed., 1898–1904); and R. HUBNER, *History of Germanic Private Law* (1918; Eng. trans. by F.S. PHILBRICK from the German, 1908).

(W.F.F.)

Mosaic

Mosaic is the art of decorating a surface with designs made up of closely set, usually variously coloured, small pieces of material such as stone, mineral, glass, tile, or shell. Unlike inlay, in which the pieces to be applied are set into a surface, which has been hollowed out to receive the design, mosaic pieces are applied onto a surface, which has been prepared with an adhesive. Mosaic also differs from inlay in the size of its components. Mosaic pieces are anonymous fractions of the design and rarely

have the dimensions of pieces for intarsia work (fitted inlay usually of wood), whose function is often the rendering of a whole portion of a figure or pattern. Once disassembled, a mosaic cannot be reassembled on the basis of the form of its individual pieces.

Technical insight is the key to both the creation and the appreciation of mosaic, and the technical aspects of the art therefore require special emphasis in the present article, both in the sections devoted to materials and methods and in the historical survey of the major periods and centres of mosaic activity. Stylistic, religious, and cultural aspects of mosaic, its role in Western art, and its appearance in other cultures are also dealt with in order to complete the picture of an art form that appears in widely separated places and at different times in history, but only in one place—Byzantium—and at one time—4th to 14th centuries—rose to become the leading pictorial art.

PRINCIPLES OF DESIGN

Interaction of painting and mosaic

Between mosaic and painting, the art with which it has most in common, there has been a reciprocal influence of varying intensity. In colour and style the earliest known Greek figurative mosaics with representational motifs, which date from the end of the 5th century BC, resemble contemporary vase painting, especially in their outline drawing and use of very dark backgrounds. The mosaics of the 4th century tended to copy the style of wall paintings, as is seen in the introduction of a strip of ground below the figures, of shading, and of other manifestations of a preoccupation with pictorial space. In late Hellenistic times there evolved a type of mosaic whose colour gradations and delicate shading techniques suggest an attempt at exact reproduction of qualities typical of the art of painting.

In Roman imperial times, however, an important change occurred when mosaic gradually developed its own aesthetic laws. Still basically a medium used for floors, its new rules of composition were governed by a conception of perspective and choice of viewpoint different from those of wall decoration. Equally important was a simplification of form brought about by the demand for more expeditious production methods. In the same period, the increasing use of more strongly coloured materials also stimulated the growing autonomy of mosaic from painting. As a means of covering walls and vaults, mosaic finally realized its full potentialities for striking and suggestive distance effects, which surpass those of painting.

The general trend towards stylization—that is, reduction to two-dimensionality—in late antique Roman painting (3rd and 4th centuries AD) may have been stimulated by experimentation with colour in mosaic and particularly by the elimination of many middle tones for the sake of greater brilliance. The central role played at that time by mosaic in church decoration, for which it is particularly well suited, encourages the assumption that the roles had shifted and painting had come under its influence. The strong, sinuous outlines and the absence of shading that came to characterize painting during certain periods of Byzantine and western European art of the Middle Ages may have originated in mosaic technique and use of materials. It is notable, however, that from the Renaissance to the 20th century, mosaic was again wholly dependent on painting and its particular forms of illusionism.

In modern mosaic practice, the main tendency is to build on the unique and inimitable qualities of the medium. Although not a few of the works created in the 20th century reveal the influence of painting, figurative or abstract, the art has gone a long way towards self-realization. By and large the modern mosaic maker shares with his medieval predecessor the conviction that there are functions to which the materials of mosaic lend themselves with particular appropriateness.

MATERIALS

In antiquity, mosaics first were made of uncut pebbles of uniform size. The Greeks, who elevated the pebble mosaic to an art of great refinement, also invented the so-called tessera technique. Tesserae (Latin for "cubes" or "dice") are pieces that have been cut to a triangular, square, or other regular shape so that they will fit closely into the grid of cubes that make up the mosaic surface. The invention of tesserae must have been motivated by a desire to obtain densely set mosaic pictures which could match, in pavements, the splendour of contemporary achievements in painting.

Invention of the tessera technique

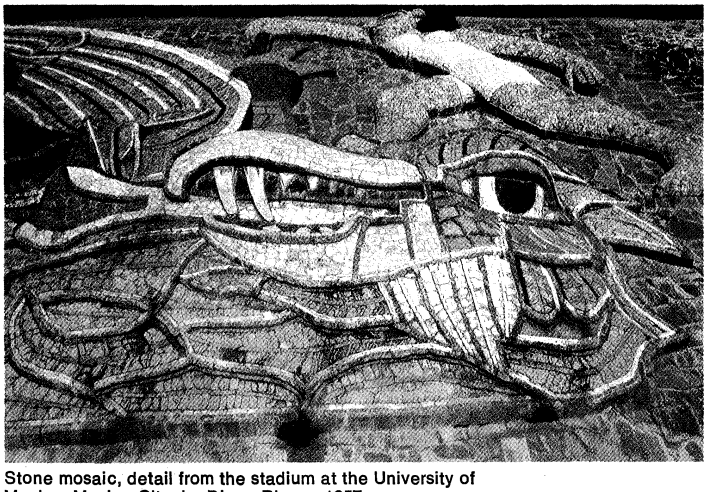
Tesserae vary considerably in size. The finest mosaics of antiquity were made of tesserae cut from glass threads or splinters of stone; ordinary floor decorations consisted of cubes about one centimetre square. Medieval works often display a differentiation in tessera size based on function: areas requiring a wealth of details, faces and hands, for instance, are sometimes set with tesserae smaller than average, while dress and jewelery are occasionally set with very large, single pieces.

As long as mosaic was a technique for the making of floors, the main requisite of its materials, besides their colour, was their resistance to wear.

Stone. Stone, therefore, was long dominant, and throughout antiquity the natural colours of stone provided the basic range of tints at the artist's disposal; they put their mark not only on the earliest Greek works but continued to determine colour schemes far into Roman times. Stone continued to be used in Christian monumental decorations but on a more limited scale and for special effects. In Byzantine mosaics, faces, hands and feet, for example, were set with stone, while cubes of marble, often of coarse crystals, were used to depict woollen garments. Stone was also used for background details (rocks, buildings), probably to bring about particular illusions. Though marble and limestone were ordinarily preferred, in a period when Roman mosaic cultivated a black and white technique, black basalt was widely employed. Marble cubes painted red, probably to substitute for red glass, have been found in many Byzantine mosaics, in 9th-century works at Istanbul, for example.

Because its granular, nonpolished surface is often preferred to the hard brilliance of other materials, stone is also widely used in modern mosaics. At the University of Mexico in Mexico City, for example, the mosaics covering the exterior of the library by Juan O'Gorman (1951–53) and the exterior of the stadium by Diego Rivera (1957) are made with natural stone.

Harrison Forman



Stone mosaic, detail from the stadium at the University of Mexico, Mexico City; by Diego Rivera, 1957.

Glass. Glass, which first appeared among the materials of mosaic in the Hellenistic period (3rd–1st century BC), brought unlimited colour possibilities to the art. In floors, however, it had to be used sparingly because of its brittleness. In floors, glass tesserae were used for the strongest hues of red, green, and blue, while softer tints were rendered with coloured stone. With the development of wall mosaic, glass largely took over the functions of stone, producing tints of unsurpassed intensity and leading to a continuing search for new coloristic effects.

Use of gold tesserae, detail from the Early Christian vault mosaics of Sta. Constanza, Rome, *c.* 337–354 AD.
SCALA, New York

Gold and silver tesserae

With little knowledge of the laws of optics but with immense practical experience, mosaic makers of the Early Christian period gave the art a completely new direction with the exploitation of gold and silver glass tesserae. Like a mirror, the glass from which this kind of tesserae was made had a metal foil applied or, better, encased in it. The metal was gold leaf or, for the "silver," probably tin. These pieces of mirror glass gave golden or white reflections of high intensity and could be used to depict objects of precious metal or to heighten the effect of other colours; but, above all, it was used as a means of rendering the light emanating from God.

Gold tesserae were first used by the Romans, in both floor and vault decoration of late antiquity. Initially, their role was simply to give a golden effect. Gold tesserae, for example, were employed to depict a golden wreath in a floor mosaic at Antioch (*c.* AD 300) and gold vessels in some of the vault mosaics in Sta. Constanza in Rome (*c.* AD 337–354). Later, when this use of gold for imitation purposes had become more refined, some spectacular effects were produced in the depiction of garments. The Good Shepherd in the Mausoleum of Galla Placidia at Ravenna (*c.* AD 450) is dressed in golden robes of densely set gold cubes shaded with stripes of light-yellow tesserae. The female saints in S. Apollinare Nuovo (*c.* AD 550–570) in the same town wear costumes set with green glass cubes among which appear both patterns and large fields of gold tesserae, producing a striking similarity to rich silk brocade. Silver was used in a similar way. Christ scenes in S. Apollinare Nuovo (AD 500–526) employ silver tesserae in the drawn sword of Peter in the betrayal, no doubt an imitation of steel. Silver tesserae are also found in the silver jug and basin in the scene of Pilate washing his hands.

Gold cubes were distributed among the ordinary tesserae to add to the shimmer of light in ornaments and background details. To avoid an uneven gleam in the surface, the mirror effect was often moderated by setting the gold tesserae in reverse, so that the visible part of the cube is the side with the thickest sheet of glass covering the gold leaf. In the now-lost mosaics of the Church of the Dormition in Nicaea, a scholar observed another exquisite effect, which he called dark gold, created by cubes from which some of the gold leaf had been chipped off, for example, in the frontal part of Mary's golden footstool (7th or 8th century AD).

An early instance of the use of gold for depicting light emanating from God is in a representation of Christ-Helios (Christ as the Sun God) in a 3rd-century mausoleum under St. Peter's at Rome. Here, a few gold tesserae are seen in the rays coming from Christ's head. The halo of gold, a feature so common in Christian art that religious pictures without it can hardly be imagined, developed in mosaic art in the 4th century AD. The gold background, signifying divine light, probably originated in Roman mosaic art, but the first preserved instances date from the advanced 4th century. The cupola mosaic of Hagios Georgios at Thessaloníki (*c.* 400), for example, has a background of gold. In Italian mosaics of the 5th century, other types of background, such as a dark-blue ground or a more naturalistic landscape setting, were dominant. Only at the beginning of the 6th century did the gold background become the rule.

In addition to this massive predilection for gold, the Christian East began to use silver to depict the symbolic light emanating from Christ. First, it was used for the entire disc of his halo, later only for the cross arms. The archangels were the only figures besides Christ for whom the silver halo was used. The light of God, appearing as rays from above in scenes of the Annunciation, Nativity, Baptism, and Transfiguration, was also depicted with silver tesserae. Finally, silver and gold were used together in Byzantine representations of the infant Jesus whose golden robes are highlighted with silver cubes (the apse and south vestibule of Hagia Sophia, Istanbul; both 9th century).

Other materials. In Christian mosaics, tesserae of mother-of-pearl or coarse-grained marble cut to round or oblong shapes were used to depict pearl. Though pieces of semiprecious stones were among the mosaic materials of antiquity, their use was rarely dictated by the wish for particular sumptuous effects. Reduced to common tessera size, bits of this strongly coloured material served as part of the general colour scheme of the mosaic pictures. Objects like those of the pre-Columbian American Indian cultures, in which, because of its exquisite materials, such as turquoise and garnet, mosaic attained the status of jewelry have not been found in Western art.

Ceramic tesserae

Among the materials that have played and continued to play a role in the production of mosaic, ceramic is the most versatile. Terra-cotta "threads" were used in Greek mosaics as contours, and tesserae of the same material were frequently used by the Byzantines for the depiction of red objects and garments. Today, glazed or unglazed ceramic is used and is one of the strongest competitors with glass and stone. Ceramic tesserae are cut from tiles or, like much modern glass mosaic material such as pressed glass, come prefabricated. Prefabricated tesserae have the advantage of a very uniform and smooth surface which harmonizes with glass, steel, and other new building materials.

TECHNIQUES

The most commonly used adhesive for mosaics was mortar, the function of which in the 20th century has been largely taken over by modern, tougher kinds of cements or glue. In Roman floors, two to three layers of mortar preceded the setting bed that was to carry a tesserae facing. The first layer rested on a thick foundation of stone that prevented settling of the mortar bed and the formation of cracks. For wall mosaics the preparation was equally painstaking, and in many cases an application of a waterproofing of resin or tar preceded the laying of the mortar. There then followed two layers of coarse, roughened mortar, the stability of which was often improved by large nails that had been driven into the joints of the wall before the work of laying started. A third and final layer was of fine consistency and frequently, like the mortar for floor mosaics, contained powdered marble and binding elements such as pounded brick.

As in fresco painting (technique of using water suspended pigments in a moist plaster surface), the setting bed was applied in patches never larger than were needed for one day's work. In a frescoed surface, the breaks between the different stages of the work can easily be detected; they are harder to discover in mosaic.

Sketches and underpainting

Numerous underpaintings discovered in wall mosaics indicate that sketches, often detailed and with the main colours suggested, were executed on the setting bed to serve as guides for the disposition of the tesserae. Similar procedures are thought to have been part of the technique of floor mosaic. In church mosaics, rough preliminary sketches have been found on layers underneath the setting bed and, in a few instances, even on the brick wall itself. This kind of preparatory sketch, for which there are parallels in wall painting, suggests that the artist was trying out the overall scheme of the decoration before making a more detailed sketch on the setting bed.

Instead of laying the tesserae one by one directly onto the mortar, another method was sometimes used. In Pompeii many of the so-called *emblēmata* (central panels of floors), which were made up of smaller than average tesserae and were often of very high artistic quality, appear to have been preset on trays of stone or terra-cotta which were then embedded in the mortar of the floor. The surrounding mosaic area was then set according to the ordinary, direct method. Although the direct method was used for wall mosaics during the Middle Ages, there are signs in at least one medieval monument of a partial use of the prefabrication—or "indirect"—method: in the cupola mosaic of the church of Hagios Georgios, Thessaloníki (late 4th century), the heads of the saints seem to have been inserted in the mortar in one piece. The indirect method is the one most used in the 20th century. In the workshop, the mosaic is first set in reverse with glue on paper or cloth and then applied to the floor or wall. The technique permits preassembling of mosaics intended even for curved surfaces, cupolas, or apses. It has been hypothesized that behind the enormous output of floor mosaics in the Roman era lay similar production methods which had developed out of the tray procedure described above. The introduction of wall mosaics led to experimentation with the spacing and angling of tesserae. The solidity of floor mosaics depended on a close-set texture, but in wall mosaics, in which the element of wear was no longer relevant, the organization of the surface could become looser. For several centuries, a very wide spacing of the tesserae was cultivated, and the placing of cubes at irregular angles was regarded as important to the over-all effect of wall mosaics. These tendencies reached the extreme in the 7th and 8th centuries, in mosaics of the chapel of S. Venanzio in the Lateran Baptistery, Rome

SCALA, New York

Wide-spaced tesserae set at irregular angles in the head of a saint, detail from mosaics in the Chapel of S. Venanzio, Lateran Baptistery, Rome, *c.* 640 AD.

(*c.* AD 640–49), and in the fragments of the decoration of Pope John VII (AD 705–07) in the old St. Peter's in the Vatican. Later periods preferred a somewhat closer setting, but the irregular surface continued to be in fashion for most of the Middle Ages.

Tilting of tesserae

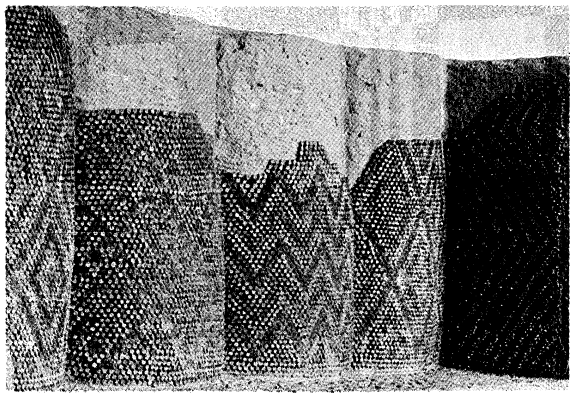
The tilting of tesserae became an art in itself. In 6th-century Byzantine mosaics, there evolved a new technique whereby gold and silver tesserae were set at extremely sharp angles to enhance reflection. By pointing their mirror ends downward in the direction of the onlooker it was possible to secure maximum light effect. In Hagia Sophia at Istanbul, the enormous gold areas in the wall mosaics of the emperor Justinian are set with cubes tilted this way. In one particularly dark corner, the tesserae are not only tilted downward but are also turned slightly sideways to catch the light from a nearby window. A similar technique, based on a high degree of tilting of the gold tesserae in unlit areas, can be observed in the mosaics of the Dome of the Rock in Jerusalem (*c.* AD 690).

Haloes set with tilted cubes that bring out the circle of light surrounding the heads of holy figures became common in Byzantine mosaics of the 6th to 7th centuries, as is seen in the mosaic panels dating from this period in the church of Hagios Demetrios, Thessaloníki. Striking examples of such haloes are also found among mosaics that were put up in Hagia Sophia in Istanbul in the 9th century, above all in a panel with the kneeling emperor (Leo VI?).

Effects such as those described above are unthinkable without the accumulated experience of the craftsman-artist. In the 20th century, mosaic increasingly has become an art divided between the inventor who furnishes the design and the worker who executes it. It may be that the dry character of many modern mosaics can be ascribed to the fact that the artist no longer puts his thumb on every tessera.

PERIODS AND CENTRES OF ACTIVITY

Ancient Near Eastern mosaics. Among the cultures of the ancient Near East there is one remarkable occurrence of a mosaic-like technique: the exteriors of some large



Columns decorated by the Sumerians in a mosaic-like technique with polychrome terra-cotta cones, from Uruk, Mesopotamia, early 3rd millennium BC. In the Staatliche Museen zu Berlin.
By courtesy of the Staatliche Museen zu Berlin

Terra-cotta cone mosaics

architectural structures, dating from the 3rd millennium BC, at Uruk (Erech) in Mesopotamia, are decorated with long terra-cotta cones imbedded in the wall surface. The blunt, outer ends of the cones, coloured in red, black, and white, form patterns consisting of zigzag lines, lozenges, and other geometrical motifs. This revetment was decorative as well as functional, for the cones shielded the core of sun-dried bricks from rain and wind. The technique, however, died out and seems to have had no influence on the later development of mosaic.

In western Asia Minor are preserved the earliest examples of the surface-covering technique that lead to mosaic in the present sense of the word. In the town of Gordion near modern Ankara in Turkey, houses have been uncovered with floors made of pebbles set in a primitive mortar. In some of these floors (dated to the 8th century BC), rows of light pebbles form awkward geometrical figures against a background of darker stones. These rudimentary elements of decoration introduced into a crude form of pavement laying spurred artistic imagination and set in motion a process that was to bring spectacular results.

Ancient Greek mosaics. Three main phases can be determined in the development of mosaic art in antiquity. The first, chiefly a Greek matter, involved the gradual perfecting of the pebble medium. The second, which saw the invention and spreading of the tessera technique, took place partly in the Hellenistic Greek world and partly on Roman soil. The third, largely a Roman phenomenon, was characterized by the popularization of mosaic and the application of the medium to new functions. By a process of diffusion, the taste in floor decoration documented at Gordion spread through the Greek-speaking world of the Mediterranean. Its first full flourishing seems to have occurred in late Classical times. Pebble mosaics are found as far west as Sicily (Motya, Morgantina) and, in the east, in the Greek colonies on the Crimea (Cherson). They are preserved in large number at only two sites, Olinthos and Pella, in Macedonian northern Greece.

In the town of Olinthos there are floor mosaics, with elaborate figures and complicated patterns, which were part of the new city culture that developed in the 5th century BC. The Olinthos mosaics also reveal that picture making with light and dark pebbles had by then evolved into an intricate art. Against a ground set with black or blue-black stones stand figures or patterns set with white or slightly tinted ones. Pebble size had become fairly uniform, with diameters from one to two centimetres, but particularly intricate areas, like faces, are set with smaller stones. Very small black pebbles serve as outlines. Although the mortar between them is visible, the pebbles are set close enough so that the pictures appear with regular, not too broken outlines and with some emphasis on detail.

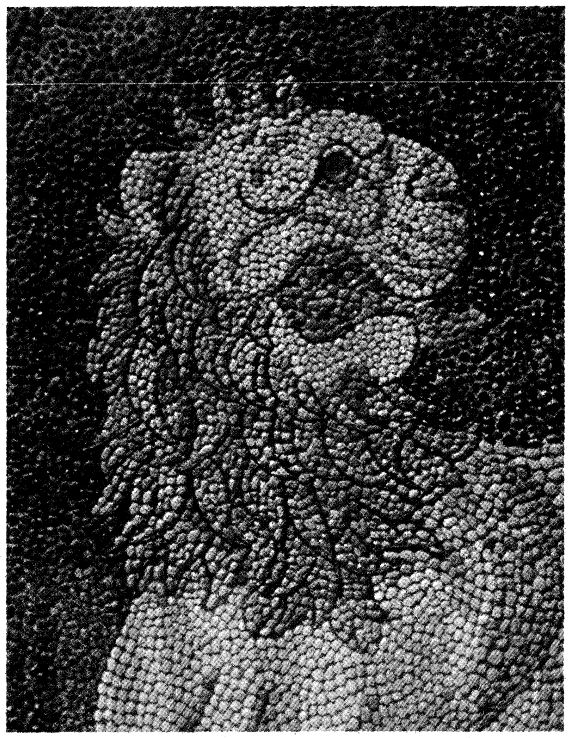
Floors in houses at Pella, dating from the 4th century BC, demonstrate a significant later development of the pebble technique. Floor mosaics then openly vied with wall painting in the rendering of space and realistic detail. This was made possible by the introduction of new materials that eliminated the shortcomings of the ordinary pebble medium. Among the basic changes was an increase in the range of colours. When the demand for particular tints could not be met with pebbles of natural colours, "artificial pebbles" were made and are found in several of the floors at Pella. These "pebbles" have been painted in the required tones—mostly strong green and red—and, to protect the film of paint, have a depression sunk in the middle. The new trend also called for smaller pebbles to permit pictures to be set more closely. To obtain precise delineation of limbs and features, outlines made not with pebbles but with long strips of terra-cotta or lead wire were employed. Pictures made in this technique reflect a taste for heroic hunt scenes and fights with wild animals, themes inspired from court art glorifying the ruler.

Hellenistic mosaics. The next innovation came at the end of the 4th or the beginning of the 3rd century BC, when the introduction of new principles led to the abandonment of the pebble technique. Cut mosaic pieces permitted the nearly complete elimination of the disturbing effects of visible mortar patches, and new materials, above all glass, offered a vast new range of colours. The acceptance of the new methods and materials seems, however, to have come about slowly. In Alexandria (Greco-Roman Museum) there is a mosaic (depicting Erotes fighting a stag and, in the outer border, a frieze of animals) in which cut, triangular tesserae are used together with both pebbles and lead wire. In somewhat later Alexandrian mosaics made with tesserae (late 3rd/early 2nd century BC) lead threads are still in use; for example, in a panel signed by the artist Sophilos and depicting a personification of Alexandria, that is the earliest known example of miniature mosaic work (called *opus vermiculatum*, meaning "wormlike work" because of the close-set, undulating rows of small tesserae).

The Pergamum School

Pergamum, another centre of the Hellenistic world, was particularly famous for its school of mosaics. According to the ancient Roman historian Pliny the Younger, Sosos, one of the most renowned mosaic artists of antiquity, worked in this city. None of his works survives but, thanks to Roman copies, the intentions that underlay his

Dimitri



Greek pebble mosaic, detail from The Lion Hunt, from Pella, Macedonia, c. 300 BC.

"Battle of Alexander and Darius at Issus," detail of the Roman mosaic done in the *opus vermiculatum* technique, from the Casa del Fauno, Pompeii, late 2nd century BC. In the Museo Archeologico Nazionale, Naples.
SCALA, New York

art can be judged. Pliny listed as his most celebrated works a representation of drinking doves and a clever imitation of the "unswept floor" of a banquet room (*asarōtos oikos*). The copies tell of Sosos' phenomenal ability to create trompe l'oeil ("fool-the-eye") effects through a shading and colouring that seems to bring the objects out in full plasticity on the ground on which they are depicted. To call the work merely an imitation of painting may be incorrect. The intense colours and the smooth texture permitted by the new setting technique paved the way for illusionistic effects that went beyond those achieved by painting.

Roman mosaics. Eager to adopt the artistic culture of the Hellenized eastern Mediterranean, the Romans introduced mosaic in this exquisite form in both their domestic architecture and their places of worship. Pompeii has yielded a host of *opus vermiculatum* works datable to the 2nd/1st century BC. Among these the most famous is the Battle of Issus, found in the Casa del Fauno in 1831 (now in the Museo Archeologico Nazionale, Naples). This is the largest of all known works, measuring about 11.22 by 19.42 feet (3.42 by 5.92 metres), in the miniature mosaic technique. This mosaic (which probably copies a work of painting, perhaps a famous picture by Philoxenus of Eretria) and other Pompeiian panels of similar quality are supposed to have been executed by Greek artists, who carried on in the tradition established at Alexandria and Pergamum.

The Romans transformed mosaic from an exclusive art to a common decorative medium. Some of the earliest examples of this new type of floor are in the late republican (2nd century BC) houses at Delos. For rooms of secondary importance and often for floors surrounding the finely designed and executed central *emblēmata* (a featured picture or ornamental motif) in the most important rooms, the Romans developed a simpler, less artistic kind of mosaic. The floors are set with fairly large tesserae with a limited range of colours, some tending toward monochrome (black-and-white). The decorative designs and motifs are also simple and uncomplicated.

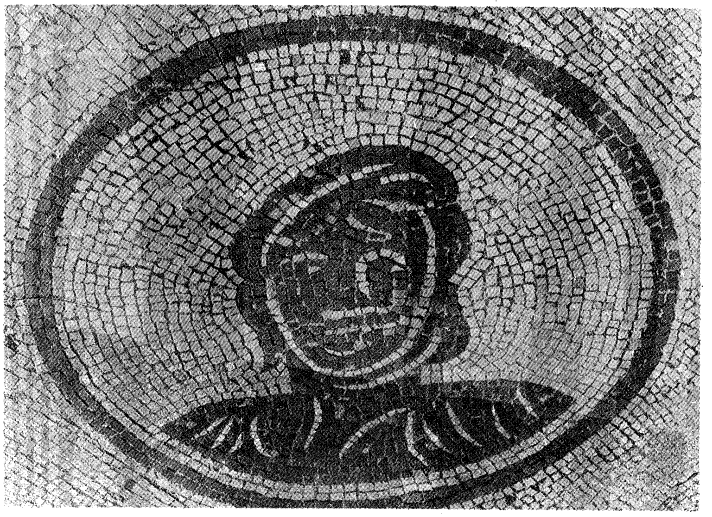
Roman floor mosaics

This new trend in mosaic floors was probably stimulated by new and functional ways of thinking about the role of floors in architecture. To the practical Romans it may have seemed illogical that floors destined for rough wear should bear delicate pictures. Moreover, the demand for large-scale mosaic making brought about by the colossal urban expansion in the 1st century AD made the development of quicker and simpler techniques imperative. The aim of the Romans seems to have been to create a style, technique, and form of composition that would be simple and functional. Competition with painting in illusionistic and coloristic refinement was therefore abandoned; *emblēmata* gave way to decorative elements distributed over the floor in one large overall pattern or to figure compositions taking the full floor plane; and polychrome gave way to monochrome mosaics (which may have been easier to produce). Enormous floors in the baths and in the courtyards of warehouses (1st to 3rd century AD) at Ostia, Rome's port at the mouth of the Tiber, are the best preserved examples of the monochrome style.

The expressionist Roman style, which flourished in Italy, penetrated into the former Greek cities in the eastern part of the empire, but polychromy and types of composition based on the framed picture persisted with especial tenacity due to strong local Hellenistic traditions. A splendid series of *emblēmata* (2nd century) with mythological representations, allegories, and scenes from the theatre have been uncovered at Antioch in southern Turkey. They prove the existence of a school there of mosaicists of particular brilliance. Recent research has pointed to the African provinces as the site of another, highly active school with a taste for larger, dramatic compositions. Influence from these areas may have been responsible for the renewed opulence, represented by a vivid polychrome pictorial mosaic, which reappeared in Roman art in late antiquity. Outstanding examples of this renewal are the mosaics in the Roman villa of Casale (*c.* AD 300) near Piazza Armerina, Sicily. The mosaic decoration of this vast palace complex culminates in the gallery of the Large Hunt, which contains a scene of animal hunting and fighting covering an area of 3,200 square feet (300 square metres).

It is generally agreed that in the course of the 3rd century the status of mosaic was radically altered. Already in Hellenistic times the medium had been employed for other ends than floor covering and had become part of the embellishment of the fantastic garden architecture of which the rulers of the period seem to have been particularly fond. Reflections of this tradition in the 1st century AD are the mosaic-covered fountains in the mansions at Pompeii and Herculaneum and mosaic panels and niches in rustic banquet halls and artificial grottoes at The Golden House of Nero in Rome and his villa at Anzio. Mosaic fragments and imprints of tesserae in the vaults of baths and buildings of similar size demonstrate that mosaic gradually was introducd into new fields. Equally important is the evidence that mosaic was used to depict sacred images. On some of the monochrome floors at Ostia are

SCALA, New York



Roman monochrome floor mosaic in the Portico delle Corporazioni, Ostia, Italy, 3rd century AD.

Cultic mosaics

scenes pertaining to animal sacrifice and to the cult of the dead. Three monuments of the 3rd century inform of another new practice introduced at this time, that of putting mosaic pictures of religious importance on walls: a niche mosaic with the god Silvanus from a temple of Mithra at Ostia; a Christian wall and vault mosaic depicting Christ as Helios, the Sun God, in a mausoleum under St. Peter's, Rome; and a decoration, now lost but recorded in a 17th-century drawing, of a chapel for the Lupercalian worship at Rome. It has been pointed out by modern scholarship that the new role gradually assumed by mosaic must be related to the corresponding decline in interest in three-dimensional representation. The cultic mosaic took over the function of the cult statue, mosaic being that two-dimensional medium which was considered most capable of convincingly expressing religious ideas in visual form.

Early Christian mosaics. Present-day insight into the crucial, early phase of this part of the history of mosaic is limited because of the loss of nearly everything that was made in the field during the first half of the 4th century. Nevertheless, as indicated above, it seems certain that wall mosaics had come into use in Roman art well before Emperor Constantine's edict of toleration of the Christian faith in AD 313. Considered to be among the earliest Christian wall mosaics in Rome are those in the church of Sta. Constanza built about AD 320–330 as a mausoleum for Constantine's daughter. The content of the pictures is almost completely Dionysiac and pagan, but a series of small format scenes from the Old and New Testaments were included among the non-Christian pictorial elements of the decoration. Obviously an independent Christian pictorial program for buildings of Sta. Constanza's size and complexity had not yet been developed; and, probably in lieu of that, a Dionysiac program had been chosen because its many allusions to the symbolism of wine lent themselves to a Christian interpretation.

Other monuments of the 4th century bear similar marks of transition. Floor mosaics in the cathedral complex at Aquileia demonstrate that the church before and immediately after Constantine's edict of tolerance of the Christian faith in AD 313 adhered to the late antique tradition of placing religious pictures in pavements. In the earliest group of Aquileia mosaics (*c.* AD 300) objects and animals symbolize the Good Shepherd, while the later group (second decade of 4th century) contains scenes from the story of Jonah, symbolic animals, such as the deer and the lamb, and a representation of the bread and the wine. Before long, pictures of this character were banished from floors, and simpler and more general symbols took their place.

The latest of the preserved transitional works, a decorated cupola in a mausoleum, possibly imperial, at Centcelles (now Constantí, Tarragona), Spain, seems to have been made not long after AD 350. This very fragmentary decoration has yielded important information about a stage of increasing mastery in the handling of the medium. The scenes from the Old and the New Testament are presented with greater self-confidence and occupy a full, broad zone in the lower part of the cupola. Yet, below it is a stag hunt, rich in symbolic content but adhering closely to the patterns of profane floor mosaics. Stone tesserae dominate in the lower zones, but glass cubes are found in large quantities in the upper. Glass, with its stronger colours, was doubtlessly concentrated in this area intentionally. The zenith of the cupola, weak in lighting and distant from the spectator, needed tesserae of strong reflecting power to make it possible to read its decoration.

A series of large, in part well-preserved mosaics make it possible to follow the progress of the art in 5th-century Italy. Ravenna and Rome have several important works, while Naples and Milan have preserved enough to suggest that workshops of high artistic standard must have existed in many of the large cities of the peninsula. In these works, the tendency to clarify and even underline the content of religious pictures with the help of colour is brought to its full peak. The swing towards a greater employment of glass reached a point at which the mosaics are almost entirely made of this material. In what must be regarded as a late but vigorous revival of the painterly illusionism of antiquity, there is an audacious blending of colours. Among the high points of this trend are the flaming visages of angels in Sta. Maria Maggiore, Rome (*c.* AD 432–440) and the spiritualized physiognomies of St. Bartholomew and his fellow apostles in the Baptistery of the Orthodox, Ravenna (*c.* 450). But the designer's mastery and sophistication are nowhere more overwhelmingly illustrated than in the glowing interior of the Mausoleum of Galla Placidia (*c.* 450) at Ravenna, with its blue star-filled mosaic dome, and in the decoration of the Naples Baptistery of S. Giovanni in Fonte (5th century), with its hypnotizing glimmer.

Development of an iconographic tradition

A Christian language of pictures (iconography) was now developed, and its grammar worked out. In cupolas, the centre tended to be reserved for depictions of Christ or the cross. In apses there was a trend toward static and symbolical representation of holy figures and a reduction of detail. On the walls of the nave of basilicas were scenes from the Old or the New Testament or both. The largely intact decoration of the church of Sta. Maria Maggiore, Rome, throws some light on the principles involved.

Old Testament scenes are distributed on the side walls of the nave in panels measuring about six by six feet (1.9 by 1.9 metres). There is one panel below each of the basilica's large windows. The pilasters (columns projecting shallowly from the surface of the wall) between the windows (restored but repeating the original disposition) serve as outer frames for these panels. Before the restoration of the church in the 16th century there were also inner frames, made of stucco; in addition, each panel was adorned with a small pediment (triangular gable) of the same material and thus appeared as if enshrined by a small aedicula (a pedimented niche). The classical rules governing the relation between the architecture of a building and its decoration may be expected to leave their mark on Christian mosaic art for a long time.

Early Byzantine mosaics. Mosaics made in Ravenna for the Ostrogoth king Theodoric (AD 493–526) are the first full manifestations of Byzantine art in the West. As seen in two of the foremost works from his time, the Baptistery of the Arians and the church of S. Apollinare Nuovo, the gold background now dominates. Accompanying it was silver, a novelty among the mosaics of Italy. In S. Apollinare Nuovo, the faces and hands in several of the Christ scenes are set not with glass tesserae but with cubes of stone. Stylistically, these mosaics are characterized by more static figures and less depth and plasticity than in those of the 5th century.

Developed integration of architecture and mosaic decoration

Another remarkable element is the movement, now fully developed, toward an integration of architecture and mosaic decoration. This is most clearly seen in the basilica. In the church of S. Apollinare Nuovo, the mosaics are no longer inserted panels but form a continuous "skin" that covers every inch of the wall. The size of the windows and even their number have been reduced, apparently to provide more wall space for pictures. Figures have grown in size, to match the dimensions of architectural components, and they seem to have taken the place of pilasters in the articulation of the room. This trend is given truly monumental expression in the choir of the church of S. Vitale at Ravenna, dating from *c.* AD 526–548. A profusion of decorative elements is spread like tapestry over the walls and vault, the panels of the emperor Justinian and his consort Theodora near the apse embodying the new spirit in their colour-laden pageantry.

In the East, the circular church of Hagios Georgios at Thessaloníki, Greece, shows Byzantine mosaic at its earliest flourishing (*c.* AD 400). Its partly preserved mosaics display a disposition related to that of the Baptistery of the Orthodox at Ravenna, with a lower zone containing Paradisiac architecture; above this a zone with standing and walking figures and in the centre of the cupola a medallion with a figure of Christ. A uniform gold background dominates the two lower zones, at a time well before it had come into general use in the West. Silver is

found in profusion, used for the background in the central medallion as well as a means to enhance the radiation of light from all parts of the mosaic. In Hagios Georgios another feature emerges fully developed: in the figures of saints the material for faces and hands is chiefly natural stone, its gentle gradations contrasting spectacularly with the violent juxapositions of coloured glass tesserae of the hair and the garments.

Effects of this kind, which are the hallmark of Byzantine mosaic technique, seem to derive from intentions wholly dissimilar to those which had determined the development of early Christian mosaic art in the West, where the illusionism of Greco-Roman art persisted for a long time. In the great Ravenna mosaics of the 5th century, pictures illustrating the narrative of the Bible or expounding the dogmas of religion were still done in the painterly style of Roman mosaics and wall painting. During the same period, mosaic art of the Eastern Empire, having abandoned conventional illustration, was boldly exploring the way that lay open, in mosaic art, toward a new kind of imagery.

In the mosaics of the 6th century are found the earliest refinement introduced by the Byzantines to enhance the brilliance of gold tesserae. This refinement, already described, involved setting gold cubes at oblique angles to direct their reflections toward the viewer. Used in haloes, the tesserae, obliquely set, convey to the holy figures a miraculous aura of light. The visages of the saints, with their dull stone surfaces and hues reminiscent of actual human skin, add a touch of mysterious reality to this theatre of effects.

Splendid mosaics from many parts of the eastern Mediterranean testify to the continuous cultivation and improvement of these effects. In the city of Thessaloníki the mosaics in the churches of Hosios David (5th century AD) and Hagios Demetrios (6th and 7th centuries) ex-

Rene Percheron—J.P. Ziolo

Gold tesserae reflecting light to the viewer; from the Byzantine votive mosaic showing St. Demetrios between two donors, from Hagios Demetrios, Thessaloníki, Greece, 7th century.

emplify the trend, which is also expressed in apse decorations preserved at Cyprus (church of the Panagia Angeloktistos, at Kiti, and of the Panagia Kanakaria near Lithrankomé; both 6th century) and in the Monastery of St. Catherine, Sinai Desert, founded by Justinian.

Apart from the gold ground, which had considerable impact, the technical subtleties essential to these mosaics met very little response outside Byzantium. When Byzantine artisans operated in foreign territory, they brought their particular techniques with them. Again and again the impact of this tradition was felt in the West, though, at its purest, mostly as short-lived episodes. To judge from a few surviving fragments, mosaics executed under Pope John VII (AD 705–707) in a chapel in St. Peter's, Rome, might have been the work of artisans summoned from Byzantium. Technical and stylistical features demonstrate that the mosaics executed under the earliest Muslim rulers, in the Dome of the Rock at Jerusalem (*c.* AD 690) and in the Great Mosque at Damascus (*c.* AD 715), are certainly the work of specialists called from Byzantium. Sources testify that even the mosaics in the mosque at Cordova, Spain (AD 965), were made by Greek craftsmen.

The floor mosaics in the great palace of the Byzantine emperors at Istanbul—with their pastoral scenes, fights with wild animals, and figure groups taken from pagan mythology—testify to an undercurrent of classical taste in Constantinople. The date, which according to archaeological evidence must be placed as late as around the year AD 600, demonstrates the tenacity of that taste in the midst of the Christian milieu of the Byzantine metropolis.

Middle Byzantine mosaics. Scholars have been concerned to discover how Iconoclasm, the dispute concerning images during the 8th and 9th centuries, may have influenced the course of Byzantine art. In some respects, at least, mosaic reflects very little change. The main source of knowledge about the state of mosaic in the time shortly after the end of Iconoclasm is Hagia Sophia at Istanbul. Parts of the redecoration that the church underwent in the last half of the 9th century have been uncovered in recent times. In their colour and technique these show a continuation of the early Byzantine tradition: the preference for rather strong, clear tints, and the effects created by such techniques as the tilting of tesserae and the turning of gold cubes. The preoccupation with light seems stronger than ever: in badly lit places in the vestibule and gallery, the gold ground displays a high percentage of silver cubes among the gold ones to add to the sparkle. Stylistically, new ground had been broken. Particularly in faces, the tesserae are set in wavy lines which break up the modelling in bandlike configurations. Linearism (the expression of form in terms of line rather than colour and tone) had taken a great step forward.

Mosaics of Hagia Sophia

In the arrangement and distribution of pictures new features are visible. In the apse of Hagia Sophia, the Virgin with the Child sits surrounded by a vast expanse of gold. She is one of the first of a family of similar majestic madonnas, the most striking of which is in the Cathedral of Torcello near Venice (12th century). The tendency to depict icon-like, motionless mosaic figures isolated on a gold background has pre-Iconoclastic precedents, but from the 9th century onward it became a leading decorative principle.

Nineteenth-century drawings show that the decoration of Hagia Sophia also included comprehensive series of saints. Of these saints, which stood in rows on the nave walls above the galleries, only a few have survived. According to the drawings, those of the middle zone represented prophets, those of the lower, holy bishops. Higher up there may have been a guard of angels and in the centre of the cupola, probably a mosaic of Christ. The disposition of the pictures, in other words, may have corresponded to that which at this time was being tried out especially for the new church architecture and which was to become the accepted system of decoration in the middle Byzantine churches.

Interplay of architecture and mosaic at Daphni

The monastery church at Daphni, near Athens, contains one of the best preserved decorations of this type (11th century). The building belongs to a category of central-plan structures that had come into fashion and was to dominate for centuries both in Byzantium and in other areas under the influence of the Orthodox Church. The interior of the church at Daphni displays a layout which, compared with the wealth of detail of the early Christian

period, appears single-minded and concentrated. In the centre of the dome is a medallion containing a colossal bust of Christ as Pantokrator, the All-ruler. In the lowest part of the dome, separated from the medallion by a broad zone of gold, stand prophets with their scrolls. Further down, there may originally have been medallions with portraits of the Evangelists. In the four arches that carry the drum of the cupola are scenes from the life of Christ which, with eight more Christological scenes in the transepts, formed a cycle devoted to the central feasts of the church. The Virgin is represented in the apse, her guard of archangels on the side walls of the sanctuary. About thirty saints, depicted either as busts or as full-length figures, fill the remaining wall space. In the vestibules are more scenes from the life of Christ and the remains of a cycle devoted to the life of the Virgin. Golden frames with floral ornaments surround the panels, and gold once covered every inch of wall between them.

Rene Percheron—J.P. Ziolo

Interplay between architecture and mosaics in the monastery church at Daphni, Greece, 11th century, crowned with a Byzantine dome mosaic Christ Pantokrator.

The ensemble represents a visualization of the Christian cosmos, its effect created by an intricately conceived interplay of pictures and architecture. The worshipper who moves within this golden shell finds its world of pictures thoroughly involved with space. Space in fact fuses the decoration into one giant image, in which the ruler, hailed by the prophets surrounding him, presides in his sphere above the host of saints that people the lower part of the room.

Subtle spatial devices animate the individual pictures; figures of saints, their two-dimensionality emphasized by their outlines, appear in niches sunk in the wall or lean forward in the interior curves of arches. The 20th-century Austrian scholar Otto Demus, in studies on the aesthetics of middle Byzantine mosaic art, has coined the term space icons for this kind of imagery, in which the forms of architecture collaborate to make the solemnly stylized figures appear with unexpected tactility. As shown by Demus, the spatial element contributes to the narrative scenes also. In the four arches, for example, the hollow plane on which the scenes from the life of Christ unfold adds a dimension of spatial realism to the total image. This is most clearly to be observed in the Annunciation scene, where Mary and the Angel face each other across a stretch of real space. The figures share or are made to appear to share the room with the beholder.

The "classical system"

The "classical system," as this close interrelation of architecture and mosaic has been called, was probably perfected in the course of the 9th to 10th centuries, but the earliest fully preserved examples are from the 11th to 12th. Besides Daphni, Greece owns two more monuments of this kind, the monastery church of Hosios Loukas in Phocis and the Nea Moni on Chios (both 11th century). Similar churches are found in such widely distant places as Kiev (Hagia Sophia, 11th century) and Palermo (Martorana, *c.* 1150), both the products of strong Byzantine influence. The system, however, is not identical in any of these. The churches belong to the same general type, but their plans and elevations vary and thus require variations in this disposition of pictures as well.

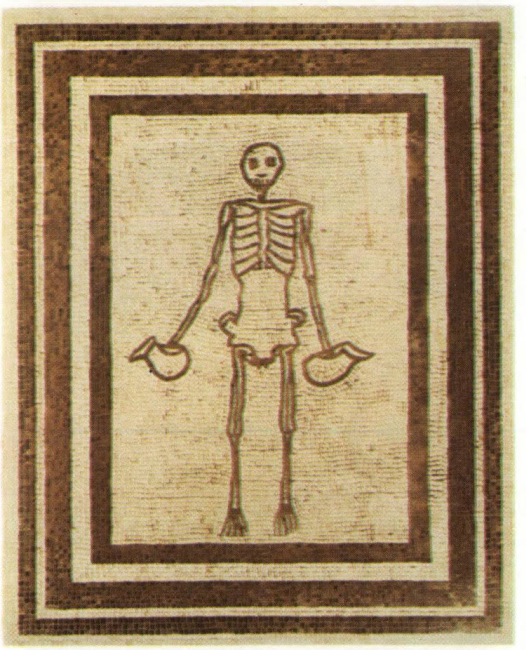
The classical system with its emphasis on totality may have led to the gradual toning down of the many splendid effects of the earlier tradition for the sake of the equilibrium and clarity of the whole. At Daphni, for example, the rich, tapestry-like character of earlier mosaic has given way to a controlled, less sparkling range of tints. The reds and yellows are restricted, their function in the all-over scheme taken over by the gold of the background. Sombre, often hard blues, greens, and violets are preferred to the lighter ones. Compared with the Hosios Loukas and the Nea Moni mosaics, which retain more of the older colour scheme (the latter almost to the point of brutality), the Daphni mosaics appear cool and intellectual, an impression further conveyed by their elegant style. Actually they belong to a new phase of Byzantine art which took its name from the dynasty of the Comnenus (AD 1081–1185). This style appears at its most refined in Hagia Sophia, Istanbul, in a panel depicting the Virgin flanked by the emperor John Comnenus II and his wife Irene. The practice of tilting the gold tesserae also seems to have been abandoned, for it is not found at Daphni nor in any of the mosaics that are examples of the fully developed classical system. Silver was reduced to the single role of depicting the light emanating from God and Christ. This drying out of the effects of light and colour was partly compensated for by a perfectionist setting and spacing of the tesserae.

Medieval mosaics in western Europe. The prestige, both cultural and political, enjoyed by Byzantium in the Middles Ages led to a widespread imitation of its arts. Art objects in great number were imported to the West from Constantinople and other Greek centres. Individuals or communities outside the realm of Byzantium, however, were able to secure Byzantine artisans for the execution of monumental mosaics. Abbott Desiderius of the abbey of Montecassino in Italy, for example, called specialists in many crafts from Constantinople to decorate his new basilica (dedicated AD 1071). Among these were mosaic workers. Of particular importance is the fact that he took care to see that young local artists were trained by the foreigners. This was the pattern that was followed where Byzantine experts were temporarily called in.

Sicilo-Byzantine mosaics

The Norman rulers of Sicily, who vied with the Byzantines for control of the Mediterranean, molded their representational arts largely on those of the great Eastern power. The existence, at Palermo, of a central-plan church (Martorana) embellished according to the classical system has already been noted. In other 12th-century churches in Sicily, the Byzantine element is blended with western Mediterranean traits. Cappella Palatina, the palace chapel of the royal residence at Palermo (*c.* 1143 and later), for example, is a synthesis of a centralized middle Byzantine church and a basilica. The building therefore called for a hybrid program. According to Western custom, the mosaics of the basilical parts depict narrative cycles: scenes from the Old Testament and from the lives of SS. Peter and Paul. In the centralized part of the church most of the features belonging to the classical system are at hand. There is a bust of the Pantokrator in the dome, surrounded by angels, but as a con-

Dionysus on a Tiger, from the House of the Faun, Pompeii, 2nd century BC. In the Museo Archeologico Nazionale, Naples.



Skeleton of a Cup-Bearer, from the House of the Faun, Pompeii, 2nd century BC. In the Museo Archeologico Nazionale, Naples.

Detail from the mosaic of the Nile, known as the "Barberini" mosaic, from the Sanctuary of Fortuna Primigenia, Palestrina, *c.* 80 BC. In the Museo Archeologico Nazionale, Palestrina, Italy.

Roman mosaics

Interior court with mosaic of Neptune and Amphitrite, from the House of Neptune and Amphitrite, Herculaneum, 1st century AD.

Plate 1: (Top left, bottom left) P. Marzari, (top right, bottom right) SCALA, N.Y.

Early Christian, early Byzantine, and Islamic mosaics

The Good Shepherd, Mausoleum of Galla Placidia, Ravenna, *c.* 450.

Mosaics in the church of S. Vitale, Ravenna, *c.* 546–547.

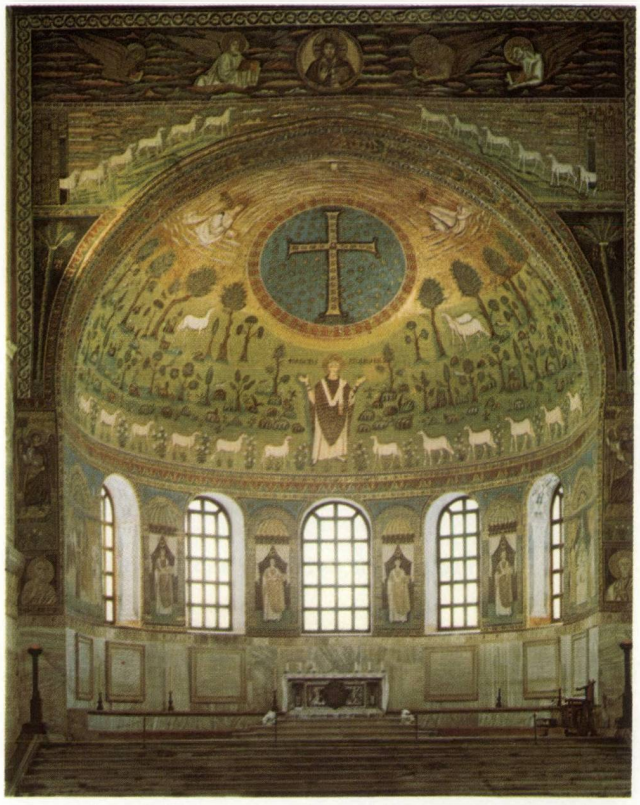
Floor mosaic from the bath of the Palace of Khirbat al-Mafjar, Jericho, *c.* 743.

Detail from the mosaics of the great mosque of Damascus, Syria, 715.

Plate 2: (Top) SCALA, N.Y., (left) Shostal, (centre right) Micha Bar-Am, (bottom right) Ciccione—Rapho Guillumette

Dome of the Baptistery of the Orthodox, Ravenna, *c.* 450.

Apse in the church of S. Apollinare in Classe, Ravenna, second half of the 6th century.



Crossing of the Red Sea, in the church of Sta. Maria Maggiore, Rome, first half of the 5th century.

Detail from the story of Jonah, pavement mosaic in the cathedral at Aquileia, second decade of the 4th century.

Plate 3: SCALA, N.Y.

Panel depicting the Virgin and Child with the emperor John Comnenus II and the empress Irene, in Hagia Sophia, Istanbul, *c.* 1118.

Detail of the floor mosaics in the Great Palace of the Emperors, Istanbul, *c.* 600.

Middle and late Byzantine and medieval mosaics

Crowning of the Virgin, detail of the apse mosaic by Jacopo Torriti in Sta. Maria Maggiore, Rome, 1295.

Apse mosaic of Germigny-des-Prés, near Orléans, France, 9th century.

Plate 4: (Top) Andre Held—J.P. Ziolo, (centre left) Ara Guller, (bottom left) Giraudon, (bottom right) SCALA, N.Y.
Plate 5: Shostal

Interior mosaics of St. Mark's, Venice, 11th through 13th centuries.

Dome of St. Peter's, Rome, mosaics after cartoons by Cavalier d'Arpino, begun in 1576.

Modern mosaics

Sculpture mosaics by Jeanne Reynal, 1970–71.

Vault of the Chigi Chapel, by Luigi da Pace after cartoons by Raphael, 1513. In the church of Sta. Maria del Popolo, Rome.

Mosaics at University City, Mexico City: (left) Central Administration Building, mosaic by David Alfaro Siqueiros, 1952–53; (right) Library, mosaic by Juan O'Gorman, 1951.

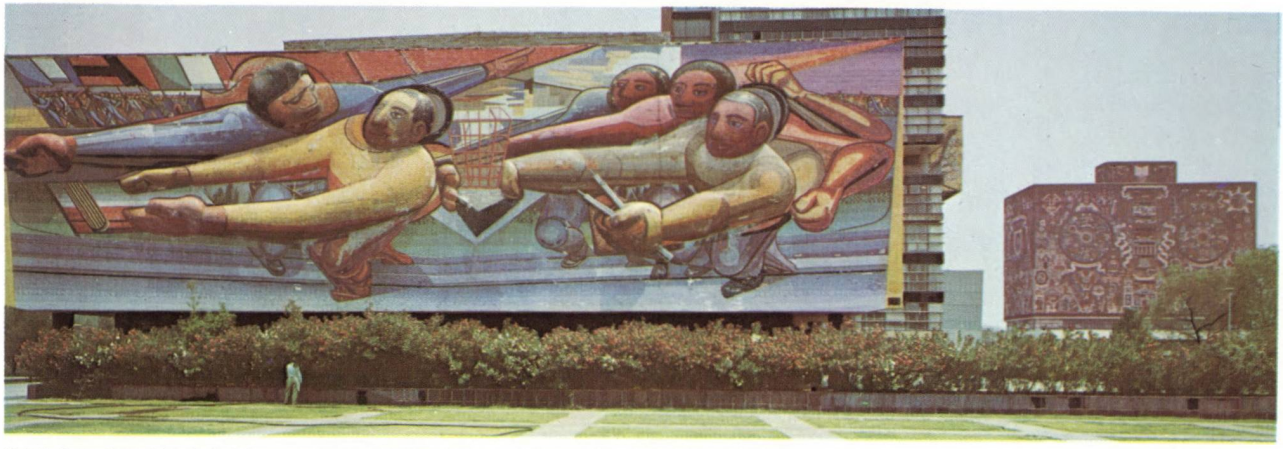


Plate 6: By courtesy of (centre left) Jeanne Reynal; photographs, (top, centre right) SCALA, N.Y., (bottom) Shostal

cession to the longitudinal disposition of the church, the Pantokrator reappears in the apse.

Martorana and Cappella Palatina were decorated by Byzantine artists, a fact borne out by the brilliant technique, the purity of the Comnenian style, and the adherence to Eastern iconographical prototypes. Two large basilicas are among the highpoints of this Sicilo-Byzantine flowering: the cathedral of Cefalù (*c.* 1148) and the church of Monreale, the last of the mosaic churches of Sicily (*c.* 1180–90). Demus has pointed out the extraordinary homogeneity of style and technique in the Monreale mosaics, which constitute the largest decoration of this kind in Italy. He has also shown that the Monreale mosaics are not executed in the refined and softly curved style that dominates in Cappella Palatina and at Cefalù. Monreale is infused with a more agitated and expressive style which, however, has nothing local or provincial about it. It was the late Comnenian style of Constantinople which had then reached Sicily—a testimony to the unbroken artistic contact that existed at this time between the Norman court and Byzantium.

Italo-Byzantine mosaics in Venice

Venice, for a long time commercially active in the eastern Mediterranean, enjoyed similar but more long-lasting artistic connections with Constantinople. A church that in the 11th century must have looked exotic as well as old-fashioned, St. Mark's was copied after the venerable Church of the Holy Apostles at Constantinople, an early Byzantine, many-domed type that had long since gone out of favour.

The mosaic work undertaken at St. Mark's lasted more than two centuries, from the end of the 11th to the middle of the 14th century. The many and marked stylistic disparities can be ascribed in part to the changes that affected Byzantine art during this period; but they may even have been caused by the freer, Western organization of the Venetian workshop, which allowed artists to develop and cultivate their own personal styles. Among the variety of styles, the Byzantine element is dominant, but it is modified by local tendencies and particularly by strong Romanesque impulses.

The workshops established in Sicily and Venice were active in the neighbouring areas. Mosaics in the cathedral at Salerno (*c.* 1190) and in the monastery at Grottaferrata near Rome (*c.* 1200) are regarded as products of the Sicilo-Byzantine school. The apse decoration of the cathedral at Ravenna (early 12th century), of which fragments survive, seems to have been the work of mosaicists from Venice and, in Florence, Venetian artists decorated the dome of the Baptistry (1225–1330). The much later mosaics on the facade of the cathedral at Prague (14th century) are also Venetian.

The Roman school

In the early Middle Ages, Rome had been able to maintain and defend a mosaic tradition of its own despite the Byzantine hegemony in the arts. At a period when Iconoclasm had loosened the ties between Byzantium and the West, at the end of the 8th century and the beginning of the 9th, the influence of the Roman school extended even into the Carolingian Empire. The apse mosaic of Germigny-des-Prés, France (between 799 and 818), is a product of this influence. In Rome, the pontificate of Paschal I (817–824) left three monumental decorations which constitute the best sources concerning the artistic intentions of this time: the sanctuary mosaics of the churches of Sta. Cecilia in Trastevere, Sta. Maria in Domnica, and Sta. Prassede (attached to the latter is the domed chapel of S. Zeno, also fully decorated with mosaics). The iconographic schemes of these 9th-century churches largely reflect local 5th- and 6th-century church decorations. Also remarkable is the return of many of the technical idiosyncrasies of the early Christian period, including the employment of large tesserae for faces and dress and the use of glass for all parts of the composition. The setting of the tesserae is, however, loose and disorganized; and their varying shapes suggest that they are reused cubes taken from older monuments. Stylistically, the mosaics of the Paschal period, with their extreme two-dimensionality, are related to Byzantine mosaics of the 9th century (the cupola of Hagia Sophia, Thessaloníki), but technically the differences could not be greater.

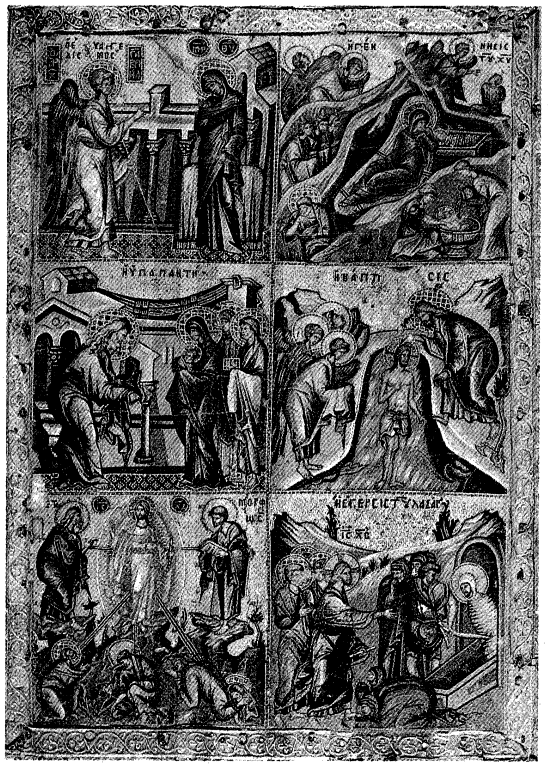
Little is known of the art in 10th and 11th century Rome. But from the 12th century, a group of works testify to continued inbreeding. The apse decoration of S. Clemente (the 1st half of the 12th century), for example, contains a scroll pattern (rinceaux) reproduced from a 4th-century decoration in a technique that in many respects resembles that of the Paschal mosaics. This isolation, or resistance to foreign influence, seems to have been broken in the 13th century. In 1218, Pope Honorius III asked the doge to send craftsmen for the decoration of S. Paolo Furori le Mura (1218). Several important mosaics from the later part of the same century reflect the trends current at that time in Byzantine and Italo-Byzantine mosaics. The mosaics by Jacopo Torriti in the apse of the basilica of Sta. Maria Maggiore (*c.* 1290–1305) are among the finest of these. They show a mingling of Western medieval and Early Christian iconographical features, such as a scene of the crowning of the Virgin surrounded by the scrolls of a fleshy, classicizing rinceaux; but the lower zone with scenes pertaining to Mary reflects Byzantine influence, an influence also seen in the technique and colour scheme of the mosaics.

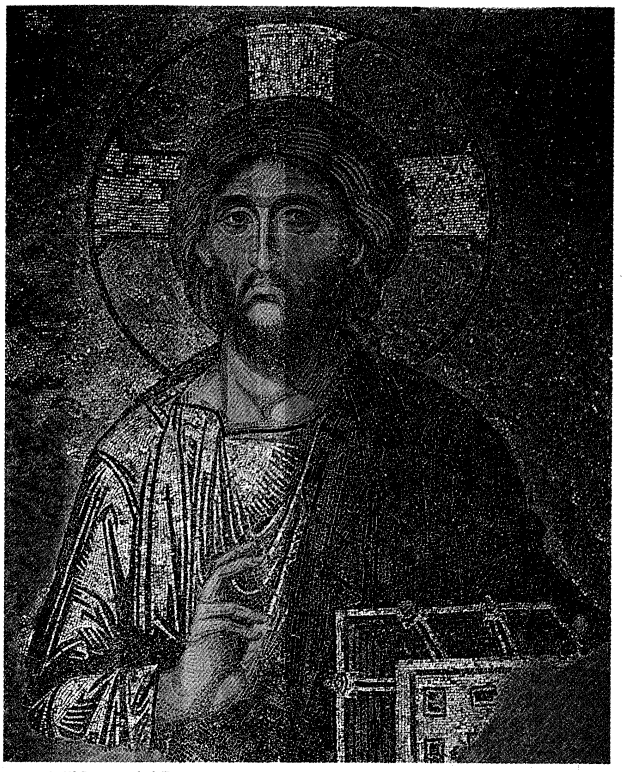
Floor mosaics had a renaissance in the West that was unmatched in the Eastern Empire. In the early Middle Ages, the Byzantines developed their particular form of floor covering consisting of a geometrical mosaic made up of pieces of marble of various sizes and shapes (*opus sectile*) usually with some tessera work added for special coloristic touches. This art, called Cosmati work, spread to the West, in which, however, there was also a revival of the tesselated pavement. This pavement, which had survived the Dark Ages in a very primitive form, reemerged in Italy in the 11th century in greater splendour. There are impressive remains of such floors in many of the larger Italian churches. The fashion spread to other parts of Italy and even to France and Germany. In France, where large floors were produced in quantity in the 12th century, fragments of outstanding quality are found; for example, in Saint-Nicolas' at Reims (last half of 12th century). Cologne seems to have been a leading centre for this art in Germany. The style of the floors is usually one of simple outlines and light colours, though in some cases figures and ornaments appear against dark ground. The programs draw their inspiration from many sources, such as textiles, early floor mosaics, and the sculptural ornamentation of churches. An exceptionally well-preserved example is found in Otranto in the Italian province of Apulia, now Puglia (1163–66), where vast floors depict scenes from the Old Testament and mythology (the ascension of Alexander), representations of the Zodiac, and of the labours of the month. A profusion of monsters and fantastic animals fills out the picture of a decoration cast in the Romanesque iconographic tradition.

Cosmati work

Late Byzantine mosaics. The phenomenon called the Palaeologian renaissance (from the dynasty of the Palaeologians, 1261–1453) led to a renewal of Byzantine mosaic art. The stylistic innovations that made themselves felt both in painting and mosaics of the late 13th and beginning 14th century bear witness to one of the most startling changes that ever took place within the framework of Byzantine culture. Bred by a vital humanism, which penetrated westward and laid the foundations for the Italian Renaissance, painting showed a predilection for perspective and three-dimensionalism. A peculiar vivacity invaded religious art, together with a sense of pathos and of the tragic. The results, as expressed in mosaics, were extraordinary.

To respond to the new trend, mosaicists recast their technique. The tessera size generally became smaller than it had been in earlier epochs; and contours lost their rigidity, became thinner, and were occasionally abolished. Colour was reintroduced in a manner that gives the Palaeologian works a striking likeness to the mosaics of the Early Christian period, which, one must suppose, in many cases served the artists as models. An interest in the optical effects of gold apparently returned but rarely, it seems, in the form of the tilting technique. On flat walls, the gold ground was sometimes set in a shell pattern,





(Left) Greek Festival Cycle miniature mosaic, 14th century. In the Museo dell'Opera del Duomo, Florence, Italy. (Right) Christ, detail from the Late Byzantine Deësis mosaic, south gallery, Hagia Sophia, Istanbul, *c.* 1300.
(Left) SCALA, New York. (right) Erich Lessing—Magnum

probably to enhance the play of light on the surface and to avoid a too-uniform brilliance. For domes, a densely ribbed form of cupola construction, which, when covered with mosaics, produces reflections of light that expand like rays from the central medallion toward the figures surrounding it, was preferred. Such domes are preserved in Kariye Cami, the former church of the Chora, at Istanbul, which was reconstructed and decorated as an act of piety by the logothete, or controller, Theodore Metochites in the second decade of the 14th century. Another superb example is found in Fetiye Cami (Church of the Virgin Pammakaristos) in the same city.

The feeling for colour, which is at its most refined in fragments from the decoration of the Church of the Holy Apostles in Thessaloníki (*c.* 1315) and at its most intense in the partly well-preserved cycles in the Kariye Cami, informs one of the greatest mosaic works of art, the Deësis panel in the south gallery of Hagia Sophia in Istanbul (*c.* 1300). In this same panel, the tilting technique reappears (in the cross arms of Christ's halo)—another indication of the retrospection inherent in late Byzantine art.

No mosaic in the true Palaeologian style has survived outside Byzantium. Reflections of it are found, however, in some of the 13th- and 14th-century works at Venice and in the mosaics executed by Pietro Cavallini in the apse of Sta. Maria in Trastevere in Rome (*c.* 1290–1300). Some of the characteristics of the style may have been brought to the attention of the Italian artists through

Portable mosaics

portable mosaics, which despite their small size (generally about two by four to eight by ten inches [five by ten to 20 by 25 centimetres]) are imbued with many of the coloristic and technical features typical of monumental mosaics. Byzantine mosaic icons, the production of which was stimulated during the early Palaeologian era, were manufactured for personal devotion more than for the embellishment of churches and were exported in considerable numbers to the West or found their way there as gifts or booty in the politically troubled 14th and 15th centuries. In works whose quality can be compared with the most splendid of the Hellenistic *emblēmata*, extremely small tesserae, some measuring less than 0.04 inch (one millimetre) square, were assembled in wax or mastic on a board of fine wood. The tesserae material is often exquisite: silver, gold, and lapis lazuli and other semiprecious stones. The icons depict single figures such as saints, Christ, or the Virgin; single Christian scenes such as the Annunciation (Victoria and Albert Museum, London) and the Crucifixion (Staatliche Museen zu Berlin); or even the full Greek Festival Cycle (Museo dell'Opera del Duomo, Florence).

Renaissance to modern mosaics. With the downfall of Byzantium in the 15th century, there perished that milieu in which mosaic had been constantly cultivated and had undergone continuous renewal in response to changing patterns of religious and cultural life. The art lost another foothold in Italy at the beginning of the same century, when changing attitudes about the world and about the function of art eliminated the very bases upon which mosaic had been built. One of the conventions against which the artists of the Renaissance, who were striving for pictorial realism, most strongly rebelled was the use of gold, the other-worldly element most typical of mosaic art.

The decline of mosaics

Although mosaic continued to be used to a certain extent as church decoration, it was a changed art. Some of its traditional glitter was retained, but essentially mosaics became imitations of painting. These imitative intentions were disastrous and led to the loss of knowledge of how to blend colours and handle materials. In earlier mosaics, there undoubtedly had been a distinction between the leading artist of the project, who drew the composition and oversaw the execution, and the ordinary setters of the tesserae. The leading artist, however, almost certainly took a hand in the setting of special parts and was thoroughly trained in the technical side of the production. Now the preparatory work was divorced from the execution: the artist submitted his cartoon and left its transposition into mosaic to artisans. This drew the lifeblood from the art and caused its degradation.

In Italy, many of the great painters of the 15th and 16th centuries delivered designs for decorations in mosaic. Best known among these decorations are the works of the Venetian Luigi da Pace after Raphael's cartoon, in the

dome of the Chigi Chapel in Sta. Maria del Popolo in Rome (1516), and the mosaics made after the cartoons of Titian, Tintoretto, Giuseppe Salviati, and Paolo Veronese to complete the decoration of St. Mark's in Venice. Among the greatest single undertakings of this kind was the decoration of the dome of St. Peter's in Rome, executed in the last quarter of the 16th century from the cartoons of Cavalier d'Arpino. St. Peter's also displays some of the most technically striking mosaic reproductions of paintings ever executed—the much admired altar pictures after originals by 16th- and 17th-century masters. Created for the completion and care of the large mosaics of the two great churches, the workshops attached to St. Peter's and to St. Mark's gradually became centres for the manufacture of mosaics. From them, artists were summoned for decorative work in all parts of Europe. The school of mosaics in the Vatican and the workshops in Venice still have a considerable share in the field, together with the school more recently set up for the restoration of the mosaics at Ravenna.

Nineteenth-century historicism and the breaking down of Neoclassicism's contempt for Byzantine art led to an increasing interest in and demand for mosaics. Improvements in the technique of prefabrication according to the indirect method and in the manufacture of tessera material led to a veritable mass production, which has put its mark on countless churches, town halls and opera houses. Shrill colours and a gleaming, metal-like surface characterize many of these works.

The modern revival of mosaics had several causes. Scholarship and tourism made the monuments of ancient and medieval mosaics available to an art-loving public. Painting, since the last third of the 19th century engaged in the exploitation of colour, at the turn of the century focussed on the problems of colour as the expression of psychological qualities rather than of the external world. Expressionism, which opened the eyes of artists to the art and artifacts of foreign and distant cultures, also turned their interest towards medieval mosaics. The abstract element which these mosaics contain and which springs from the latent conflict between the design and the tessera pattern made them particularly attractive to artists of the earliest decades of this century such as Marc Chagall and Giovanni Serverini. The texture of mosaic was also an attraction. An American mosaicist, Jeanne Reynal, for example, created abstract compositions in which texture is emphasized by a combination of granulated, pebble-sized, and normal tesserae, sparsely spread over a coloured base of portland cement. Many of these mosaics are small and are hung on the wall like paintings.

Renaissance of the mosaic art in Mexico

Mosaic's smooth yet faceted surface is ideal for decorating the large, unbroken surfaces of modern architecture. The greatest modern use of mosaic as architectural decoration is in Mexico, a country with a long tradition of folk mural painting. Realizing the potential of the medium for public enjoyment and education, the government in the 1930s and 1940s commissioned many murals with historical and political themes for public buildings. Later, it became desirable to decorate the exterior walls of buildings, and mosaic was the logical alternative to the less durable murals. Often mosaics were designed by mural painters such as Diego Rivera who, in 1953, designed the immense mosaic on the facade of the Teatro de los Insurgentes. Francisco Eppens also used historical themes in his mosaic decorations of the schools of medicine and dentistry at the National Autonomous University of Mexico (1957), as did Xavier Guerro in the Cine Ermita in Mexico City. Carlos Mérida, however, created abstract mosaic designs in the Reaseguras Alianza in Mexico City. Among the most prolific Mexican mosaicists was the architect-muralist Juan O'Gorman. Of his many mosaic works, the most important is on the exterior walls of the library of the National Autonomous University of Mexico (1951–53), which exemplifies the monumentality of which mosaics are capable. Other works executed by O'Gorman include mosaics on the SCOP, or Secretaría de Communicaciones y Obras Públicas (1952), and a stone mosaic on the facade of the Posada de la Misión Hotel in Taxco. In 1950, O'Gorman began to decorate his own house in Mexico City with phantasmagoric images and symbols from Aztec mythology.

Mosaic in the strict sense of the word is an art in transition. The conventional tessera mosaic is still largely in use, mainly because of the efficient production methods of modern mosaic firms; yet the distinction between this kind of mosaic and other, mosaic-like techniques is slowly being dissolved. A mixture of cubes and larger pieces of glass, ceramics, or stone is one of the variants that occurs, cubes and *objets trouvés* ("found objects") another. After World War II, experiments seem to have moved in the direction of an employment of larger, less regularly cut units. Today, mosaic is just one out of a very wide range of techniques that have as common principle the piecing together of a surface covering with the use of different durable or nondurable materials. (P.J.N.)

Pre-Columbian mosaics. The art of mosaic in pre-Columbian Central America was marked by a combination of great technical skill and widespread use. The representation of a mosaic mask on a stela at Seibal (an upright, freestanding stone slab functioning as a commemorative monument), Guatemala (AD 590), established the early use of the technique in Maya territory, but it became best known from the few specimens surviving from the time of the Aztec empire (*c.* 1376–1519) and from descriptions of others left by the Spanish conquerors. The Mexican lapidaries worked with obsidian, garnet, quartz, beryl, malachite, jadeite, marcasite, gold, mother-of-pearl, and shell, but turquoise above all constituted their favourite material; the excessive richness of the religious ceremonial gave wide range to its employment in ritual paraphernalia of all sorts. The incrustation was laid upon wood, stone, gold, shell, pottery, and possibly leather and native paper and was held in place by a tenacious vegetal pitch or gum or a kind of cement.

Masks, shields, helmets, knife handles, staffs, collars, medallions, ear plugs, leggings, mirrors, animal figures, and cult statues received a covering, in whole or part, of small and irregularly shaped pieces of highly polished turquoise, cut to fit tightly together so as to form a brilliant green surface, varied at times with cabochons (a gem cut in convex form, highly polished but not faceted) of turquoise or other material. Most striking and best preserved of the surviving two dozen major specimens of this art are a mask in the British Museum, London, and a shield in the Museum of the America Indian, Heye Foundation, New York City. Minute pieces of turquoise, stud-

Mask of Quetzalcoatl, turquoise tesserae on wood with mother-of-pearl, Mixtec, from Mexico, 14th–15th century. In the British Museum.
By courtesy of the trustees of the British Museum

ded with cabochon turquoises, completely cover the cedar mask and are laid in symmetrical lines around the eyes and mouth and on the nose; eyes and teeth are of shell inlay. Over the wood of the shield, one panelled and three circular borders of mosaic frame a scene that may relate to the worship of the planet Venus; it is estimated that nearly 14,000 pieces of turquoise make up the decoration. A Nahuatl story of a hall at the Toltec city of Tula, the walls of which were covered with fine mosaic, may belong to legend, but a monumental use of the technique was achieved in the mosaic-like treatment of the exterior wall casing of certain buildings. Those at Mitla in the state of Oaxaca are outstanding; bands and panels of simple but striking geometric ornamentation were produced by fitting together small stones of different shapes and sizes, tenoned back into the rubble mass of the wall. Each stone was cut for the spot it occupied, and some were more deeply imbedded than others so that the designs stand out in sharp relief. The effective simplicity of design and precision of workmanship at Mitla are not matched on the elaborate Maya facades—at Uxmal and Chichén-Itzá in Yucatán, for example—where, along with geometrical designs, animal forms also occur. (F.O.Wa.)

BIBLIOGRAPHY

General works: A. BLANCHET, *La Mosaique* (1928); E.W. ANTHONY, *A History of Mosaics* (1935, reprinted 1968); P. FISCHER, *Das Mosaik* (1969). H.P. L'ORANGE and P.J. NORDHAGEN, *Mosaikk* (1958; Eng. trans., 1966), deals with the history of the art from its beginnings to *c.* AD 900.

Greek and Roman mosaic: Of the literature on Greek and Roman mosaic, the following works are of particular importance: (*pebble mosaic*): R.S. YOUNG, "Gordion 1956: Preliminary Report," *Am. J. Archaeology*, 61:319–331 (1957); D.M. ROBINSON, *Excavations at Olynthus*, vol. 5, 8, and 13 (1933, 1938, 1950). (*pebble-tessera mosaic*): C.M. ROBERTSON, "Greek Mosaics," *J. Hellenic Stud.*, 85:72–89 (1965); B.R. BROWN, *Ptolemaic Paintings and Mosaics and the Alexandrian Style* (1957); J. CHAMONARD, *Les Mosaïques de la maison des masques* (1933). (*floor mosaic—Roman*): *Colloque sur la mosaïque gréco-romaine, Actes* (1965); M.E. BLAKE, articles on Roman mosaics in *Mem. Am. Acad. Rome*, 8:7–160, 13: 69–214, and 17:81–130 (1930, 1936, 1940); K.M. PHILLIPS, "Subject and Technique in Hellenistic-Roman Mosaics: A Ganymede Mosaic from Sicily," *Art Bull.*, 42:243–262 (1960); R.P. HINKS, *Catalogue of the Greek, Etruscan and Roman Paintings and Mosaics in the British Museum* (1933); D. LEVI, *Antioch Mosaic Pavements* (1947); I. LAVIN, "The Hunting Mosaics of Antioch and Their Sources," *Dumbarton Oaks Papers*, no. 17 (1963); H. STERN, "Origine et débuts de la mosaïque murale," *Études d'archéologie classique*, vol. 2 (1959).

Early Christian mosaic: M. VAN BERCHEM and E. CLOUZOT, *Mosaïques chrétiennes du IV^e^ au X^e^ siècle* (1924); G. MATTHIAE, *Mosaici medioevali delle chiese di Roma* (1967); W.F. OAKESHOTT, *The Mosaics of Rome, from the Third to the Fourteenth Centuries* (1967); H. STERN, "Les Mosaïques de l'Église de Sainte-Constance à Rome," *Dumbarton Oaks Papers*, no. 12 (1958); C. CECCHELLI, *I mosaici della basilica di S. Maria Maggiore* (1956); C. RICCI, *Tavole storiche dei mosaici di Ravenna*, 4 vol. (1930–37); F.W. DEICHMANN, *Frühchristliche Bauten und Mosaiken von Ravenna* (1958); G. BRUSIN and P.L. ZOVATTO, *Monumenti paleocristiani di Aquileia e di Grado* (1957); J.L. MAIER, *Le Baptistère de Naples et ses mosaïques* (1964); B. BRENK, "Die ersten Goldmosaiken der christlichen Kunst," *Palette* (Basel), 38:16–25 (1971), a work concerning the earliest use of gold cubes in mosaics.

Early Byzantine mosaic: H. TORP, *Mosaikkene i St. Georgrotunden i Thessaloniki* (1963): G. BRETT *et al.*, *The Great Palace of the Byzantine Emperors* (1947); C. DIEHL, M. LE TOURNEAU, and H. SALADIN, *Les Monuments chrétiens de Salonique* (1918); T. SCHMIT, *Die Koimesis-Kirche von Nikaia* 1927); P.A. UNDERWOOD, "The Evidence of Restorations in the Sanctuary Mosaics of the Church of the Dormition at Nicaea," *Dumbarton Oaks Papers*, no. 13 (1959); M. AVI-YONAH and M. SCHAPIRO, *Israel: Ancient Mosaics* (UNESCO 1954–60); E. KITZINGER, *Israeli Mosaics of the Byzantine Period* (1965); K. WEITZMANN, "The Mosaic in St. Catherine's Monastery on Mount Sinai," *Am. Phil. Soc. Proc.*, 110:392–405 (1966); K.A.C. CRESWELL, *Early Muslim Architecture*, 2nd ed., vol. 1 (1969), including M. VAN BERCHEM, "The Mosaics of the Dome of the Rock in Jerusalem and of the Great Mosque in Damascus"; G. BOVINI, *Mosaici di Ravenna* (1956; Eng. trans., *Ravenna Mosaics*, 1956), and *La vita di Cristo nei mosaici di S. Apollinare Nuovo di Ravenna* (1959); P.J. NORDHAGEN, "The Mosaics of John VII (705–707 A.D.)," *Acta, Institutum Romanum Norvegiae*, 2:121–166 (1965).

Middle and late Byzantine: O. DEMUS, *Byzantine Mosaic Decoration* (1948); A. GRABAR, *La Peinture byzantine* (1953; Eng. trans., *Byzantine Painting*, 1953); T. WHITTEMORE, *The Mosaics of Haghia Sophia at Istanbul*, 4 vol. (1933–52); P.A. UNDERWOOD and E.J.W. HAWKINS, "The Mosaics of Hagia Sophia at Istanbul: The Portrait of the Emperor Alexander," *Dumbarton Oaks Papers*, no. 15 (1961); E.J.W. HAWKINS, "Further Observations on the Narthex Mosaic in St. Sophia at Istanbul," *ibid.* 22 (1968); C.A. MANGO "Materials for the Study of the Mosaics of St. Sophia at Istanbul," *ibid.* 8 (1962); P.A. UNDERWOOD, *The Kariye Djami*, 3 vol. (1967); E. DIEZ and O. DEMUS, *Byzantine Mosaics in Greece, Hosios Lucas and Daphni* (1931); A. GRABAR and M. CHATZIDAKIS, *Greece: Byzantine Mosaics* (UNESCO 1960); V.N. LAZAREV, *Old Russian Murals and Mosaics: From the 11th to 16th Century* (1966).

Medieval mosaics in the West: W.F. OAKESHOTT, *The Mosaics of Rome, from the Third to the Fourteenth Centuries* (1967); O. DEMUS, *The Mosaics of Norman Sicily* (1950); E. KITZINGER, "The Mosaics of the Cappella Palatina in Palermo," *Art Bull.*, 31:269–292 (1949); *The Mosaics of Monreale* (1960); A. DE WITT (ed.), *I mosaici del Battistero di Firenze*, 4 vol. (1954–59); H. KIER, *Der mittelalterliche Schmuckfussboden* (1970); A. PRANDI, "Pietro Cavallini a S. Maria in Trastevere," *Rivista dell'Istituto Nazionale d'Archeologia e Storia dell'Arte*, 1: 282–297 (1952); R. OERTEL, "Wandmalerei und Zeichnung in Italien," *Mitteilungen des Kunsthistorischen Institutes in Florenz*, 5:217–314 (1937–40), on the technique; F. FORLATI, "La tecnica dei primi mosaici marciani," *Arte veneta*, 3:85–87 (1949).

Renaissance to modern: The best surveys are those contained in E.W. ANTHONY, *A History of Mosaics* (1935, reprinted 1968); and P. FISCHER, *Das Mosaik* (1969).

(P.J.N.)

Moscow

One of the world's great cities, Moscow (Moskva in Russian) is the capital of the Soviet Union and also of the Russian Soviet Federated Socialist Republic, its largest constituent unit. As capital of the largest and one of the most powerful countries in the world, Moscow is a city of outstanding international significance. Since it was first mentioned in chronicles dating from 1147, Moscow has played a vital role in Russian history; indeed, the history of the city and of the state are closely interlinked. For more than 600 years Moscow has also been the spiritual centre of the Orthodox Church of Russia. Today it is not only the political focus of the Soviet Union but the country's leading city in population size, in industrial output, and in cultural, scientific, and educational importance, and the name of the Kremlin, the seat of government in Moscow, has become a synonym worldwide for Soviet authority.

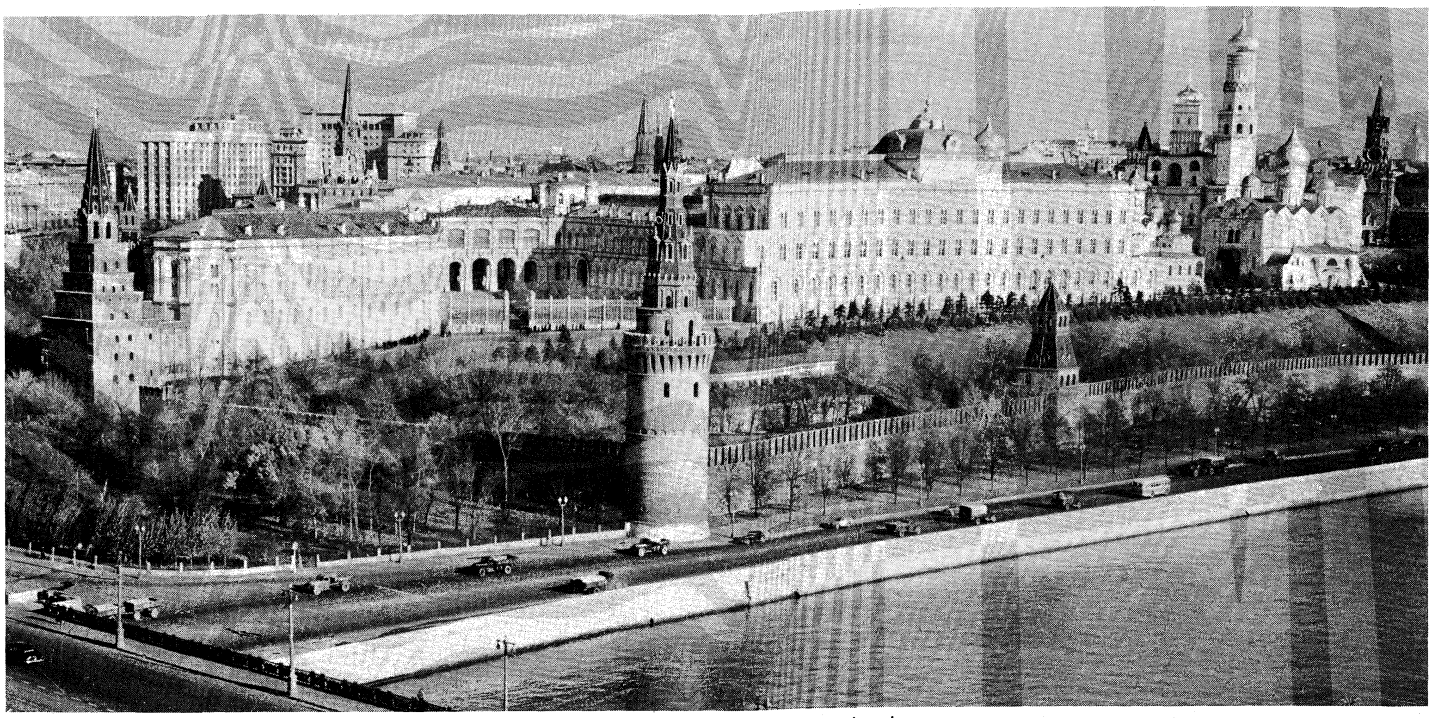
Greater Moscow

Greater Moscow, formed in 1960, covers an area of 339 square miles (879 square kilometres), its outer limit being delineated by the Moscow Circular Motorway. Beyond the motorway, an outer zone has been designated as a Forest-Park Zone, or green belt. Occupying 666 square miles (1,725 square kilometres), this is intended as an amenity area, with severely restricted urban development and industry. In 1976 the population of Greater Moscow was 7,734,000, of whom 7,563,000 lived in the city itself.

For information on related subjects, see RUSSIA AND THE SOVIET UNION, HISTORY OF; RUSSIAN SOVIET FEDERATED SOCIALIST REPUBLIC; and SOVIET UNION.

THE PHYSICAL ENVIRONMENT

Setting and relief. Moscow stands on the Moskva River, a tributary of the Oka and thus of the Volga, in the centre of the vast plain of European Russia. This focal region, as a consequence of the developments of the last 100 years, is known as the Central Industrial Region and is one of the most highly developed and densely populated parts of the Soviet Union. The city lies in the broad, extremely shallow valley of the Moskva and its tributaries. The valley itself overlies a deep downfold of the Russian Platform of very ancient crystalline rocks, a granitic basement some 5,740 feet (1,750 metres) below the present-day surface. The trough in the platform has been



The Kremlin in Moscow, on the Moskva River, with the Water Tower in the foreground and (from the left) the Borovitskaya Gate, the Armoury Palace, the Kremlin Great Palace, and the Archangel Cathedral.
Novosti Press Agency

filled through long geological time by sedimentary rocks, mostly of Devonian and Carboniferous age, with Jurassic series and some Cretaceous rocks nearest the surface. Hardly anywhere is the bedrock exposed on the surface.

The advances of the glaciers in the Ice Age deposited a thick mantle of boulder clays and morainic sands and gravels into which the sinuous Moskva River cut its wide valley in successive stages, marked by four corresponding levels. Recent alluvial deposits cover their surfaces. Beside the river itself is the narrow belt of floodplain; a few feet above this is the first terrace, which yields to the successively higher second and third terrace levels. The last of these, rising up to 100–115 feet (30–35 metres) above the river, is the most extensive, and much of Moscow is built on it. Northward the third terrace merges imperceptibly with a morainic plain of clays and sands, which slopes up very gradually to the Klin-Dmitrov morainic ridge some 40 miles (64 kilometres) north of the city. Eastward and southeastward the surface equally gradually merges into the vast, almost completely flat and very swampy clay plain of the Meshchera Lowland, which extends far beyond the city limits.

Almost everywhere surface relief is minor; only in the southwest of the city is there an upland area on Cretaceous rocks, covered by glacial morainic material. This is the Teplostanskaya Upland, which rises more than 400 feet above the Moskva River and on which is the highest point within Moscow's limits, 830 feet (253 metres) above sea level. One of the sweeping bends of the Moskva has cut into the edge of the Teplostanskaya Upland a steep cliff, the Lenin Hills (formerly the Sparrow Hills), from the top of which there are fine panoramic views of the city. Today the Lenin Hills are crowned by the buildings of Moscow State University.

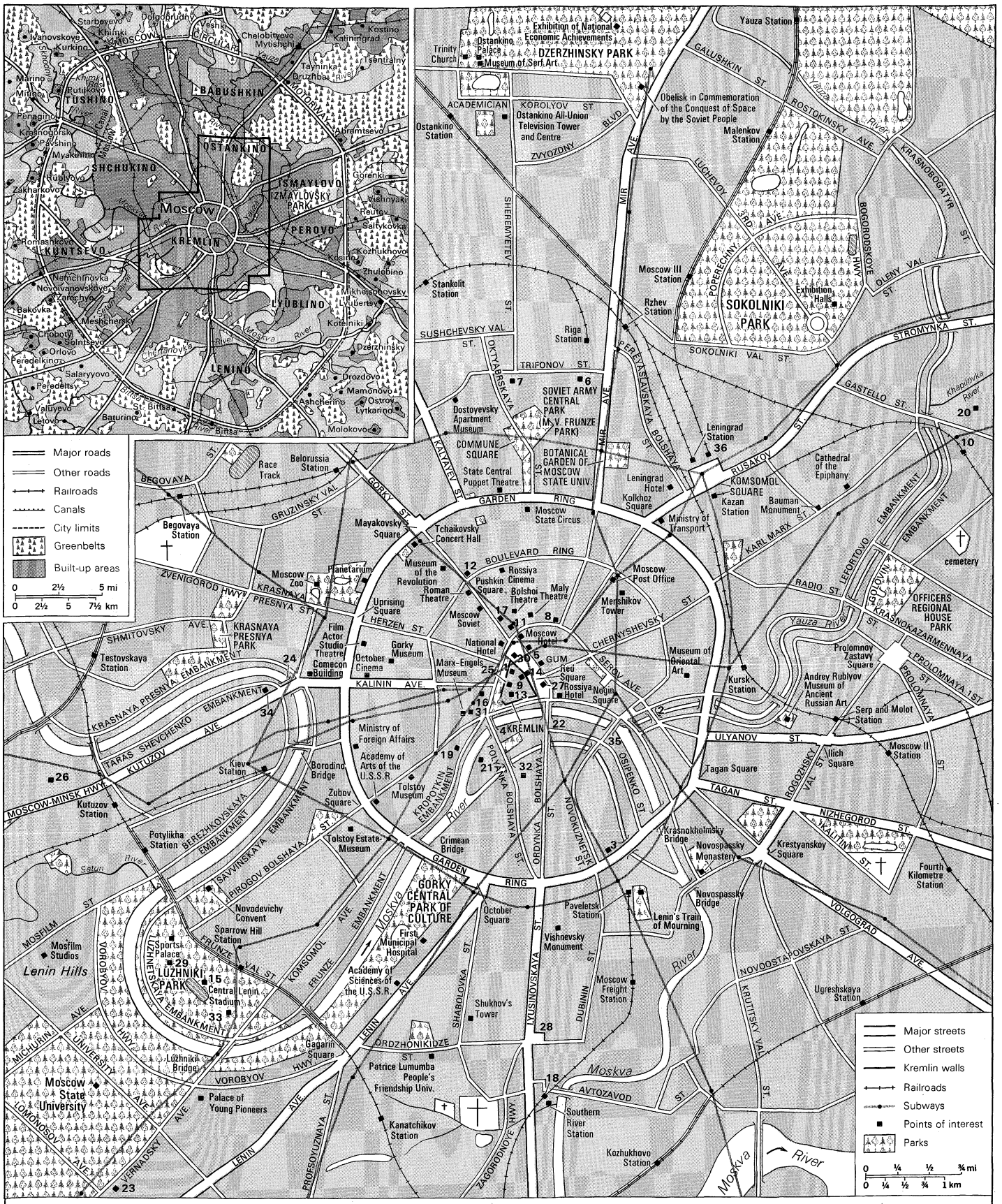
The human imprint. Long occupation by man has extensively altered the natural setting. The "cultural layer," consisting of debris of buildings demolished long ago and of other materials deposited by man, is up to 50 feet deep in some parts of central Moscow. Almost all the small rivers and streams that once flowed into the Moskva through the city area have now been put into underground conduits or filled in. The only tributaries still visible are the Yauza on the northern (left) bank and, on the southern bank, the Gorodnya, Kotlovka, and Setun, with its tributary the Ramenka. The Yauza, largest of these, and the Moskva itself are controlled by stone embankments for most of their winding courses through the city. The Moskva has been diverted in places, with cuts made through the necks of its loops, and it has also been both widened and deepened; in places it is 800 feet wide. The Yauza now receives additional water from the Volga, by way of the Moscow Canal and its branch, the Likhobory Canal. Two dams on its lower course have raised the level of the Yauza and made the lower reaches navigable.

Rivers

Some of Moscow's water supply comes from some 1,000 deep bores in the city, tapping the artesian water of the underlying Carboniferous beds. Overuse has greatly lowered the levels of these underground waters (90–115 feet during 1900–58), and today most water needs are met by surface sources, from the reservoirs north of the city built in connection with the Moscow Canal, in particular the Ucha Reservoir. Water is also drawn from the Moskva River and pumped into underground storage reservoirs. Up to the early 1960s, the discharge of untreated sewage and industrial effluent had brought about extremely severe pollution of the Moskva, together with increasing pollution of underground waters. So serious was the problem that in the mid-1960s a major effort was made to correct it by heavy investment in anti-pollution and water purification measures. Although the work was not complete by the mid-1970s, a great improvement was discernible. The oxygen content of the Moskva had doubled, fish reappeared in its waters, and bathing was permitted at the foot of the Lenin Hills. Slow but steady progress is being made in controlling the discharge of industrial effluent into the river and into bore wells, and the discharge of untreated sewage will end.

Water pollution

Climate. The climate of Moscow is of continental type, modified by the temperate influence of westerly winds of Atlantic origin. Precipitation is moderate, 22.9 inches (582 millimetres) a year, with a marked summer maximum that reaches a peak in July. A considerable part of the precipitation falls as snow, usually beginning about mid-October though permanent snow cover is not normally established before the beginning of November; it lasts on average 146 days, until mid-April. Winters are thus long, dark, and gloomy, although southerly airstreams not uncommonly bring thaw days, with temperatures above freezing. Conversely, northerly airstreams



1 Arsenal
2 Astakhov Bridge
3 Bakhrushin Theatre Museum
4 Bolshoy Kamenny Bridge
5 Central Lenin Museum
6 Central Museum of the Soviet Army
7 Central Theatre of the Soviet Army
8 Children's World
9 Council of Ministers
10 Elektrozavodskaya Station
11 House of Trade Unions
12 Izvestiya Offices
13 Kremlin Great Palace
14 Lenin Mausoleum
15 Lenin Monument
16 Lenin State Library
17 Moscow Art Theatre
18 Moscow Southern River Station
19 Moscow Swimming Pool
20 Moscow Television Theatre
21 Moscow Variety Theatre
22 Moskva River Bridge
23 New Moscow State Circus
24 Novoarbatsky Bridge
25 Palace of Congresses
26 Panorama Museum of Borodino Battle
27 Pokrovsky Cathedral
28 Serpukhovskoy Zastavy Square
29 Sports Palace Arena
30 State Historical Museum
31 State Pushkin Museum of Fine Arts
32 State Tretyakov Gallery
33 Summer Swimming Stadium
34 Ukraina Hotel
35 Ustyunsky Bridge
36 Yaroslavl Station

Central Moscow and (inset) its metropolitan area.

from the Arctic bring very sharp drops in temperature, often accompanied by clear, brisk weather with a low relative humidity. December, January, and February are in fact the driest months. Thus, although the January average temperature is 13.6° F (—10.2° C), there can be considerable variation; the minimum recorded temperature is —44° F (—42° C). Moscow employs a large number of workers and a fleet of mechanical devices to keep its streets clear of snow throughout the winter. Spring is relatively brief, and the temperature rises rapidly during late April. Summers are warm, and July is the warmest month, with an average temperature of 64.6° F (18.1° C); a maximum temperature of 99° F (37° C), however, has been recorded in August. Rainy days are not uncommon, but the summer rainfall often comes in brief, heavy downpours and thunderstorms. Autumn, like spring, is short, with rapidly falling temperatures.

Urban climate

Human activity has modified the climatic conditions as well as the relief and hydrology. The vast number of buildings creates the so-called urban heat island effect, causing the average annual temperature in the central city to be up to 2.5° F (1.4° C) higher than the temperature in open country outside the city. Plants in central parks and gardens blossom a week or more earlier than those in the outer suburbs. The temperature differential is most marked on clear winter nights as the buildings radiate heat, but foggy weather also traps heat by preventing radiation.

As with the water, up to the late 1950s there was increasing pollution of Moscow's atmosphere. Smogs were common, and often heavy concentrations of sulfur dioxide occurred. Again a major campaign to control noxious emissions was launched, assisted greatly by a changeover from coal to natural gas as the principal fuel. Some seriously polluting factories were moved out of the city. The resulting improvement in the air of the city has been very marked, though further improvements are still necessary. Monthly deposits of dust and ash per square metre vary from 135 grams (0.38 ounces per square foot) in industrial sectors to as little as 15 grams (0.05 ounces per square foot) on the Lenin Hills. The growing number of motor vehicles—though they are far fewer than in comparable North American cities, and low octane gasoline is used—is increasing concentrations of carbon monoxide in certain parts of Moscow.

Red Square, with the Kremlin wall (right), the Lenin Mausoleum (centre), the Cathedral of St. Basil the Blessed, and, in the background, the Moskva River.
Sovfoto

HISTORY

Foundation and early growth. First documentary reference to Moscow is found in the early monastic chronicles under the year 1147, when on April 4 Yury Vladimirovich Dolgoruky, prince of Suzdal, was host at a "great banquet" for his ally the Prince of Novgorod-Seversky "in Moscow." This is traditionally taken as the date of Moscow's foundation, although archaeological evidence shows that a settlement had existed on the site since Neolithic times. Archaeological work has revealed the remains of corduroy roads and evidence of iron and leather working dating from the 11th century. The disturbed times made defense essential, and in 1156 Prince Dolgoruky built the first fortifications—ditches and earthen ramparts topped by a wooden wall with blockhouses. This was the Kremlin, or Citadel. The origin of the word *kremlin* is disputed; some authorities suggest Greek words for "citadel" or "steepness," others the early Russian word *krem*, meaning a conifer providing timber suitable for building. The Kremlin was sited on the relatively high spit of land between the Moskva River and a small tributary, the Neglinnaya. The triangular piece of land between the rivers was protected on the eastern side by a moat joining them. The Neglinnaya today flows through an underground conduit, but part of its course is traced by a street of the same name. This Kremlin site was the core of the young town, and it still is.

First fortifications

Moscow soon developed as one of the more important towns of the principality of Vladimir-Suzdal. A trading settlement, or *posad*, grew up to the east of the Kremlin, along the Moskva in the area known as Zaradye. Like most other Russian towns, Moscow was captured and burned by the Mongols in their great invasion of 1236–40, and its princes had to accept Mongol suzerainty. It soon recovered, though the Mongols sacked it once again in 1293. Three years later the Kremlin was strengthened with a new earthen wall and oak palisade. Thereafter Moscow grew in importance, in trading and artisan activity, and in size, overtaking the older and previously more important centres of Suzdal and Vladimir. The town was fairly centrally placed in the system of rivers and portages that formed the trade routes across European Russia. The area east of Moscow between the Oka and Volga moreover enjoyed rather better soils than most of northern Russia and formed a region of more developed agriculture and prosperous towns. Moscow's authority was greatly enhanced when in 1326 the metropolitan of the Russian Orthodox Church transferred his seat from Vladimir to Moscow. Thereafter the town was to remain the centre of Russian Orthodoxy, and after the fall of Constantinople to the Turks (1453) it claimed the title of the Third Rome. Under Ivan I Kalita (1325–41) the principality of Vladimir was incorporated into that of Moscow. Gradually the princes of Moscow extended their rule over the other surrounding Russian princedoms, and the town became the leader in the long struggle against Mongol hegemony.

Seat of the metropolitan

The struggle at first fluctuated. In 1378 a Muscovite army repulsed a Mongol attack on the Vozha River south of the town, and in 1380 Prince Dmitry of Moscow inflicted a crushing defeat on the Mongols under the great khan Mamai at Kulikovo on the Don, for which victory he was thereafter known as Dmitry Donskoy ("of the Don"). The Kremlin had been enlarged and given walls and towers of white limestone in 1367, but the new fortifications were unable to withstand a renewed Mongol attack in 1382, when, despite a heroic defense, the khan Tokhtamysh captured and plundered Moscow. Yet another attack in 1408 under Khan Yedigei, however, was beaten off. Steadily Moscow grew in size and importance, as it continued to absorb the surrounding princedoms. Within the Kremlin the first stone cathedral, the Assumption (Uspensky), was built in 1326. Palaces for the prince and leading boyars, monasteries, and churches

were erected. Outside the Kremlin walls, the trading and artisan quarter to the east grew in size and became known as the Kitay-gorod; this name, which originated in the 16th century, probably derives from the word *kita*, a binding of poles used in the fortifications before stone walls were built, and does not mean "Chinese City" as it is sometimes translated.

The rise of Moscow as capital. By the second half of the 15th century, especially after the annexation of Novgorod in 1478, Moscow had become the undisputed centre of a unified Russian state. During the reign of the grand prince of Moscow Ivan III the Great (1462–1505), the Kremlin was again enlarged and given fire-resisting brick walls more than a mile in length and in places up to 60 feet high, with 19 brick towers (1485–95). From this period also date the rebuilt Cathedral of the Assumption (Uspensky Sobor) and the equally beautiful Annunciation (Blagoveshchensky) and (also rebuilt) Archangel cathedrals, the Palace of Facets (Granovitaya Palata), and the soaring white bell tower of Ivan III the Great. In 1534–38 the Kitay-gorod, previously protected only by earth banks and palisades, was also surrounded by a brick wall, with 12 towers. The town continued to grow and spread outside the walls to form what became known as the Bely-gorod (White City) in a semicircle around the Kremlin and Kitay-gorod.

Despite its new fortifications, Moscow remained subject to disaster and attack. In 1547 two great fires destroyed the greater part of the town. In the mid-16th century Ivan IV the Terrible conquered the Tatar (Mongol) khanates of Kazan (1552) and Astrakhan (1556), but the Crimean Tatars succeeded once more in 1571 in capturing Moscow and burned it, except for the Kremlin. The annals record that only 30,000 of 200,000 inhabitants survived. A further attack was launched by the Crimean Tatars in 1591, but they failed to overcome Moscow's stubborn resistance. The defense was helped by the new walls, built between 1584 and 1591 to protect the Bely-gorod, made of stone and some five miles long, with 29 towers and 11 gates. The line of these fortifications is marked today by the broad, tree-lined Boulevard Ring (Bulvarny koltso) of streets. In 1592 an outer earth rampart with 50 towers was thrown up all round the city, including an area on the southern (right) bank of the Moskva. This encompassed a further extension of Moscow that had grown up beyond the Bely-gorod; known at first as Skorodom, this outer sector came to be called the Zemlyanoy-gorod, or Earthen City. The line of its fortifications is also traced today by a ring of very wide avenues, the Garden Ring (Sadovoye koltso). As an outermost line of defense, a chain of strongly fortified monasteries was established beyond the ramparts to the south and east, principally the Novodevichy Convent and Donskoy (Don), Danilovsky, Simonov, Novospassky, and Andronikov monasteries, most of which survive today as museums.

Zemlyanoy-gorod

With much improved security, trade and craft manufacture flourished. Distinct quarters were occupied by particular trades; for example, the suburbs of Bronnaya by armour makers, Kuznetskaya by blacksmiths, and Kotelniki by kettle makers. Across the Moskva was the weavers' suburb. These artisan sectors are commonly commemorated today by street or quarter names. State workshops cast cannon and made weapons and gunpowder. The tsar's court and its attendant nobility provided patronage for luxury crafts. Increasingly the boyars took over the Kitay-gorod, with artisans and traders moving to the outer parts; the Kremlin became solely the seat of temporal and ecclesiastical authority. The centre of commercial activity was the market in Red Square between the Kremlin and the Kitay-gorod, where were rows of stalls, each handling a specific variety of goods; the name Red Square derives from the old Russian word for "beautiful." Trade with western Europe (especially England and Holland), as well as with Central Asia, Transcaucasia, Persia, and the Black Sea coast, was brisk, furs forming a major staple in this international commerce. Foreign merchants lived in the Nemetskaya (German) suburb. Many churches were built in the Kitay-gorod and Bely-gorod. Flourishing cultural life was marked by the growth of the book trade and the founding in 1553 of the first printing house.

The opening of the 17th century saw Moscow, like the rest of Russia, suffer severely during the Time of Troubles. In the reign of Boris Godunov there were severe famines from 1601 to 1603. After Boris' death in 1605, the False Dmitry seized Moscow with Polish help, and though he was killed in 1606 and the Poles were driven out, they reoccupied Moscow with a second False Dmitry in 1608–10. In May 1611 the Muscovites attacked the Poles, and the invaders retreated into the Kremlin. Under the energetic leadership of a boyar, Prince Dmitry Mikhaylovich Pozharsky, and a merchant, Kuzma Minin, the Russians forced the Poles to surrender in October 1612.

With the establishment in 1613 of the Romanov dynasty under Michael, relative peace returned to Moscow and with it further economic advance. Nevertheless the conditions of the poor of the town often led to riots and uprisings; such events were nothing new, having occurred also in 1382, 1445, and 1547. In 1648, as a result of an increase in the salt tax, and again in 1662 (the so-called Copper Riots) there were disturbances by artisans, labourers, and tradesmen. The great revolt of Stenka Razin in southern Russia (1667–71) was echoed by unrest in the capital, and in 1671 Razin was executed in Moscow as a warning to its inhabitants. The revolts were put down by the hereditary militia, the *streltsy*, who in 1698, early in the reign of Peter I the Great, themselves revolted and were suppressed only with great slaughter.

Riots and uprisings

Despite the frequent upheavals, however, culture could flourish. Russia's first higher educational institution, the Slavonic-Greek-Latin Academy attached to the Zaikonospassky Monastery in the Kitay-gorod, dates from 1687. In 1701 Peter founded a School of Mathematics and Navigation. The first newspaper in Russia began publication in Moscow in 1703.

The 18th and 19th centuries. The year 1703 saw Peter I the Great begin construction of St. Petersburg (now Leningrad) on the Gulf of Finland, and in 1712 he transferred the capital to his new, "westernized," and outward-looking city. Members of the nobility were compelled to move to St. Petersburg; many merchants and artisans also moved. Both population growth and new building in Moscow languished for a time, but even during Peter's reign the city began to recover from the loss of capital status. Peter himself stimulated economic growth by establishing new factory industries, and private entrepreneurs followed suit. By 1725 there were 32 new factories employing 5,500 workers; 23 of the factories were textile mills, including a crown enterprise making sailcloth. At the same period there were about 8,500 craft workers.

During the 18th century Moscow retained its major role in the cultural life of Russia. In 1755, on the initiative of the great man of letters and science M.V. Lomonosov, Moscow University (now Moscow M.V. Lomonosov State University) was founded, the first university in Russia; a medical and surgical college was opened in 1786. Although serious fires did much damage in 1737, 1748, and 1752, many splendid new buildings appeared, designed by such architects as Giacomo Quarenghi, Vasily Bazhenov, Matvey Kazakov, and Vasily Stasov. Among many outstanding buildings were the beautiful Pashkov House (now part of the Lenin State Library), the old university buildings, the former Senate in the Kremlin, and the former assembly of nobles (now the House of Trade Unions) famous for its Hall of Columns. In 1741 Moscow was surrounded by a barricade 25 miles (40 kilometres) long, the Kamer-Kollezhsky barrier, at whose 16 gates customs tolls were collected; its line is traced today by a number of streets called *val* (rampart, or wall) and by place-names such as Kaluga Zastava (Customs Gate). Industry flourished, and by the end of the 18th century there were about 300 factories in Moscow, more than half of them textile mills. The population grew to 275,000 in 1811.

Moscow University

In 1812 Napoleon invaded Russia; after a bitter 15-hour battle on September 7 at Borodino on the approaches to Moscow, the Russian commander in chief, Gen. M.I.

Kutuzov, evacuated troops and civilians from the city, which was occupied by the French a week later. A fire broke out and spread rapidly, eventually destroying more than two-thirds of all the buildings. Looting was rife. The lack of supplies and shelter and continual harassment by Russian skirmishing forces made it impossible for Napoleon to winter in Moscow, however, and on October 19 the French troops began their catastrophic retreat.

Rebuilding

In 1813 a Commission for the Construction of the City of Moscow was set up, under the direction of the architect Osip Bove. It launched a great program of rebuilding, which included a partial replanning of the city centre. Among many buildings constructed or reconstructed at this time were the Kremlin Great and Armoury (Oruzheynaya) palaces in the Kremlin, the university, the Manezh (Riding School), and the Bolshoi Theatre. Industry also recovered rapidly and continued to develop through the 19th century. In 1837 the Moscow stock exchange was established. The emancipation of serfs in 1861 and the beginning of the railway era with the opening of the line to St. Petersburg in 1851 greatly increased labour mobility, and large numbers of peasants from the villages began moving into Moscow. The population, which had reached 336,000 in 1835, had almost doubled, to 602,000, in 1871 and by 1897 had reached 978,000. Moscow became the hub of Russia's railways, with trunk lines to all parts of European Russia. A ring of main line termini was built, mostly on or near the Kamer-Kollezhsky barrier at the limits of the built-up area. Outside the barrier many new factories, particularly those concerned with textiles, began operation. In the 1890s heavy engineering and metalworking industries also developed. Between 1897 and 1915 Moscow yet again doubled in size, to a population of 1,983,700.

The later 19th century was a period of ostentatious building by public bodies and wealthy private persons, in various imitative "old Russian" styles and the so-called modern style. From this period date the old Town Hall (meeting place of the Gorodskaya Duma, now the Central Lenin Museum), the State Historical Museum, and the Upper Trading Rows (now the State Department Store, GUM).

The growth of an industrial proletariat in Moscow, together with the generally low living standards of the workers, brought growing unrest and strikes. Various revolutionary groups were active. In the Revolution of 1905 a small-scale insurrection took place in Moscow, and an attempt was made to seize the Nikolayev (now Leningrad) station; the revolt was ruthlessly crushed. In 1917, although a Council of Workers' and Soldiers' Deputies was set up in Moscow, the city remained relatively quiet until after the Bolshevik seizure of power in Petrograd (formerly St. Petersburg) on October 25 (November 7, new style), which was immediately followed by fighting in Moscow. Military cadets held out for a time in the Kremlin, but by November 3 (16, N.S.) they were overcome and Soviet power was firmly established.

The Bolshevik Revolution

Moscow in the Soviet period. In March 1918 Lenin and the Soviet government moved to Moscow, which thereby resumed its former status as capital. This status was formally ratified on December 30, 1922, when the first All-Union Congress of Soviets met in the Bolshoi Theatre and passed the legislation setting up the Union of Soviet Socialist Republics. In the civil war period Moscow, like other Soviet cities, suffered greatly, with grave food shortages, loss of population (falling to 1,027,300 in 1920), and reduction of industry. In the years following the final establishment of Soviet power and peace, recovery was swift, and the beginning of the series of five-year plans in 1928 brought great industrial progress. The city's existing plant and labour force formed one of the main springboards for industrialization elsewhere in the Soviet Union. Between the censuses of 1926 and 1939, Moscow once more doubled its population, from 2,029,425 to 4,182,916 (within the 1959 boundaries). Priorities in investment went to industry, and housing construction was very limited; in consequence, exceptional overcrowding of existing housing developed, with extremely high population densities. Living standards in all respects were low.

During World War II, the Germans in late 1941 reached the outskirts of Moscow, less than 25 miles from the Kremlin. Many factories were evacuated, together with much of the government. From October 20 the city was declared to be in a state of siege. Its remaining inhabitants built and manned anti-tank defenses, while the city was bombarded from the air. A desperate counterattack on December 6 threw the German forces back from the outskirts and saved Moscow. The end of the war once more saw a rapid recovery and further growth of the city's economy. By the mid-1950s the housing shortage had reached grave proportions. Within the Garden Ring, population densities in 1959 surpassed 132,000 persons per square mile. Under N.S. Khrushchev a major program of construction was begun. Although some old housing, sometimes single-storied and made of wood, still remains, much has been cleared in renewal schemes, and extensive new suburbs of large apartment buildings have sprung up all round the historic core of the city. In central areas much urban renewal has taken place, and high-rise buildings now dominate the skyline. In 1960 Greater Moscow was created, incorporating into the city a group of adjacent satellite towns. There have been several development plans for Moscow, the first dating from 1935; the General Plan for the Development of Moscow was adopted in 1971. In accordance with Soviet practice of making awards to cities, Moscow has been given the title of Hero City for its part in World War II, two Orders of Lenin, and the Order of the October Revolution.

LAYOUT AND ARCHITECTURE

The pattern of rings and radials that marked the historical stages of Moscow's growth remains evident in its layout today. Successive epochs of Moscow's development are traced by the Boulevard Ring and the Garden Ring (both following the line of former fortifications), the Greater Moscow Ring Railway, and the Moscow Circular Motorway, which is the administrative limit of Greater Moscow. Each concentric zone carries its own particular stamp of function and architectural style.

The Kremlin. The hub of the layout, and today as throughout its history the heart of Moscow, is the fortified enclosure of the Kremlin, the symbol of first Russian and later Soviet power and authority. Its present crenellated red brick walls and 19 towers were built at the end of the 15th century. Of the most important towers, the Spasskaya (Saviour's) Gate Tower leading to Red Square was built in 1491 by Pietro Solario, who designed most of the main towers; its belfry was added in 1624–25 by an Englishman, Christopher Galloway. The chimes of its clock, dating from 1852, are broadcast by radio as a time signal to the whole Soviet Union. Also on the Red Square front is the Nikolskaya Tower, built originally in 1491 and rebuilt in 1780. The two other principal gate towers lie on the western wall, the Troitskaya (Trinity) Gate, with a bridge and outer barbican (the Kutafya Tower), and the Borovitskaya Gate.

Kremlin gate towers

Within the walls is one of the most striking and beautiful architectural ensembles in the world. Around the central Cathedral Square are grouped three magnificent cathedrals, superb examples of Russian church architecture at its height in the late 15th and early 16th centuries. The white stone Cathedral of the Assumption (Uspensky Sobor) is the oldest, built in 1475–79 by the Italian Aristotele Fioravanti in the Italianate-Byzantine style and modelled on the Cathedral of the Assumption in Vladimir. Its pure, simple, and beautifully proportioned lines and elegant arches are crowned by five golden domes. The interior frescoes and paintings of the iconostasis date from the 17th century. The Orthodox metropolitans and patriarchs of the 14th to 18th century are buried here. Across the square is the Cathedral of the Annunciation (Blagoveshchensky Sobor) built in 1484–89 by craftsmen from Pskov; although burned in 1547, it was rebuilt in 1562–64. Its cluster of chapels is topped by golden roofs and domes. Inside are a number of early 15th-century icons attributed to Theophanes the Greek and to Andrey Rublyov, the greatest of all Russian icon painters. The third cathedral, the Archangel (Arkhangel-

sky), was rebuilt by Alevisio Novyi in 1505–09; in it are buried the tsars of Russia up to the founding of St. Petersburg.

Just off the square stands the splendid, soaring, white bell tower of Ivan III the Great; built in the 16th century and destroyed in 1812, it was rebuilt a few years later. At its foot is the 200-metric ton Tsar Bell (Tsar Kolokol), cast in 1733–35 but never rung; in the fire of 1737 it fell from its scaffolding and a fragment weighing 11.5 tons broke off. It is 20 feet (6.1 metres) high and 22 feet (6.6 metres) in diameter. Nearby is the Tsar Cannon (Tsar Pushka), cast in 1737 but never fired for fear of explosion; intended to project one-ton cannonballs, it has a calibre of 35 inches (89 centimetres) and weighs 40 tons. Beside the gun is the Cathedral of the Twelve Apostles (Dvenadtsati apostolov), built in 1653–56, and the adjoining Patriarch Palace.

On the west of Cathedral Square is a group of palaces of various periods; the Palace of Facets (Granovitaya Palata)—so called from the exterior finish of facetted, white stone squares—was built in 1487–91. Beside it is the Terem Palace of 1635–36, which incorporates the Church of the Nativity of Our Lady, dating from 1393.

Kremlin Great Palace

Both abut on the Kremlin Great Palace, built as a royal residence in 1838–49 by Konstantin Thon and today used for sessions of the Supreme Soviet of the U.S.S.R.; its long, yellow-washed facade dominates the river front. It is connected to the Armoury (Oruzheynaya) Palace, built by the same architect in 1844–51 and now the Armoury Museum, housing a spectacular collection of treasures of the tsars. Along the northeast wall of the Kremlin is another group of buildings, including the Arsenal (1702–36) and the former Senate building (1776–88), designed by M.F. Kazakov and now housing meetings of the Council of Ministers. The one modern building within the Kremlin is the Palace of Congresses (1960–61), with a vast auditorium, seating 6,000, used for Communist Party congresses and as a theatre.

The Kitay-gorod. Along the east wall of the Kremlin lies Red Square, the ceremonial centre of the capital and scene of the May Day and October Revolution parades. Beneath the walls stands the squat bulk of the Lenin Mausoleum. At the southern end of Red Square is the Intercession (Pokrovsky) Cathedral, or Cathedral of St. Basil the Blessed. Built by the architects Barma and Posnik Yakovlev in 1554–60 to commemorate the defeat of the Tatars (Mongols) of Kazan and Astrakhan by Ivan IV the Terrible, it is a unique and magnificent architectural fantasy, each of its 10 domes differing in design and colour. Along Red Square facing the Kremlin is the State Department Store, GUM, in the revivalist imitation "old Russian" style; in the same style is the slightly earlier State Historical Museum (1875–83) closing off the northern end of the square.

Many old churches survive along the narrow streets of the Kitay-gorod. Of particular note are the Church of the Trinity (1628–34) in Nikitniki, the 15th-century Church of St. Anne of the Conception, and the cathedral (1812) of the former Epiphany Monastery (1693–96) in the so-called Moscow Baroque style. Much of the Kitay-gorod near the river was demolished in the 1960s to make room for the huge square block of the Rossiya Hotel, with its accommodations for 6,000 guests. Along the northern front of the hotel is a row of preserved ancient buildings, including the Church of St. Barbara (1794) and the 16th-century house of the Romanov boyars. East of the hotel is a surviving section of the Kitay-gorod wall; another section with a tower can still be seen in the northeast corner of the Kitay-gorod, near Marx Prospekt.

The inner city. In the remainder of the central part of Moscow, within the Garden Ring, are buildings representative of every period of Moscow's development from the 15th century to the present day. Scattered through the inner city are several fine examples of 17th-century church architecture, notably the Church of All Saints of Kulishki, built in the 1670s and '80s to replace a 14th-century edifice, and the Church of the Nativity of Putniki (1649–52). This was the period of development of the Moscow Baroque style, well represented in the Church of the Trinity in the former Sheremetyev estate, which dates from the end of the 18th century. Buildings of the Classical period, beginning about the middle of the 18th century and covering the rebuilding of Moscow after the fire of 1812, abound within the Garden Ring. Notable examples are the old university and the former assembly of nobles with its Hall of Columns (now the House of Trade Unions), both built by Kazakov in the 1780s; Bazhenov's elegant Pashkov House (1785–86); the former Merchants' Court (Gostiny Dvor), designed by Quarenghi and built in 1790–1805; the State Pushkin Museum of Fine Arts (1814); the Manezh (1817), which is now used as an exhibition hall; and the magnificent Bolshoi Theatre (1821–24), rebuilt by A. Cavos in 1856 after a fire. Toward the end of the 19th century and continuing into the early 20th century, buildings in the revivalist "Old Russian" style, often more quaint than beautiful, were built, including the State Department Store and State Historical Museum on Red Square, the State Tretyakov Gallery (1906), and the former Morozov mansion, now the House of Friendship (1894).

In the Soviet period much of the old, dilapidated part of the inner city has been swept away and replaced by large office and apartment buildings. Side by side with the old appeared new buildings in the modern, functional style of the 1920s and '30s, in the ponderous, often over-ornate style of the later Stalin period, and in the high-rise concrete and glass of the 1960s and '70s. Cheek by jowl with the Pashkov House stands the severe main building of the Lenin State Library (1927–29). Among more interesting and imaginative examples of later architecture is the Rossiya Cinema (1960–61) in Pushkin Square.

In general, however, the Soviet period has added little of outstanding architectural merit to the central areas of Moscow, but it has created far more space, with wide, new squares such as the Square of the 50th Anniversary of October (formerly Manezhnaya Square) permitting the older buildings to be seen far more effectively. Many streets have been widened, in particular Gorky Street, now one of Moscow's principal radial roads, lined with large shops, hotels, and offices. The Garden Ring itself has been widened to form a broad inner ring motorway, with up to four lanes in each direction and with overpasses where it is intersected by the main radial routes. In the 1960s a new radial street, Kalinin Avenue, was driven through an area of poorer housing westward from the Kremlin to the Moskva River. The showpiece of the modern city, it is lined by high-rise office and apartment buildings, linked at street and second-floor levels by a shopping mall. At its outer end rises the lofty, three-winged Comecon (Council for Mutual Economic Assistance) building overlooking the river. Beside it a new Council of Ministers building was under construction in the mid-1970s. Although most of the historic buildings of central Moscow have been preserved, and the 1960s and '70s in particular have seen much careful restoration and repair, some architectural monuments disappeared in the early Soviet period. Where before the Revolution the Cathedral of the Saviour stood, near the river just above the Kremlin, there is now a vast open-air swimming pool.

Kalinin Avenue

Inner Moscow has the functions typical of a central business district. Here are concentrated most of the government offices and central administrations of state bodies, most of the hotels and larger shops, and the principal theatres, museums, and art galleries. But the function of residence has not been squeezed out; the many very large apartment buildings of modern times have ensured a large population continuing to live in the centre. Moscow has no dead heart, empty of people at night. Also, a number of industries remain within the Garden Ring.

The middle zone. Beyond the Garden Ring and approximately as far as the Ring Railway lies a zone mostly of late 18th- and 19th-century development. Within it are many factories and the principal railway stations and freight yards. The Likhachyov automobile works, with its associated housing, occupies much of the southeast sector. Enveloped within this zone are further examples of the

Kalinin Avenue in Moscow, with administrative offices (left), apartment houses (right), and the Ukraina Hotel in the background.
V. Yegorov—Tass from Sovfoto

best of Classical Moscow, such as the 18th-century palace that now houses the Presidium of the Academy of Sciences of the U.S.S.R. on Lenin Avenue. Also to the southwest, on the banks of the Moskva, are the most important of the fortified monasteries, the 16th-century Novodevichy Convent, with its beautiful Smolensk Cathedral, whose tall bell tower (1690) dominates the churches and buildings within the crenellated walls and towers of the convent. The cathedral now houses the Novodevichy Convent Museum, and the complex includes a cemetery where, among other prominent Soviet citizens, Khrushchev is buried. Just beyond Novodevichy, within the large loop of the Moskva and facing the Lenin Hills, is the sports complex known as Luzhniki Park, dominated by the huge Central Lenin Stadium (1955–56), built to accommodate 103,000 spectators.

Novodevichy Convent

The middle zone has seen the most urban renewal in Soviet times. Among the features of the present Moscow skyline are the ornate "wedding cake" skyscrapers along the Garden Ring, built in the late 1940s and early 1950s. In the same Stalin period style are the Ukraina Hotel across the river and the gigantic new building in the Moscow State University complex on the Lenin Hills. Most of the renewal has taken place since 1960 in extensive neighbourhoods of wide streets and rather monotonous apartment buildings. A number of areas still await redevelopment, with narrow streets of 19th-century housing, many smaller factories, and air pollution higher than average.

Outer Moscow. This is the zone of modern factory development and of extensive housing construction in the 1960s and '70s. Closer to the centre are the micro-regions, or neighbourhood housing units, of the Khrushchev period, built predominantly of yellowish brick and in five- to nine-story apartment buildings. Farther out, the 1970s neighbourhoods are characterized by high-rise buildings (up to 20 or more stories), made of standardized, prefabricated concrete sections. Commonly the street levels of the buildings are occupied by shops. Streets are broad and tree-lined. Between the densely populated micro-regions are "green wedges" of open land, notably the extensive Izmaylovsky Park in the east, Sokolniki Park and large forest tracts to the northeast, and on the north, in Dzerzhinsky Park, the grounds of the permanent Exhibition of National Economic Achievements. Nearby, at Ostankino, is the television tower of the 1960s, 1,758 feet (536 metres) tall.

In the sea of new building, individual monuments of the past, such as the Church of the Intercession in Fili, built in 1693, survive. Moscow has engulfed a number of former country estates, with mansions mostly dating from the period of Classical architecture. On the east side of Moscow is Kuskovo, once the estate of the Sheremetyev family, with a palace built in the 1770s, a church, hermitage, and Baroque grotto. To the south is the Znamenskoye-Sadki mansion, formerly of the Trubetskoy family; to the north are the Petrovsky Palace (built by Kazakov in 1775–82) and, best known of all, the Ostankino Palace (1790–98). In the southeastern suburbs is the former village of Kolomenskoye, once a summer residence of the princes of Moscow. Its most remarkable architectural ensemble of buildings is dominated by the Church of the Ascension (1532). The Kazan Church and the gatehouse both date from the later 17th century. The surrounding park has a collection of examples of early Russian wooden architecture, brought from various parts of the country. In the nearby village of Dyakovo is the ornate Church of St. John the Baptist, built in 1557.

Beyond the most recent suburbs are remaining areas of open land and forest within the Circular Motorway, together with satellite industrial towns and dormitory suburbs, which were included within the extended boundaries of Greater Moscow in 1960. The principal satellites are Babushkin to the north, with textile mills and a large number of commuters, Perovo to the east, with major marshalling yards and engineering and chemical factories, Lyublino to the southeast, Kuntsevo to the west, and Tushino to the northwest. One part of Moscow lies beyond the Circular Motorway; this is the overspill town of Zelenograd, a new town of the 1960s modelled on the lines of the British New Towns built around London after World War II. Population and industry from the capital were moved north to Zelenograd; in 1975 it had 117,000 inhabitants.

Satellite towns

POPULATION

The program of housing construction since the late 1950s has greatly reduced the excessive population densities of central Moscow, but densities are nevertheless still high, largely as a consequence of the construction of high-rise apartment buildings. Within the Garden Ring the average density in the early 1970s was more than 76,000 persons per square mile (29,000 per square kilometre). This was higher than the overall density for Manhattan (about 69,000 per square mile). Even Greater Moscow, including large areas of park, forest, and other open land, has an overall density exceeding 20,000 persons per square mile (7,700 per square kilometre), lower than the overall density of New York City (about 26,000 per square mile). Moreover, the constant influx of people into the capital steadily, if slowly, increases these figures. A considerable proportion of the inhabitants are first-generation Muscovites, born outside the city. For most of the Soviet period there has been a deliberate attempt to check the growth of Moscow's population by rigid restrictions on the number of police residence permits granted. This policy has not been altogether successful. On the one hand, it has caused a recurrent labour shortage in the city, by no means made good by developments in automation; on the other hand, it has only checked, but not stopped, inmigration. During the 1970s Moscow's population was growing by more than 100,000 persons each year; of this number, less than 20,000 came from natural increase (*i.e.*, the excess of births over deaths), leaving an excess of arrivals over departures totalling some 80,000–90,000 annually. In addition, there is an unknown number of illegal, unregistered arrivals. This increment puts additional demand on housing.

Population imbalance

The population, in common with that of the Soviet Union as a whole, displays a marked imbalance of males and females. In 1976 there were 774,000 more females than males in Moscow, males constituting 45 percent and women 55 percent of the total. In large measure the imbalance is one of the long-term consequences of losses of men in World War II and other periods of population decimation. In terms of age groups, the able-bodied groups (age 20–60) and more particularly the fertile age groups (20–40) form a larger share of Moscow's population than in the Soviet Union as a whole. Yet the birth rate in the city (12.5 per thousand during 1971–75) is one of the lowest in the country; in 1967 it was only 10.5. At the same time the death rate is slowly and steadily rising, from 8.5 per thousand in the mid-1960s to 10 in 1975, a figure higher than the national average. In consequence, the rate of natural increase is extremely low, 2.5 per thousand in 1975, the lowest figure for any major city in the Soviet Union. At the same time the proportion of the population over retirement age is slowly growing, while the proportion of young people (under 20 years old) is far below the national average (24.8 percent for Moscow, as against 38 percent for the nation). This disparity is widening and is most marked in the under-10 age group.

The inhabitants are overwhelmingly Russian, 89.2 percent in 1970. Jews then formed 3.6 percent, Ukrainians 2.6, and Tatars 1.6. No other group formed more than 1 percent, but, as expected in the capital of a multinational state, the remainder include representatives of the more than 100 nationalities within the country.

ADMINISTRATION AND GOVERNMENT

In the period after the Revolution, Moscow was divided into 11 administrative districts, or wards. With the creation of Greater Moscow in 1960 the number of administrative units was increased to 17 *rayony* (districts; singular *rayon*). The subsequent huge building program and redistribution of population from central areas to the suburbs necessitated a new division in 1968 into 29 *rayony*. The *rayony* tend to be wedge-shaped, pointing in toward the city centre, with an inner ring of 13 *rayony* and an outer ring of 16. The *rayony* vary in population from some 140,000 to 340,000, the inner *rayony* usually being somewhat smaller in population and much smaller in area than the outer. Zelenograd forms a separate, 30th *rayon* of the city. The *rayony* elect a total of 1,160 delegates to the Moscow City Soviet of Workers' Deputies (Mossovet), which has its headquarters on Gorky Street. The council handles the city's annual budget, which is chiefly derived from central government funds. Since Moscow is a "city of republic subordination," the City Council is responsible to the government of the Russian S.F.S.R. rather than to a smaller administrative unit.

Twofold capital

Moscow is a twofold capital, of the U.S.S.R. as a whole, and of the Russian S.F.S.R., which comprises about three-fourths of the area of the Soviet Union. Although the centre of national government is the Kremlin, which houses the meetings of the Presidium of the Supreme Soviet and party congresses, the buildings of various ministries and government departments are scattered fairly widely over the inner city. The Ministry of Foreign Affairs is in a skyscraper of the Stalin period on Smolensk Square on the Garden Ring. Several ministries are housed in the new tower buildings on Kalinin Avenue. The new Comecon and Council of Ministers buildings (at the latter of which the council's Secretariat meets) at the outer end of Kalinin Avenue still further disperse the functions of government. A number of foreign embassies have been transferred to the southwestern suburbs on the Lenin Hills, far from the city centre. Naturally, Moscow as capital has the headquarters of most national bodies and organizations. The Central Committee of the Communist Party has its headquarters on the east side of the Kitaygorod, facing Staraya (Old) Square.

ECONOMY

Employment. Although Moscow's role in the administration of the Soviet Union is of prime importance, administration as a source of employment in the city is relatively minor, if steadily growing, both absolutely and relatively. In 1960 3.2 percent of the employed population was engaged in administration; by 1975 the figure had risen to 5.0. A further 5 percent in 1975 was employed in matters relating to housing (other than construction) and ancillary services and another 5 percent in the various social services such as health. Education and cultural activities, including entertainment, accounted for 6 percent. For a major city, a notably small proportion of the working population is engaged in retailing and catering, 7.9 percent in 1975. Conversely an unusually large proportion (19 percent) was engaged in science and research. The building trade employed 10 percent, transport and communications 9.5 percent. Industry remains the dominant source of livelihood for Muscovites, although its share of total employment is gradually falling, from 36.4 percent in 1960 to 27.8 percent in 1975.

The imbalance between males and females in the population is reflected in the preponderant role of women in the city's labour force. Overall, women form 56 percent of the total number employed, but in certain occupations their share is far greater—in education and culture 72 percent, in retailing and catering 80, and in health and social services 86. Even in industry, more than half the employees (52 percent) are female. Only in transport and the building trade is their role less important, with in each case slightly more than 30 percent of the labour force.

Industry. Moscow is the largest industrial centre of the Soviet Union. The contribution to the national budget made by deductions from the profits of Moscow's enterprises makes up a fifth of all such deductions in the Soviet Union. In addition, Moscow provides to the national budget a tenth of the total turnover tax. The city is the focus of an extensive area of industrial development, which includes not only Moscow and its satellites but also many other textile and engineering towns that have linkages with the capital. This Central Industrial Region stretches east and northeast to the Volga between Yaroslavl and Gorky. Moscow's own industries are mainly of the kind that require a skilled and educated labour force but relatively limited raw materials and that are nonpollutant.

Central Industrial Region

At the time of the Revolution, Moscow's industry was dominated by textiles, but nowadays engineering and metalworking are by far the most important, employing 56.6 percent of the industrial work force. Pride of place is taken by the production of automobiles and trucks.

Mass housing and woods in southwest suburban Moscow.
I. Denisenko—Novosti Press Agency

The largest plant in the city is the Likhachyov Auto Works. The second automobile works is the Lenin Komsomol Automobile Plant, making some 200,000 small cars of the Moskvich type each year and employing about 17,000 workers. Closely linked with these two plants are a number of factories producing component parts, both elsewhere in Moscow (carburetors, shafts, and bodies) and in other towns of the Central Industrial Region, together with some as far away as the Urals. Ball bearings, both for the vehicle industry and for other purposes, are a product of prime importance; more than 3,000 varieties of bearings are made, the largest weighing six tons. The manufacture of machine tools is another major branch of engineering, particularly grinding lathes at the Red Proletariat Plant, precision cutting tools at the Stankoliniya Plant, and machinery for the textile industry. Precision engineering is highly developed and is Moscow's fastest growing industry; it is noted for measuring and other instruments (especially at the Kalibr and Frezer factories), watches, and platinum needles. Other products include pumps, compressors, ventilators, pressure gauges, and hydraulic units for tractors. Moscow is a large centre of electrical, electronic, and radio engineering; the wide range of items made includes transformers, computers, radio and television sets, refrigerators, vacuum cleaners, and electric motors. Aviation engineering and the manufacture of bicycles are also carried on.

Precision engineering

Some of the specialized, high-grade steels required by this formidable array of engineering industries are made in the city at the Hammer and Sickle Steel Works; specialized steels also come from electric hearths in the satellite town of Elektrostal, 36 miles east of Moscow. More steel comes great distances from the heavy industrial regions of the Donets Basin (Donbass) and the Urals.

Second place in terms of employment is taken by the numerous branches of the textile industry, for which Moscow is the largest centre in the country. This includes the manufacture of all types of cloth, worsted (Krasnokholmsky Kamvolny Complex), cotton (Trekhgornaya Mill), and linen and silk (Red Rose Factory). Nowadays artificial fibres are also made. There is considerable finishing of fabrics at specialized enterprises, as well as production of clothing.

Moscow's large chemical industry, employing some 3.1 percent of industrial workers, was originally geared to the textile industry, producing dyestuffs. To this has been added the making of synthetic industrial rubber and rubber tires, paints, plastics, pharmaceutical goods, and perfumes. Many of its chemical products are derived from Moscow's oil refinery, which processes petroleum piped from the Volga–Urals oil field; the automobile industry is a major consumer of the petrochemical products.

Chemical industry

There is a wide range of consumer goods industries, serving primarily the Moscow region but also a wider hinterland. These include foodstuffs, especially confectionery and processed dairy and meat products. Footwear and pianos are specialties of Moscow. Furniture making is part of a varied timber-processing industry, which also makes pulp and paper. Wood processing employs 3.1 percent of the city's workers. Some of the timber is used in the vast construction industry, workers in which include not only the large numbers actually employed in building but also those engaged in making building components, such as reinforced concrete sections, glass, and bricks. An industry of particular note is printing and publishing. Three-quarters of all books and journals and more than half the newspapers in the Soviet Union are printed in Moscow.

Under Moscow's general plan for development, a number of factories were scheduled to be relocated out of the central areas of the city, especially those factories that cause the most harm to the urban environment. At the same time the location of new industries in Moscow was prohibited. The pressures on industrial ministries to keep up and indeed to increase production year by year, however, have operated against the aims of the plan. By the later 1970s there was only limited evidence of any significant move of industries out of Moscow; exceptions were the manufacture of dump trucks, which moved to Mytishchi, just beyond the Circular Motorway, and of buses,

which shifted to Likino-Dulevo, east of Moscow. Established factories were expanded and new industries brought in, among them the Khromotron Works, making colour television tubes. During the 1970s more than 10,000,000 square feet (930,000 square metres) of production space were brought into operation annually. The majority of Moscow's factories are not large; in 1974, 51 percent of them employed fewer than 500 workers. Many of these smaller works are long established, and many make highly specialized items. The 6.8 percent of factories with more than 3,000 workers account for nearly two-fifths of the city's industrial output by value.

Small factories

Power. Moscow draws on a very wide area for its power supply. Electrical generating stations in the city are now based on natural gas, which is piped from Saratov on the Volga and Gazli in Central Asia, from Stavropol in the North Caucasus, from the Volga–Urals region, and from Shebelinka and Dashava in the Ukraine. Large gas storage facilities have been constructed near Moscow, and a linking ring pipeline surrounds the city. The changeover from coal to natural gas as the principal fuel source has been a major factor in the improvement of the city's atmosphere. To the east of Moscow the large peat resources of the Meshchera Lowland are converted to electricity in stations near Shatura. High-voltage transmission lines (500 kilovolts) supply power from the hydroelectric stations on the Volga at Uglich, Rybinsk, Gorky, Kuybyshev, and Volgograd. Other sources are the very large thermal stations at Konakovo to the northwest and Kashira to the south of Moscow. A by-product of the power industry is hot water, which is piped from the 22 district central stations to apartment buildings for heating and domestic supplies.

Transport. *Rail.* The Soviet Union is heavily dependent for the transport of freight on its railways, which in 1975 carried 62 percent of the total national freight turnover. The hub of this network is the capital, which handles about a tenth of total freight. The large marshalling yards cope with some 450 freight trains daily and about 71,000,000 metric tons of freight each year, of which 54,000,000 tons are incoming. Eleven trunk lines radiate out from the city: the Leningrad line, first to be opened (1851); the Savyolovo line (1902), running north to the Volga and on as a secondary route to Leningrad; the Yaroslav line (1862), which continues on to Kirov, Perm, the Urals, and, by the Trans-Siberian route, to Vladivostok; the Gorky line (1861), also linking to Kirov; the Kazan line (1912), the most direct route to the Urals and Siberia; the Ryazan line (1862), leading to the southern Urals, Central Asia, the lower Volga, and the Caucasus; the Pavelets line (1900), a secondary route to the European south and the Caucasus; the Kursk line (1866), the main route south to Kharkov, the Donbass, and the Crimea; the Kiev line (1898) to the southwest; the Smolensk line (1870) to Minsk, Warsaw, and Berlin; and the Riga line (1911) to the Baltic. The most heavily used of these lines are now electrified—notably those to Leningrad, Kiev, the Donbass, and the Trans-Siberian—and all lines in the suburban zone are electrified to carry the heavy commuter traffic. To link the radial lines the inner Circle Railway was built in 1908 within the city; this has been supplemented by the Greater Moscow Ring Railway at a distance of some 25 to 40 miles from the city.

Passenger trains connect to destinations throughout the Soviet Union and Europe; through coaches run to Berlin, Vienna, Paris, and the Hook of Holland. Several hundred long-distance passenger trains each day move some 170,000 people, and the 1,500 suburban trains carry a further 850,000 persons daily. The period since World War II has seen a marked growth of commuting into Moscow from satellite towns, and by the early 1970s daily commuters numbered about 500,000.

Waterways. Moscow is a major river port. The canalized Moskva can take only smaller craft, but the Moscow Canal, built by forced labour in 1932–37, can take ships of seagoing size. The canal runs from the Moskva, upstream of the city, northward to the Khimki and Ucha reservoirs (the latter linking to the Klyazma Reservoir and the navigable Klyazma River) and thence north through the Klin-Dmitrov Ridge to the Volga at Ivankovo, in all 80 miles (128 kilometres). The Volga's various canal links open up Moscow to all the seas surrounding European Russia. Minimum depth on the Moscow Canal is 18 feet (5.5 metres), and lock dimensions are 950 by 100 feet (290 by 30 metres). The capital has three river ports, the Southern River Station on the Moskva downstream, the Kiev Station upstream, and the Northern River Station on the Khimki Reservoir. River freight turnover is about 29,000,000 metric tons annually. Passengers are catered for by short- and long-distance cruises.

Moscow Canal

Air. Moscow is similarly the hub of the Soviet Union's airline network, and the number of passengers carried is rising steadily year by year. The original Tushino Airport has long been abandoned, but Moscow now has four airports. Sheremetyevo to the north is the main airport for international flights, with direct links to most European capitals, to Cuba, New York City, Montreal, Tokyo, and other foreign cities. Internal flights are handled at Vnukovo, 15 miles to the southwest; at Bykovo, 20 miles to the southeast; and at Domodedovo, the newest and largest, to the south.

Intracity transport. Compared with other large cities in most developed countries, Moscow has very few privately owned cars, though their number is slowly and steadily rising. On the one hand, this means relative freedom from serious traffic problems, while, on the other, it necessitates heavy reliance on public transport. In this respect Moscow is in general well served. In 1975 nearly 5,300,000,000 people were transported, an increase of 25 percent in a decade. The most important element in Moscow's city transport is the Metropolitan (Metro) subway, with 103 stations and some 295 miles of line. It carried 1,966,000,000 passengers in 1975, or almost 5,400,000 each day. The system, begun in 1935, is renowned for the elaborate architecture of its stations, more especially the older ones, which are highly decorated with marble, stained glass, statuary, and chandeliers. One of the stations built later, Lenin Hills, forms the glass-sided lower level of a two-tier bridge over the Moskva. Train service throughout the day is frequent. The network copies the city street pattern, with a series of radial lines linked by a single ring subway.

Subway system

Next in importance are the motor bus services, carrying 1,781,000,000 people in 1975 in more than 5,000 buses on 200 routes. Moscow still relies heavily on streetcars (trams) and trolleybuses, which in 1975 moved 637,000,000 and 910,000,000 passengers, respectively, in some 1,500 streetcars and 2,000 trolleybuses. The inflexibility of these fixed-line forms of transport poses a potential problem as traffic gradually increases.

Although Moscow has for the most part a dense network of surface transport, new routes sometimes tend to lag behind housing construction, so that certain new suburbs are rather poorly served for some time after their completion. On the credit side, fares on all forms of transport are very low and are standard regardless of distance.

Commercial functions. As capital and as the largest city, Moscow not surprisingly is the largest commercial centre of the Soviet Union. The headquarters of nationwide state insurance and state banking organizations are there. Its retailing facilities are heavily used, not only by the inhabitants of Moscow itself and its surrounding satellite towns but also by people from all over the Soviet Union. Its shops are better stocked with a wider range of goods than those of any other city and are always crowded with customers. Yet in comparison with major cities of western Europe and North America, Moscow is poorly provided with retail outlets. In 1976 there were about 10,500 retail outlets in all Greater Moscow (compared with 37,500 in New York City), employing more than 65,000 people (1970). Of these, 5,370 were stalls or kiosks and only 5,150 were shops; the figure 5,370 represented an increase of only 330 over 1966. Many of the stores are fairly large, including some of the central food shops (*gastronomy*) and department stores (*univermagy*). The most important of the 66 department stores are Children's World, the Central Department Store (TsUM), and, best

known and most heavily patronized, the State Department Store (GUM) on Red Square. GUM, the direct descendant of the medieval trading rows, is alone responsible for about 10 percent of Moscow's total retail trade in goods other than foodstuffs. Moscow does not have shopping centres with local concentrations of retailing; the usual practice is for shops to be widely dispersed on the ground-floor levels of apartment buildings, which tends to add to the time taken up by shopping. A very important part of Moscow's retail trade is carried on at the "collective markets," where collective farms and individual members of such farms bring fruit, vegetables, and meat for direct sale to the populace; although often more expensive than the state shops, they tend to excel in quality and availability of produce.

There are almost 7,800 eating places in Moscow, but most are canteens in places of work, as are most of the 2,300 cafés, buffets, and snack bars; there are 355 restaurants, including those in hotels. As the capital and as the principal focus for foreign tourists, Moscow has always a number of transient inhabitants, but hotel accommodation is rather limited for so large and important a city. In 1975 there were 25 public hotels with places for 33,361 people, with accommodation for another 14,000 in hotels run by the armed services and other bodies for their own personnel. Leading hotels include the huge, modern Rossiya, the Ukraina (one of the Stalin period skyscrapers), and the pre-Revolutionary National and Metropole. In the late 1970s plans were being put into effect for a considerable increase in hotel accommodation to meet the expected influx of visitors for the Olympic Games in 1980; the plans include a very large hotel at Ostankino in the north.

Hotels

EDUCATION, CULTURAL LIFE, AND SOCIAL SERVICES

Education. Moscow has an exceptionally large concentration of educational establishments. At the lowest level, serving the city's own population, are more than 1,050 schools with more than 850,000 pupils (1975–76). Of these, 65 are for handicapped children, some 75 are special foreign language schools, and 36 are boarding schools. For children below school age (seven years old in the Soviet Union) there are about 1,300 nurseries and kindergartens accommodating 156,000 children; many are attached to individual places of work. Moscow's 76 higher educational institutions draw students from all over the Soviet Union, totalling 635,200 in 1975–76. Nearly 44 percent of the students were registered for correspondence courses. Although six out of every seven were Russians, the total included representatives of almost every nationality in the Soviet Union, together with a number of overseas students.

Leading place is taken by the Moscow M.V. Lomonosov State University, founded in 1755 and in 1975–76 having more than 26,000 students, of whom approximately a tenth are from foreign countries. Of its 17 departments, some of the humanities are still housed in the old university buildings facing the Square of the 50th Anniversary of October, near the Kremlin. Science departments are located in the complex of buildings on the Lenin Hills, which date from the early 1950s and are dominated by a huge 34-story Stalinesque skyscraper; this building houses the central administration, three departments, the Museum of Earth Science, and accommodation for thousands of students. In 1970 a new humanities building was constructed on the Lenin Hills site. From 1960, students from African, Asian, and Latin-American countries were segregated in Moscow's newly founded second university, the Patrice Lumumba People's Friendship University, which now has about 7,500 students, including Russians.

Of the many specialist higher educational foundations, outstanding importance attaches to the Moscow Timiryazev Academy of Agriculture in the north of Moscow and the world-famous Moscow P.I. Tchaikovsky State Conservatoire, where many of the world's finest musicians received their training. Also important are the Moscow D.I. Mendeleyev Institute of Chemical Technology, the Moscow N.E. Bauman Higher Technical School, and the State A.V. Lunacharsky Institute of Dramatic Art. Many institutes, among them the Moscow Institute of Aviation Technology and the All-Union Extra-Mural Institute of Railway Engineers, produce specialists for particular industries.

Heading Moscow's libraries is the State V.I. Lenin Library of the U.S.S.R., the principal copyright library of the Soviet Union and one of the world's largest. Some 40,000 different books are published annually in the city (in all, nearly 1,200,000 copies), together with nearly 3,600 journals.

Lenin Library

Scientific research. In addition to the large volume of research undertaken in the university and in other teaching establishments, there is a formidable array of scientific research institutions. Between them they constitute the second largest employer in Moscow after industry, with 17.6 percent of the total employed population and including some 234,000 scientists. More than a third of all Soviet citizens holding the doctor of science degree are employed in Moscow. The main focus of research is the Academy of Sciences of the U.S.S.R., the Presidium of which is located on Lenin Avenue. There are also national academies of arts and medical, pedagogic, and agricultural sciences.

Many scientific institutes, such as the Experimental Research Institute of Metal-Cutting Machine Tools, are concerned with highly specialized industrial research. These are often connected closely with Moscow's own industries, as in the case of the Central Research Institute of Automobile Engineering. Linked to the research bodies are the many design bureaus, such as the All-Union Planning, Surveying, and Scientific Research Institute, which designs hydroelectric power projects, and the State All-Union Institute for the Planning of Metallurgical Plants.

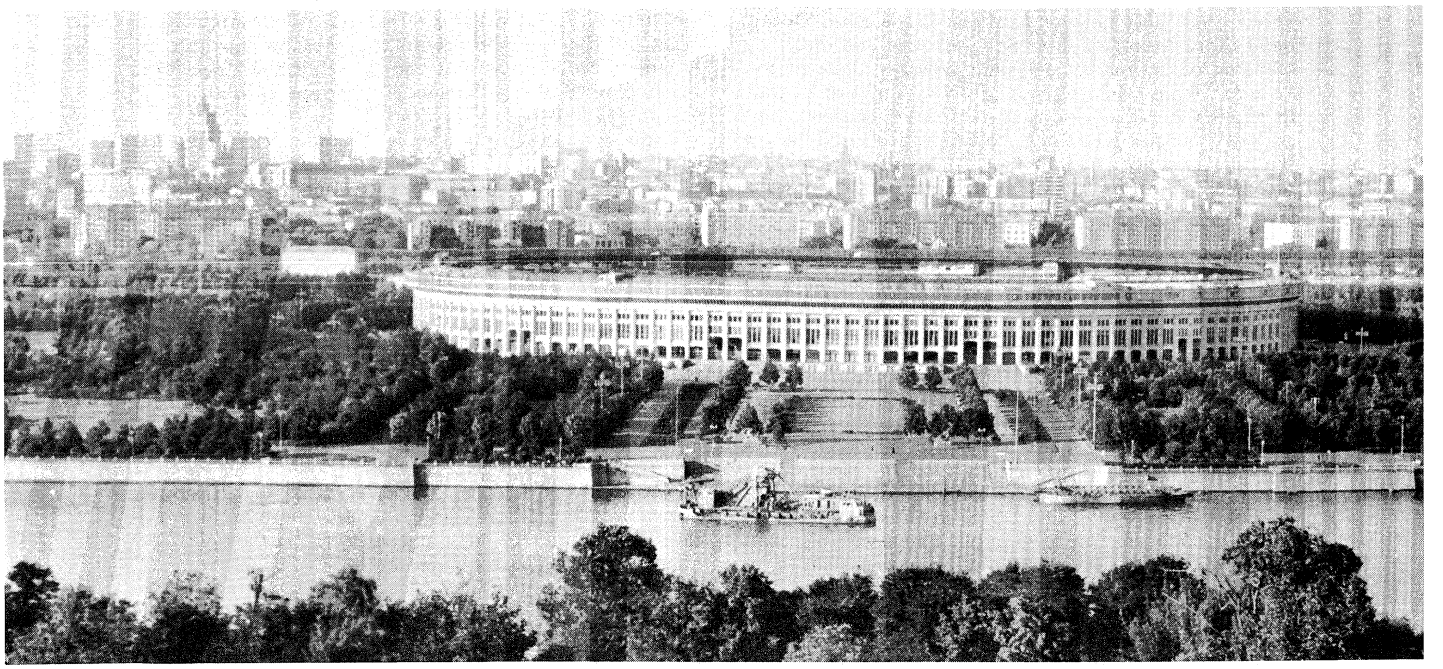
Entertainment. Moscow has 28 theatres, headed by the justly renowned State Academic Bolshoi (Great) Theatre, which was founded in 1825, though its present splendid building facing Sverdlov Square dates from 1856. Also on Sverdlov Square is the Maly (Little) Theatre. The city's principal drama theatre, the Moscow Art Theatre, is particularly noted for its productions of Chekhov and other classical playwrights; founded in 1898, it is especially associated with the name of K.S. Stanislavsky. Also of worldwide fame are the State Central Puppet Theatre and the Moscow State Circus, which in 1971 acquired new quarters on the Lenin Hills. The resident companies of all these theatres are frequently on tour, both in the Soviet Union and abroad. There are several concert halls, notably the Tchaikovsky Concert Hall and the two halls of the Conservatoire. The U.S.S.R. State Philharmonic Orchestra has won international repute, as have a number of Moscow-based folk dance and choral ensembles.

Moscow has 119 cinemas, with further facilities for showing films in several hundred clubs and houses of culture. The October and Rossiya cinemas are the largest in the Soviet Union, both with seating for more than 2,800. There are three studios producing motion pictures in the city. Four television channels are available, the main studios being at Ostankino near the soaring television tower. Moscow radio broadcasts not only in Russian but also in many of the other languages of the Soviet Union and in more than 60 foreign languages. There are 34 newspapers, headed by the two principal national newspapers, the morning *Pravda* and evening *Izvestiya*.

Cinema

Museums and galleries. The 65 museums and art galleries in the capital include several of international rank. In first place are the State Pushkin Museum of Fine Arts, with a very fine international collection, and the State Tretyakov Gallery. The latter, which began in 1856 as the private collection of a connoisseur, Pavel Tretyakov, is noteworthy for its superb collection of icons, including several by Rublyov. Other notable museums are the Armoury Museum in the Kremlin, the State Historical Museum on Red Square, and the nearby Central Lenin Museum on Revolution Square.

Recreation and health facilities. The Luzhniki Park complex heads Moscow's provision for sports and is intended to form one of the main arenas for the 1980 Olympic Games. The Central Lenin Stadium, accom-



Central Lenin Stadium, as seen from the Lenin Hills in Moscow.
Novosti Press Agency

modating 103,000, is flanked by a smaller arena, swimming pool, and indoor Sports Palace. Altogether there are nearly 70 stadiums in the city and 26 swimming pools, including a large, heated, open-air pool that is in use all year round. In addition there are some 290 soccer fields, 1,200 gymnasiums, and nearly 1,600 volleyball and basketball courts; most of these are attached to individual places of work or to sports clubs.

Parks

Outside the Garden Ring, Moscow is well provided with parks and open spaces. Gorky Central Park of Culture along the right bank of the Moskva is the closest to the centre and, with its fun fair, is very popular. By far the largest park, covering nearly 3,000 acres (1,200 hectares), is Izmaylovsky Park on the east side. To the northeast is the more formal Sokolniki Park, which leads on to an extensive tract of forest. North of the city are the Botanical Gardens of the Academy of Sciences and the grounds of the Moscow Timiryazev Academy of Agriculture. The Moscow Zoo, one of the world's leading zoos, is visited by more than 3,800,000 people every year. The Khimki Reservoir is used for boating and aquatic sports, but even more popular are the other reservoirs to the north, just outside Greater Moscow. The surrounding forest–park zone provides extensive space for recreation.

The health of the Muscovites is attended to at the neighbourhood level by almost 1,000 polyclinics, giving medical, dental, and maternity services. There are 193 hospitals, with a total of 93,500 beds, the largest being the Botkinskaya Hospital with 2,650 beds. Among specialist institutions are 21 maternity, 12 isolation, 16 mental, and 7 tuberculosis hospitals, supported by a number of specialist medical research institutions.

OUTLOOK

Development plan

The General Plan for the Development of Moscow published in 1971 covers the next 30 years. It assumes that by the year 2000 Moscow will have a stable population of about 8,500,000. Further housing construction will continue to be based on the micro-region neighbourhood unit but will provide more advanced services; there will be more variety in building styles, but emphasis will still be placed on high-rise buildings. The micro-region will be fitted into a more clearly defined functional structure. To this end the plan envisages eight city zones, a central zone and seven peripheral zones, each with local administrative, retailing, and entertainment functions. Dilapidated areas in the central zone, which serves the entire city area, will be redeveloped, but emphasis will be placed on preservation of its historic character and architectural monuments. The other zones will form wedges of housing and industry, separated by green wedges of open space. Existing plans to prevent new industrial development and to remove noxious factories from the city are retained as features of the plan; the intent will be to spread industry more evenly through the city in order to reduce the length of journeys to and from work. At the same time there will be major improvements to the public transport system, including doubling the length of subway lines, providing express services, and building airport links. Road widening and the building of express highways to cope with projected increases in car ownership also form part of the plan. Much of the plan is concerned with the links between Moscow and its surrounding metropolitan region, which (outside the strictly protected forest–park amenity zone) will take the brunt of future industrial and population growth; it is expected that the population of the surrounding area will double by the end of the century.

(R.A.F.)

BIBLIOGRAPHY. Академия наук СССР, *История Москвы,* 7 vol. (1952–67), the most complete survey of Moscow, including detailed and illustrated analyses of the economy, culture, science, arts, architecture, education, public health, and other aspects of the city's life; and *Москва за 50 лет Советской власти. 1917–1967* (1968), statistics of the development of the city after the October Revolution; Владимир Федорович Промыслов, *Развитие индустриального строительства в Москве* (1967), a history of building and the construction industry in Moscow, with emphasis on the methods used in mass construction of both residential and industrial buildings; Петр Васильевич Сытин, *Из истории московских улиц; очерки,* 3rd ed. (1958), a broad popular account of the history and topography of Moscow's streets and squares; Ю.Г. Саушкин, *Москва; географическая характеристика,* 4th ed. (1964), a geographical-historical survey of the city; Музей архитектуры, *Москва; архитектурный путеводитель* (1960), a guidebook that provides detailed descriptions of the most important architectural sites; Борис Сергеевич Земенков, *Памятные места Москвы; страницы жизни деятелей науки и культуры* (1959), on architectural monuments, buildings, and sites; *Москва в цифрах* ("Moscow in Figures") is a short statistical compendium containing data on most aspects of the city and its economy.

(I.K.M./Ed.)

General descriptive works in English include: IVAN K. MIACHIN and VLADIMIR TCHERNOV, *Moscow: Tourist Guide* (Eng. trans., 1967); A.P. KOVALEV, *Moscow: A Short Guide* (Eng. trans., 1957); DAVID D. DUNCAN, *Great Treasures of the Kremlin,* rev. ed. (1968); ARTHUR VOYCE, *The Moscow Kremlin: Its History, Architecture, and Art Treasures* (1954);

V. ALEXANDROV, *The Kremlin* (Eng. trans., 1963); B.A. RYBAKOV, *Treasures in the Kremlin* (Eng. trans., 1962); KAREL NEUBERT, *Portrait of Moscow* and *Palaces and Churches of the Kremlin* (Eng. trans., both 1965); VLADIMIR CHERNOV and MARCEL GIRARD, *Splendors of Moscow and Its Surroundings* (Eng. trans., 1967); and two photographic views: HENRI CARTIER-BRESSON, *The People of Moscow* (1955); and JAN LUKAS, *Moscow* (1964).

Moses

Moses was the gifted Hebrew leader who, in the 13th century BCE (before the Common Era, or BC), delivered his people from Egyptian slavery. In the Covenant ceremony at Mt. Sinai, where the Ten Commandments were promulgated, he founded the religious community known as Israel. As the interpreter of these Covenant stipulations, he was the organizer of the community's religious and civil traditions. In the Judaic tradition, he is revered as the greatest prophet and teacher, and Judaism has sometimes loosely been called Mosaism, or the Mosaic faith, in Western Christendom. His influence continues to be felt in the religious life, moral concerns, and social ethics of Western civilization, and therein lies his undying significance.

THE HISTORICAL PROBLEM

Historical views of Moses. Few historical figures have engendered such disparate interpretations as has Moses. Early Jewish and Christian traditions considered him the author of the Torah ("Law," or "Teaching"), also called the Pentateuch ("Five Books"), comprising the first five books of the Bible, and some conservative groups still believe in Mosaic authorship.

View of Martin Noth

Opposing this is the theory of the German scholar Martin Noth, who, while granting that Moses may have had something to do with the preparations for the conquest of Canaan, was very skeptical of the roles attributed to him by tradition. Although recognizing a historical core beneath the Exodus and Sinai traditions, Noth believed that two different groups experienced these events and transmitted the stories independently of each other. He contended that the biblical story tracing the Hebrews from Egypt to Canaan resulted from an editor's weaving separate themes and traditions around a main character Moses, actually an obscure person from Moab.

This article, following the lead of the biblical archaeologist and historian W.F. Albright, presents a point of view somewhere between these two extremes. While the essence of the biblical story (narrated between Exodus 1:8 and Deuteronomy 34:12) is accepted, it is recognized that during the centuries of oral and written transmission the account acquired layers of accretions. In general, the reconstruction of the documentary sources of the Pentateuch by literary critics is considered valid, but here the sources are viewed as varying versions of one series of events (see BIBLICAL LITERATURE: *Old Testament literature*). Other critical methods (studying the biblical text from the standpoint of literary form, oral tradition, style, redaction, and archaeology) are equally valid. The most accurate answer to a critical problem is thus likely to come from convergence of various lines of evidence. Despite the aid of critical scholarship, the sources are so sketchy that the man Moses can be pictured only in broad outline.

The date of Moses. According to the biblical account, Moses' parents were from the tribe of Levi, one of the groups in Egypt called Hebrews. Originally the term Hebrew had nothing to do with race or ethnic origin. It derived from Habiru, a variant spelling of Ḫapiru (Apiru), a designation of a class of people who made their living by hiring themselves out for various services. The biblical Hebrews had been in Egypt for generations, but apparently they became a threat, so one of the pharaohs enslaved them. Unfortunately, the personal name of the king is not given, and scholars have disagreed as to his identity and, hence, as to the date of the events of the narrative of Moses. One theory takes literally the statement in I Kings 6:1 that the Exodus from Egypt occurred 480 years before Solomon began building the Temple in Jerusalem. This occurred in the fourth year of his reign, about 960 BCE; therefore, the Exodus would date about 1440 BCE.

This conclusion, however, is at variance with most of the biblical and archaeological evidence. The storage cities Pithom and Rameses, built for the pharaoh by the Hebrews, were located in the northeastern part of the Egyptian delta, not far from Goshen, the district in which the Hebrews lived. It is implicit in the whole story that the pharaoh's palace and capital were in the area, but Thutmose III (the pharaoh in 1440) had his capital at Thebes, far to the south, and never conducted major building operations in the delta region. Moreover, Edom and Moab, petty kingdoms in Transjordan that forced Moses to circle east of them, were not yet settled and organized. Finally, as excavations have shown, the destruction of the cities the Hebrews claimed to have captured occurred about 1250, not 1400.

Inasmuch as tradition figured about 12 generations from Moses to Solomon, the reference to 480 years is most likely an editorial comment allowing 40 years for each generation. Since an actual generation was nearer 25 years, the most probable date for the Exodus is about

By courtesy of Staatliche Museen Preussischer Kulturbesitz Gemaldegalerie, West Berlin

"Moses Showing the Tables of the Law to the People," oil painting by Rembrandt, 1659. In the Staatliche Museen Preussischer Kulturbesitz, Berlin.

1290 BCE. If this is true, then the oppressive pharaoh noted in Exodus (1:2–2:23) was Seti I (reigned 1318–04), and the pharaoh during the Exodus was Ramses II (*c.* 1304–*c.* 1237). In short, Moses was probably born in the late 14th century BCE.

YEARS AND DEEDS

The formative years. One of the measures taken by the Egyptians to restrict the growth of the Hebrews was to order the death of all newborn Hebrew males. According to tradition, Moses' parents, Amram and Jochebed (whose other children were Aaron and Miriam), hid him for three months and then set him afloat on the Nile in a reed basket daubed with pitch. The child, found by the pharaoh's daughter while bathing, was reared in the Egyptian court. While many doubt the authenticity of this tradition, the name Moses (Hebrew Moshe) is derived from Egyptian *mose* ("is born") and is found in such names as Thutmose ([The God] Thoth Is Born). Originally,

it is inferred, Moses' name was longer, but the deity's name was dropped. This could have happened when Moses returned to his people or possibly even earlier, because the shortened form Mose was very popular at that time.

The young prince or courtier

Moses' years in the court are passed over in silence, but it is evident from his accomplishments later that he had instruction in religious, civil, and military matters. Since Egypt controlled Canaan (Palestine) and part of Syria and had contacts with other nations of the Fertile Crescent, Moses undoubtedly had general knowledge of life in the ancient Near East. During his education he learned somehow that he was a Hebrew, and his sense of concern and curiosity impelled him to visit his people. According to the biblical narrative, Moses lived 120 years and was 80 when he confronted Pharaoh, but there is no indication how old he was when he went to see the Hebrews. Later Jewish and Christian tradition assumed 40-year periods for his stay in the Egyptian court, his sojourn in Midian, and his wilderness wanderings.

Most likely Moses was about 25 when he took the inspection tour among his people. There he saw the oppressive measures under which they laboured. When he found an Egyptian taskmaster beating a Hebrew, probably to death, he could control his sense of justice no longer. After checking to make sure that no one was in sight, he killed the tough Egyptian overlord. As a prince in the court, Moses was probably in excellent physical condition, and apparently he knew the latest methods of combat.

The flush of victory pulled Moses back the next day. He had removed one threat to his people and was determined to assist them again. This time, however, he found two Hebrews fighting. After parting them, he questioned the offender in an attempt to mediate the disagreement. Two questions jolted him: "Who made you a prince and a judge over us? Do you intend to kill me as you killed the Egyptian?" The confidence of the self-appointed deliverer turned into fear. One of his own knew his "secret" and soon Pharaoh would, too. Realizing that he would have to flee, he went to Midian (mainly in northwest Arabia).

Moses in Midian. In noting the flight to Midian the narrative says nothing of the difficulties involved. Like Sinuhe, the Egyptian court official whose flight in about 1960 BCE was narrated in a famous story, Moses undoubtedly had to filter through the "Wall of the Ruler," a series of forts at the eastern border, approximately where the Suez Canal is now located. From there he made his way southeast through very desolate country. Unfortunately, the Bible does not specify the part of Midian in which Moses resided. Midian proper was east of the Gulf of Aqaba, in the northern section of Hejaz in Arabia, but there is evidence that some of the Midianite clans crossed over the Arabah (the great valley south of the Dead Sea) and settled in the eastern and southern sections of the Sinai Peninsula.

While Moses was resting at a well, according to the biblical account, seven daughters of the Midianite priest Jethro came to water their father's flocks. Other shepherds arrived and drove the girls away in order to water their own flocks. Again Moses showed his courage and prowess as a warrior because he took on the shepherds (perhaps with the girls' help) and routed them. Moses stayed on with Jethro and eventually married Zipporah, one of the daughters. In assuming the responsibility for Jethro's flocks, Moses roamed the wilderness looking for pasture.

The burning bush

One day at the base of a mountain, his attention was attracted by a flaming bush, but, oddly, it was not consumed. He had seen bushes brilliant with flamelike blossoms, but this phenomenon was different, and so he turned aside to investigate it. Before he could do so, he was warned to come no closer. Then he was ordered to remove his sandals because he was standing on holy ground.

Regardless of how one interprets the burning bush, the important fact is that Moses was conscious of an encounter with Deity. This God, who claimed to be the God of Abraham, Isaac, and Jacob, was calling him to deliver the Hebrews from Egypt. Although on his own he had previously been zealous to help his own people, now that he was being commissioned to deliver them he expressed doubt concerning his qualifications. The underlying reason was probably fear—he had fled from Sethos I, and he did not relish confrontation with Ramses II. God reassured Moses that in the future he and the Hebrews would worship at this mountain. Then Moses asked to know the name of the Deity commissioning him. The God of the fathers had been known mostly as El 'Elyon (God Most High) or El Shaddai (God of the Mountain or Almighty God), but he identified himself to Moses as Yahweh and gave instructions that he was to be called by his new name from then on. As the causative form of the verb "to be," Yahweh means He Who Creates (Brings Into Being). This revelation enabled Moses to understand the God of the Hebrews as the sovereign Lord over nature and the nations of the world.

Even after further assurances, Moses was still reluctant to accept Yahweh's call; therefore, he pleaded for release because he was a stammerer. Yahweh acknowledged the defect but promised to help him express himself. Awed by his assignment, Moses made a final desperate plea, "Oh, my Lord, send, I pray, some other person." Although angry at Moses, Yahweh would not yield. Moses would still be Yahweh's representative, but his golden-tongued brother Aaron would be the spokesman. Apparently Moses was ready to play the role of God to Pharaoh as long as Aaron would serve as his prophet. He returned to Jethro and requested permission to visit his people in Egypt, but he did not disclose that he had been commissioned by Yahweh.

Moses and Pharaoh. Ramses II became king as a teenager and reigned for 67 years. He aspired to defeat the Hittites and control all of Syria, but in the fifth year of his reign Ramses walked into a Hittite trap laid for him at Kadesh, on the Orontes River in Syria. By sheer determination he fought his way out, but in the light of his purpose the battle was an utter failure. Yet Ramses, like all the pharaohs, claimed to be divine; therefore, the defeat had to be interpreted as a marvellous victory in which he alone subdued the Hittites. His wounded ego expressed itself in massive building operations throughout Egypt, and before his reign ended the boast of his success literally filled acres of wall space.

Confrontation with Ramses

It was probably only a few years after the Kadesh incident that Moses and Aaron confronted Ramses with their demand, "Thus says the Lord, the God of Israel, 'Let my people go.'" As a god in human form Ramses was not accustomed to taking orders from lesser gods, let alone an unknown like Yahweh. "Who is the Lord," he inquired, "that I should heed his voice and let Israel go? I do not know the Lord, and moreover I will not let Israel go." Thus the stage was set for a long struggle between a distrustful ruler with an outsize ego and a prophet with a new understanding of Yahweh and his power.

Ramses increased the oppression of the Hebrews by the fiendish plan of requiring them to gather the straw binder for the bricks and still produce the same quota each day. Some of the Hebrews rebuffed Moses, and in frustration he asked Yahweh, "Why didst thou ever send me?" Moses' doubt was allayed by Yahweh's promise to take action against Pharaoh. Scholars differ widely concerning the narrative about the plagues. Some claim that three sources have been combined, but more recent scholarship finds only the two traditions. While granting that some of the plagues had a core of historicity, older critics tended to discount the present accounts as fantastic stories with pious decorations. A recent school of research suggests that, notwithstanding some later additions, all the plagues probably had a historical core.

The basic cause, according to one interpretation, was an unusually high flooding of the Nile. The White Nile originates in the lake region of east central Africa, known today as Uganda. The flow is fairly even throughout the year because of consistent equatorial rains. The Blue Nile, on the other hand, originates in the headwaters of the Ethiopian highlands, and it varies from a small stream to a raging torrent. At the time Moses was bargaining with

Ramses, excessively heavy summer rains in Ethiopia washed powdery, carmine-red soil from the slopes of the hills. Around the Lake Tana region the blood-red torrent picked up bright red algae (known as flagellates) and their bacteria. Since there were no dams at that time, the Nile flowed blood-red clear to the Mediterranean. It probably reached the delta region in August. Thus, this rare natural event, it is held, set in motion a series of conditions that continued until the following March.

During these months Moses used the plagues of the frogs, gnats, mosquitoes, cattle murrain, boils, hail, locusts, and thick darkness to increase the pressure on Ramses. At first the King was adamant. The Hebrews were not the only disgruntled slaves, and, if he agreed to let them go, then other groups would want the same privilege. To protect his building program, he had to suppress the slave rebellion at its outset. Yet he could not discount the effect of the plagues, and grudgingly he began to acknowledge Yahweh's power. As an expedient attempt to restore order, he offered to let the Hebrews sacrifice in Goshen. When this failed, he suggested that they make offerings to Yahweh at the edge of the Egyptian border. Moses, however, insisted on a three-day journey into the wilderness. Pharaoh countered by allowing the Hebrew men to make the journey, but this, too, was rejected. As his final offer Pharaoh agreed to let the people go. He would keep the livestock, however, as the guarantee of their return. Moses spurned the condition, and in anger Pharaoh drove him out. After nine rounds with Pharaoh it appeared that the deliverance of the Hebrews was no nearer, but, in contrast to his earlier periods of doubt and frustration, Moses showed no despair. Apparently he had an inner assurance that Pharaoh would not have the last word.

From Goshen to Sinai. Chapters 11–14 of Exodus comprise an exceedingly complex section, and at times the traditions have contradictory statements. The drama is more blurred than usual, and scholars vary tremendously in their interpretation of the material. One tradition notes that Pharaoh was shaken when death took his son and that he ordered the Hebrews to leave. Another source indicates that Moses used the period of mourning for the first-born son as the occasion for fleeing secretly from the country. In either case, it is clear that Pharaoh finally had his forces pursue the Hebrews. Although tradition interpreted the Hebrew text to claim that about 2,000,000 people left Egypt, interpretation by critical methods reduces the number to 15,000 or so.

Yahweh's victory at the Reed Sea

The Egyptian army cornered them at the Sea of Reeds (papyrus), which barred their exit to the east. Later Jewish tradition understood the body of water to be the Red Sea, and this erroneous interpretation persists today, even in some of the most recent English translations of the Bible. Scholars disagree as to the precise location of the Reed Sea, but, since papyrus grows only in freshwater, it was most probably a shallow lake in the far northeastern corner of Egypt.

Hemmed in by the Egyptians, the people vented their complaints on Moses. According to one tradition, Moses shared their uneasiness, and he called to Yahweh for help. Another account claims that Moses confidently challenged them to be calm and watch for Yahweh's deliverance. A strong east wind blew all night, creating a dry corridor through the lake and permitting the Hebrews to cross. The pursuing Egyptians were destroyed when the waters returned. The timing of this natural event gave the final answer to Pharaoh's arrogant question, "Who is Yahweh?" Safely on the other side, Moses and his sister Miriam led the people in a victory song of praise to Yahweh (Ex. 15:1–21). The style of the poetry is similar to that of 14th-century Canaanite literature, and there is every reason to believe that the poem virtually preserves the original form of the song, with its refrain, "Sing to the Lord, for he has triumphed gloriously; the horse and his rider he has thrown into the sea."

The route of the Hebrews is contested by scholars, but the most likely possibility is the southern route to Jabal Mūsā, the traditional location of Mt. Sinai (Horeb), in the granite range at the southern tip of the Sinai Peninsula. The journey there traversed some very desolate country, and Moses had to contend with bitter complaints about the lack of water and food. Finally, however, he brought the people to "the mountain of God," where Yahweh had appeared to him in the burning bush.

The Covenant at Sinai. During the 14th century BCE the Hittites of Asia Minor made a number of treaties with neighbouring rulers who came under their control. The agreement was not between equals, but between the Hittite king (the suzerain) and a subordinate ruler (the vassal). In the prologue the Hittite ruler described himself as "the great king," the one granting the treaty. Then followed a historical survey of relationships between the Hittite suzerain and his vassal. Special attention to the kindnesses shown the underling by the overlord was intended to remind the vassal of his obligation to abide by the treaty stipulations. The basic requirement was an oath of loyalty. Since Egypt was involved with the Hittites in the international politics of the time, Moses probably learned about the Hittite treaty form during his years in the Egyptian court.

The revelation at Sinai: the Ten Commandments

The appearance of Yahweh in a terrific storm at Mt. Sinai, narrated in chapters 19 and 20 of Exodus, was a revelatory experience for Moses, just as the burning bush had been. Somehow he realized that the Hittite treaty was an accurate analogy of the relationship between Yahweh and the Hebrews. Yahweh had a claim upon them because he had delivered them. The only proper response to his love and care would be a pledge of obedience to his will. Scholars have tended to date the Decalogue (contained in the revelation at Sinai) after the conquest of Canaan, but there is absolutely nothing in these guidelines to indicate their origin in an agricultural context. More likely they were the stipulations in the covenant ceremony at Mt. Sinai.

Because Yahweh was proclaimed the only true God, one of the first commands was appropriately a ban against all other gods. Authorities have debated whether or not this understanding was interpreted as monotheism. Most certainly it was not the philosophical monotheism of later periods, but it was a practical monotheism in that any gods recognized by other nations were under Yahweh's control. Inasmuch as he had brought them into being and authorized their presence in his council, he was Lord over all gods and nations.

Another early command has been taken to mean a ban on making images of other gods, but originally the prohibition applied to representations of Yahweh himself. Worship in the ancient world was unthinkable without some idol or image; therefore, the uniqueness of Moses' restriction is all the more evident. Yahweh is the unimagable Deity who cannot be represented in material forms. Since Yahweh had revealed the meaning of his name to Moses, it was fitting that the Decalogue should also prohibit any magical or unethical use of his name. Undoubtedly the ideas underlying the other commands came from the religious culture of his day, but they were raised to a significantly higher level because of the holy, righteous character of Yahweh. Moses realized that, if the Covenant people were to have a stable, just society, they would have to emulate their God. Concern for his creatures would mean respect for them as persons. Murder, adultery, theft, lying, and covetousness would never be legitimate because they lead to chaos and breakdown of the community. Moreover, inasmuch as Yahweh had been concerned to protect the powerless Hebrews in Egypt, they in turn would have to guarantee justice for the orphans, widows, resident aliens, and any other disadvantaged persons under their jurisdiction.

Interpretations and elaborations

On confirmation of the Covenant, Moses and the people faced the task of living by the stipulations. This called for interpretations of the commands, and so Moses began issuing ordinances for specific situations. Many of these he drew from the case law of his day, but insight as to their selection and application probably came in the "tent of meeting" (a simple sanctuary tent pitched outside the camp), where Yahweh spoke to Moses "face to face, as a man speaks to his friend." Breaches of the Covenant necessitated means of atonement, which in

turn meant provision of a priesthood to function at sacrifices and in worship. In short, the rudiments of the whole Hebrew cult, according to tradition, originated at Sinai. At Jethro's suggestion Moses instituted a system of judges and hearings to regulate the civil aspects of the community. It was at Sinai, perhaps, where the people were organized into twelve tribes.

One of Moses' most remarkable characteristics was his concern for the Hebrews, in spite of their stubborn, rebellious ways. When they reverted to worshipping a young bull (probably made of limestone, instead of gold, as the biblical text says), Yahweh was ready to disown them and begin anew with Moses and his descendants. Moses rejected the offer, however, and later, when pleading for the forgiveness of the people, he even asked to have his own name blotted out of Yahweh's book of remembrance if the Lord would not forgive them.

From Sinai to Transjordan. After leaving Mt. Sinai, Moses faced increasing resistance and frustration, according to the narrative in the book of Numbers. Apparently his virility did not diminish during these years because he took a Cushite woman as his second wife. But Miriam, with the support of Aaron, opposed the marriage. At Kadesh-barnea the pessimistic majority report of the spies who had been sent out to reconnoitre thwarted Moses' desire to march north and conquer the land of Canaan. When he urged the people to reconsider their action they almost stoned him. But here again, according to tradition, Moses interceded for the people with Yahweh, who threatened to destroy them and raise up another and greater nation. In one instance, however, tradition recalled that Moses' anger overrode his compassion. At Meribah, probably in the area of Kadesh-barnea, Moses addressed the complaining people as rebels and struck a rock twice in anger, whereupon water flowed forth for the thirsty people. He had been angry before in defense of Yahweh's name, honour, and cause, but this time his anger stemmed from utter frustration with his contentious people. Although tradition interpreted this lapse as the reason why Yahweh would not permit Moses to enter Canaan, the remarkable fact is that Moses was able to bear up under such continuous pressure.

In Transjordan the new states of Edom and Moab, vassals of the Midianites, rejected Moses' request for passage. He wisely circled east of them and moved north to conquer Sihon, king of the Amorites, and Og, king of Bashan. After Moses permitted some of the tribes to settle in Transjordan, it evoked opposition from the Moabites and their Midianite overlords. They hired the Syrian diviner Balaam to put a curse on the Hebrews, but instead he pronounced a blessing. Some scholars interpret this as proof that Balaam was a convert to Yahwism. If this be true, he must have reverted later on, because the biblical tradition implies that Balaam incited his former employers to weaken the Hebrews by religious seduction. Moses responded to the enmity of the Midianites with a successful holy war against them shortly before his death.

Final conflict against the Midianites

As his last official act Moses renewed the Sinai Covenant with those who had survived the wilderness wanderings. From his camp in the Jordan Valley, Moses climbed to a vantage point on Mt. Pisgah. There he viewed the land of promise. The Hebrews never saw him again, and the circumstances of his death and burial remain shrouded in mystery. Tradition claimed that Yahweh buried him in the valley opposite Beth-peor, the shrine of the people's apostasy.

MOSES THE MAN

Although time undoubtedly enhanced the portrait of Moses, a basic picture emerges from the sources. Five times the narratives claim that Moses kept written records (Ex. 17:14; 24:4; 34:27–28; Num. 33:2; and Deut. 31:9, 24–26). Even with a generous interpretation of the extent of these writings, they do not amount to more than a fifth of the total Pentateuch; therefore, the traditional claim of Mosaic authorship of the whole Pentateuch is untenable. Moses formulated the Decalogue, mediated the Covenant, and began the process of rendering and codifying supplemental interpretations of the Covenant stipulations. Undoubtedly he kept some records, and they served as the core of the growing corpus of law and tradition. In a general sense, therefore, the first five books of the Hebrew Bible can be described as Mosaic. Without him there would have been no Israel and no collection known as Torah.

Moses was a gifted, well-trained person, but his true greatness was probably due to his personal experience of and relationship with Yahweh. This former stammering murderer understood his preservation and destiny as coming from the grace of a merciful Lord who had given him another chance. Moses had an understanding spirit and a forgiving heart because he knew how much Yahweh had forgiven him. He was truly humble because he recognized that his gifts and strength came from Yahweh.

The multiple roles of Moses

Because of the uniqueness of his situation, Moses had to function in a number of roles. As Yahweh's agent in the deliverance of the Hebrews, he was their prophet and leader. As mediator of the Covenant, he was the founder of the community. As interpreter of the Covenant, he was an organizer and legislator. As intercessor for the people, he was their priest. Moses had a special combination of gifts and graces that made it impossible to replace him. Although his successor, Joshua, and the priest Eleazar, the son of Aaron, tried to do so, together they did not measure up to him. Later prophets were great men who spoke out of the spirit that Moses had, but they were not called to function in so many roles. As tradition claimed, he was indeed the greatest of the prophets, and, as history shows, few of mankind's great personalities outrank him in influence.

BIBLIOGRAPHY. WILLIAM F. ALBRIGHT, *From the Stone Age to Christianity*, 2nd ed. with new introduction (1957), classic synthesis of Israel's history and religion in the setting of the ancient Near East (see pp. 11–17, 200–272); *The Biblical Period from Abraham to Ezra* (1963), a popular historical survey (see pp. 1–23); *Yahweh and the Gods of Canaan* (1968), a technical analysis contrasting Israelite and Canaanite region (see pp. 64–109, 153–182); ALBRECHT ALT, *Kleine Schriften zur Geschichte des Volkes Israel*, 2nd ed. (1959; Eng. trans., *Essays on Old Testament History and Religion*, 1967), see the classic article, "The God of the Fathers," pp. 1–86; DEWEY M. BEEGLE, *Moses, the Servant of Yahweh* (1972), a wide-ranging account for laymen and students; WALTER BEYERLIN, *Herkunft und Geschichte der ältesten Sinaitraditionen* (1961; Eng. trans., *Origins and History of the Oldest Sinaitic Traditions*, 1965), technical study of biblical sources in Exodus 19–20, 24, 32–34; JOHN BRIGHT, *A History of Israel*, 2nd ed. (1972), a standard work mediating scholarly extremes; MARTIN BUBER, *Moses, the Revelation and the Covenant* (1946), a sympathetic treatment with philosophical emphasis but weak in details of ancient Near East; GEORGE W. COATS, *Rebellion in the Wilderness* (1968), a detailed critical study of the murmuring motif in the wilderness traditions; FRANK MOORE CROSS, JR., "Yahweh and the God of Patriarchs," *Harvard Theological Review*, 55:225–259 (1962), a scholarly treatment of issues raised by Alt's classic article; DELBERT R. HILLERS, *Covenant: The History of a Biblical Idea* (1969), an excellent popular study; GRETA HORT, "The Plagues of Egypt," *Zeitschrift für die alttestamentliche Wissenschaft*, 69:84–103 (1957) and 70:48–59 (1958), indicates the historical basis of the plagues; YEHEZKEL KAUFMANN, *The Religion of Israel*, translated and abridged by MOSHE GREENBERG (1960), considers Moses the founder of Israel and its religion but doubts details of the historical tradition—neglects much modern archaeological and linguistic research; GEORGE E. MENDENHALL, "The Mask of Yahweh," in *The Tenth Generation: Essays in Early Biblical History* (1972); MURRAY LEE NEWMAN, JR., *The People of the Covenant* (1962), a popular study of Israel from Moses to the monarchy (see pp. 13–101); MARTIN NOTH, *Geschichte Israels* (1956; Eng. trans., *The History of Israel*, 2nd ed. rev., 1960), a basic study, but treatment of Hebrew history prior to the conquest is very radical (see pp. 110–138); *Überlieferungsgeschichte des Pentateuch* (1948; Eng. trans., *A History of Pentateuchal Traditions*, 1972), a technical study of the biblical sources in the Pentateuch—doubts the accuracy of many of the truths presented there (see pp. 156–188); H.H. ROWLEY, *From Joseph to Joshua* (1950), a serious attempt to relate the data of biblical traditions to the findings of archaeology; J.J. STAMM and M.E. ANDREW, *The Ten Commandments in Recent Research*, 2nd ed. rev. (1967).

(D.M.Be.)

Mössbauer Effect

The Mössbauer effect is a nuclear process permitting the resonance absorption of gamma rays. It is made possible by fixing atomic nuclei in the lattice of solids so that energy is not lost in recoil during the emission and absorption of radiation.

The Mössbauer effect is one of the important discoveries of the 20th century and has become as useful a scientific tool as nuclear magnetic resonance apparatus or the laser. Since the announcement of its discovery by Rudolf L. Mössbauer in 1958, more than a thousand research papers making use of the principle have been published in fields as diverse as physics, chemistry, metallurgy, and biology. In order to understand the basis of the Mössbauer effect and appreciate its significance it is necessary to understand several fundamental principles.

Doppler shift

General principles. The first is the Doppler shift. When a locomotive whistles, the frequency, or pitch, of the sound waves is increased while the whistle is approaching a listener and decreased as the whistle recedes. The Doppler formula expresses this change, or shift in frequency, of the waves as a linear function of the velocity of the locomotive. Similarly, when the nucleus of an atom radiates electromagnetic energy in the form of a wave packet known as a gamma ray photon it is also subject to the Doppler shift. The frequency change, which is perceived as an energy change, depends on how fast the nucleus is moving with respect to the observer.

The second concept, that of nuclear recoil, may be illustrated by the behaviour of a rifle. If it is held loosely during firing, its recoil, or "kick," will be violent. If it is firmly held against the marksman's shoulder, the recoil will be greatly reduced. The difference in the two situations results from the fact that momentum (the product of mass and velocity) is conserved: the momentum of the system that fires a projectile must be opposite and equal to that of the projectile. By supporting the rifle firmly, the marksman includes his body, with its much greater mass, as part of the firing system and the backward velocity of the system is correspondingly reduced. An atomic nucleus is subject to the same law. When radiation is emitted in the form of a gamma ray, the atom with its nucleus experiences a recoil due to the momentum of the gamma ray. A similar recoil occurs during absorption of radiation by a nucleus.

Resonance absorption

Finally, it is necessary to understand the principles governing the absorption of gamma rays by nuclei. Nuclei can exist only in certain definite energy states. For a gamma ray to be absorbed its energy must be exactly equal to the difference between two of these states. Such an absorption is called resonance absorption. A gamma ray that is ejected from a nucleus in a free atom cannot be resonantly absorbed by a similar nucleus in another atom because its energy is less than the resonance energy by an amount equal to the kinetic energy given to the recoiling source nucleus.

The Mössbauer effect may be defined as recoil-free gamma ray resonance absorption. As a graduate student at the Max-Planck-Institut in Heidelberg, while investigating the scattering of gamma rays from iridium-191 at liquid air temperature, Mössbauer discovered that he had achieved resonant absorption. By using atomic nuclei fixed in the lattice of a crystalline solid so that the whole mass of the solid was involved in the momentum transfer, the recoil energy loss of the nuclei had become negligible. It was Mössbauer's perception of the mechanism that leads to the absence of recoil and Doppler shift that constitutes the essence of his scientific contribution.

Principles of application. By imbedding the same kind of nuclei in two solids, those in one solid being in a radiating or excited state and those in the other being in an absorbing or ground state, it is possible to use gamma radiation from the first solid for resonant absorption in the second. The method is so sensitive that if one solid moves relative to the other with a velocity as small as a tenth of a centimetre per second the resonance is destroyed by the Doppler shift in the gamma ray energy. In this fact lies the usefulness of the Mössbauer effect as a tool of experimental physics. If an energy state of the radiating or absorbing nucleus is disturbed by any external effect, resonance absorption will also be nullified. Then, by varying the relative velocity of the radiating and absorbing solids until resonant absorption is re-established, the precise magnitude of the disturbing effect can be determined.

Apparatus. The Mössbauer effect is usually observed by measuring the transmission of gamma rays from a radioactive source through an absorber containing the same isotope in the ground state. The device used for this experiment is termed a Mössbauer effect Doppler-velocity spectrometer because it utilizes the Doppler shift to scan a region of energy in the vicinity of the energy of the gamma ray itself. A spectrometer utilizing this concept is shown schematically in Figure 1. A radioactive source is

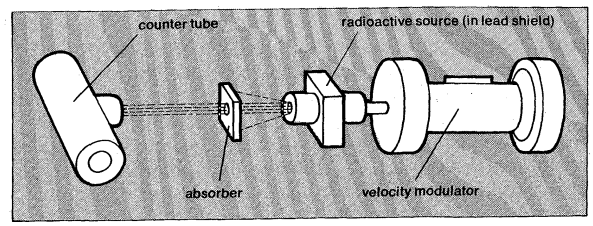


Figure 1: Spectrometer utilizing Mössbauer effect concept. Effect is usually observed by measuring transmission of gamma rays from radioactive source through absorber containing resonant isotope.

mounted on an electromechanical transducer that moves with a well-defined velocity. Generally, velocities of a few centimetres per second are required. The energy of the emitted gamma rays is shifted, due to the Doppler effect, by an amount proportional to the velocity. In this example, nuclei of the radioactive source and the resonant absorber are identical, not having been perturbed by an outside field. The gamma rays pass through an absorber containing the resonant isotope and are detected by a proportional counter. The gamma rays detected per second are plotted as a function of the Doppler velocity, resulting in a Mössbauer effect absorption spectrum like that shown in Figure 2. The drop in counting rate in the

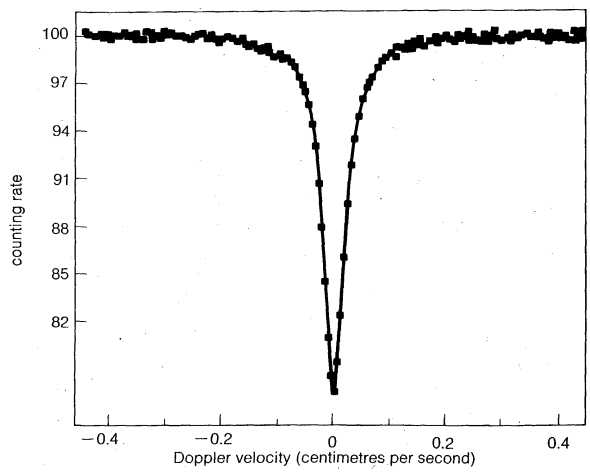


Figure 2: Plot of intensity of gamma ray flux passing through absorber as function of Doppler velocity applied to source. Dip in centre is caused by nuclear resonant absorption (see text).

centre is due to resonant absorption; *i.e.*, the Mössbauer effect. At high positive or negative velocity, the resonant absorption has been destroyed by the Doppler shift.

Isotope suitability. The Mössbauer effect has been observed in more than 35 isotopes. Suitable isotopes must have a stable or long-lived ground state and a low-lying excited state that decays to an appreciable extent by gamma ray emission. The condition on the energy of the excited state arises from need to achieve recoil-free gamma ray emissions. Only if the free-atom recoil energy is small compared to the characteristic lattice vibration energy will there be a finite probability that momentum conservation is satisfied by the recoil of the crystal as a whole with negligible loss of energy. The resulting gamma

ray has the full energy of the nuclear transition and is not broadened by thermal vibrations. It consequently has the exact energy for recoil-free resonant reabsorption by a nucleus in its ground state. The condition on the magnitude of the free-atom recoil energy generally restricts the Mössbauer effect to gamma rays with energy less than 150 KeV. The natural width of a resonance or nuclear level, Γ, is related to its lifetime, τ, by $\Gamma\tau = h/2\pi$, in which h is Planck's constant. The lifetime of the excited state must be greater than 10^{-11} second; otherwise the resonance level would be so broad that the distinction between ordinary and recoil-free emission is lost. The longest useful lifetime is about 10^{-5} sec because the corresponding width is comparable to that due to other mechanisms.

Applications. Mössbauer's discovery has had far-reaching consequences because it has made available electromagnetic radiation (gamma rays) whose frequency is more precisely defined than any other known to date and has provided a new technique for measuring the interaction of nuclei with their environment. Energy resolution better than one part in 10^{12} has been achieved with recoil-free gamma rays. This has made possible experiments in areas as diverse as chemistry, nuclear and solid-state physics, and relativity. Mössbauer received the Nobel Prize in 1961 for his work.

Resolution

The Mössbauer effect has found application in many areas of science. In relativity studies the high precision with which the energy of the gamma ray can be measured has made possible a direct demonstration of the gravitational red-shift; *i.e.*, the change in the energy of a quantum of electromagnetic radiation as it moves through a gravitational field. This was accomplished by measuring the Doppler shift required to compensate for the change in the energy of the gamma ray resulting from a change in vertical position of 2,260 centimetres through the earth's gravitational field. The measured change amounted to 2.5 parts in 10^{15} (a Doppler velocity of 2.7 millimetres per hour) and was in close agreement with theoretical predictions. That is, it was found that a photon of energy E behaves as though it had a mass of E/c^2, in which c is the velocity of light. In a related experiment it was found that the energy of the recoil-free gamma ray decreases with increasing source temperature. This thermal red-shift can be interpreted in a number of ways. It may be viewed as a second-order relativistic Doppler effect resulting from the mean square velocity of the atoms in the solid; *i.e.*, the thermal motion. On the other hand, it can be viewed as a direct demonstration of the relativistic time dilation; *i.e.*, the slowing down of the clock in a moving coordinate system, here the atom, when viewed by a stationary observer. It has been argued that the Mössbauer effect thermal red-shift provides direct experimental resolution of the famous twin paradox of relativity by showing that a space traveller will be younger upon return to Earth than his stay-at-home twin.

Nuclear physics. Applications in nuclear physics are manifold. The Mössbauer effect makes possible the direct measurement of a gamma ray line width, which corresponds to the width of the decaying nuclear level. Results are in close agreement with the measured decay time, indicating that the width of recoil-free gamma rays in fact is determined entirely by the lifetime of the decaying state. The isomer shift, the change in the energy of a nuclear gamma ray due to the electrostatic interaction between nuclear and electronic charge, provides a measurement of the change in the nuclear-charge radius when the nucleus is raised to an excited state. The splitting of nuclear levels into hyperfine components by electric field gradients in crystals of low symmetry or by magnetic fields in ferromagnets makes possible the measurement of nuclear electric quadrupole and magnetic dipole moments. Both isomer shifts and hyperfine structure splittings are readily resolved in Mössbauer spectra. The energy width of a Mössbauer resonance provides a direct measurement of the width of the excited state involved in the gamma ray emission and absorption process. From the width, the lifetime of the excited state can be directly obtained.

Solid-state physics. Applications in solid-state physics fall broadly into the categories of lattice dynamics and hyperfine interactions, although contributions have been made in other areas. The probability that a gamma ray emission process will be recoil free depends on the amplitude of the thermal vibrations compared to the wavelength of the gamma ray. A measurement of the fraction of emission events that are recoil free as determined by their ability to be resonantly absorbed provides the mean square amplitude of the thermal motion in the solid. Using single crystals, the amplitude of the motion can be measured in specific crystallographic directions, providing a stringent test of lattice dynamical models. The thermal red-shift mentioned above gives in addition the mean square thermal velocity.

Magnetic hyperfine interactions have been particularly useful in the study of magnetically ordered materials; *i.e.*, ferromagnets, ferrimagnets, and antiferromagnets. The hyperfine interaction gives an indirect measure of the magnetization of the lattice of magnetic ions and has been used to elucidate the details of magnetic interactions as well as their temperature dependence.

Chemistry. Applications in chemistry are based largely on the isomer shift and quadrupole splitting. The former measures the s-electron charge density at the nucleus and gives information on the character of chemical bonds; *e.g.*, valence and covalence. The latter is sensitive both to the symmetry of the structural environment and to the wave function of the outer electrons of the atom containing the Mössbauer nucleus. The technique has been applied to the study of metal-organic compounds of tin and iron, including hemoproteins; to inorganic compounds of iron, tin, iodine, and the rare earths; as well as to clathrates, catalysts, and glasses containing dilute Mössbauer isotopes.

Theory. The emission of a gamma ray by a nucleus is governed by the physical laws of conservation of momentum and energy. The former requires that the recoil momentum of the nucleus, p, be equal to that of the gamma ray, E/c, in which c is the velocity of light. This gives the free-atom recoil energy, $E_R = p^2/2m$, as being $E^2/2mc^2$. Conservation of energy then requires that the energy of the gamma ray be equal to the energy available in the nuclear transition minus the recoil energy.

Energy calculation

The fraction of total energy lost in recoil is somewhat small, ranging from two parts in 10^8 to one part in 10^6 for isotopes that can be used in Mössbauer experiments. This recoil-energy loss is, however, much greater than the natural width of resonant absorption, so that gamma rays emitted from the nuclei of free atoms with recoil are not resonantly absorbed. Characteristic values of the fractional width Γ/E for natural widths range from one part in 10^{14} to one part in 10^9. The recoil-free process that is fundamental to the occurrence of the Mössbauer effect takes place only for atoms bound in solids and viscous liquids. In the classical sense, one may think of momentum conservation as being satisfied by the recoil of a macroscopic piece of solid. The recoil-energy loss, being inversely proportional to the mass of the recoiling entity, then is vanishingly small.

A more accurate picture of the origin of the recoil-free process is obtained within the domain of quantum mechanics. An atomic solid is pictured as a complex vibrational system consisting of elastically coupled masses in a regular three-dimensional array. Such a system is known to have a well-defined vibrational spectrum. An impulse applied to an individual atom, such as the recoil due to gamma ray emission, can in general change the state of excitation of the solid only by discrete amounts or phonons. Phonons are the quanta corresponding to the intervals between allowed states of the vibrational spectrum. A detailed evaluation shows that as the magnitude of the impulse is reduced, the probability increases that the quantum state of the solid will remain unchanged. These so-called zero-phonon events correspond to the recoil-free events that give rise to the Mössbauer effect.

The probability, f, for recoil-free emission depends on the amplitude of the thermal motion of the emitting atom compared to the wavelength of the gamma radiation, λ:

Recoil-free emission

$$f = \exp\left[-4\pi^2\overline{x^2}/\lambda^2\right]$$

in which $\overline{x^2}$ is the mean square vibrational displacement in the direction of gamma ray emission. This expression can be evaluated for specific models of lattice vibrations. Using the model introduced by Peter J.W. Debye in his theory of heat capacity of monatomic solids, the recoil-free fraction has the form

$$f = \exp\left[-\frac{E_R}{k\Theta_D}\left(\frac{3}{2}+\frac{\pi^2T^2}{\Theta_D{}^2}\right)\right];\ T << \Theta_D,$$

in which Θ_D is the Debye temperature, T is the temperature of the solid, E_R the free-atom recoil energy, and $k\Theta_D$ the energy of the most energetic phonon that can exist in the lattice. Examples are known in which f is very close to unity, making the Mössbauer effect a prominent part of the gamma ray absorption process.

The most important property of the recoil-free gamma rays is that their width is entirely determined by the width of the decaying nuclear state. Doppler broadening by the thermal vibrations of the emitting atom is completely absent in the recoil-free part. It has the expected value, however, when the entire emission spectrum is considered. It is interesting to note that in the years before Mössbauer's discovery, resonant scattering of nuclear gamma rays that *had* suffered a recoil energy loss had been repeatedly studied by using various techniques to compensate for energy shift. These experiments were subject to thermal Doppler broadening and did not provide the high resolution that is the main reason for the importance of the Mössbauer effect.

BIBLIOGRAPHY. The discovery of the effect was reported by R.L. MOSSBAUER in papers that presented evidence for the existence of the effect, its theoretical interpretation, as well as a discussion of its possible applications. His papers (all in German) may be found in these journals: *Z. Phys.*, 151:124–143 (1958); *Naturwissenschaften*, 45:538 (1958), and *Z. Naturf.*, 14a:211–216 (1959). His Nobel Lecture, "Recoilless Nuclear Resonance Absorption of Gamma Radiation," was reprinted in *Science*, 137:731–738 (1962). Introductory treatments are given in H. FRAUENFELDER, *The Mössbauer Effect* (1962); G.K. WERTHEIM, *Mössbauer Effect: Principles and Applications* (1964); and in V.I. GOLDANSKII, *The Mössbauer Effect and Its Applications in Chemistry* (1964). More detailed treatment of selected topics are available in V.I. GOLDANSKII and R.H. HERBER (eds.), *Chemical Applications of Mössbauer Spectroscopy* (1968); and in I.J. GRUVERMAN (ed.), *Mössbauer Effect Methodology*, 4 vol. (1965–68).

(G.K.W.)

Motion-Picture Industry

The term motion-picture industry as it is generally used refers to the *theatrical* motion-picture industry, which produces works of dramatic art or of mass entertainment for national and international markets.

Scope of the industry

Although the industry was once associated primarily with Hollywood, theatrical film production there had come to a virtual standstill in the late 1960s. Asian nations produced more films, especially Japan, which made some 700 films a year, and India, which made about half that many. American-financed and produced films, however, continued to dominate the screens of most of the world. In each of the nations of Western Europe, for example, U.S.-backed films earned far greater revenues than the films of any other nation; and in many places, such as the Scandinavian countries, they earned more than domestically made films. From the 1920s onward, American films so monopolized the international film market that severe import restrictions were imposed on them virtually everywhere to protect domestic film industries. The American patterns of production, distribution, and exhibition, however, were generally adopted by the film industries of other nations, and this article focusses on these patterns.

In addition to the theatrical motion-picture industry, there is an industrial motion-picture industry, which produces films under contract to a single buyer-user, typically a large corporation or institution, which may utilize the film either for internal purposes or for external purposes, such as consumer-product advertising or public relations. There is also an educational film industry, which also produces, under contract, material to be used in government or private classrooms, in technical training facilities, or for teaching and vocational aids for the underprivileged. Another part of the motion-picture industry as a whole consists of the basic raw material and equipment companies, which develop and refine products marketed to all industrial and private customers. Finally, there is an amateur and semiprofessional market that produces personal films, some of which receive public exhibition, such as the so-called underground films often containing artistic, social, or political protest. Individual experimentalists, however, often gravitate to the theatrical film industry.

THE INDUSTRY AND SOCIETY

From the time films began telling a story in the first years of the 20th century, their enormous potential influence on mass audiences has been recognized. In times of crisis the national film industries everywhere have been mobilized to mold public opinion. When Germany was cut off from the major sources of films in World War I, for example, the German government considered films important enough for purposes of morale and propaganda to establish its own industry, despite severe wartime scarcities. Later, under the regime of Adolf Hitler, outright political propaganda again became its major aim in films such as *Triumph of the Will* (1936) by Leni Riefenstahl. At various times the influence of the film has been thought so powerful as to require control by society, usually through censorship. Nonetheless, there are many who feel that films do not influence the behaviour of their audiences as much as they reflect it. This question lies at the root of the unending debate between those who would censor films and those who would have no restraints whatever. Historically, nations with strong central governments have acted as if films caused important changes in those who watch them and therefore must be regulated.

In democratic societies, points of view on the issue of controls vary. Film artists, even though they may control their own films, contend that films mirror life, and not vice versa. Producers, by financing genres as long as the public responds to them and by dropping them when they lose popularity, seem to support the view that audiences determine film content. Many critics, however, attribute a far more causal influence to films in relation to audiences. Film stars are modern folk heroes, who sanction such matters as clothing styles and such habits as smoking or drinking. Like no other medium, magnified pictures and amplified sound give powerful authority to even the most insignificant actions and phrases.

This uncertainty over their power is reflected in the films of all nations. Most films are curious mixtures of appeals to the broadest possible audience and cautious accommodations to official or unofficial pressures. In all nations where films are important, producers have attempted to attract audiences by offering, more than ever before, sensational material, and such material in turn has brought demands for censorship. The fluctuations of the U.S. film industry are illuminating in this regard.

Censorship

Federal censorship was never imposed, but state and municipal censorship was widespread. The industry's first major trade organization, the Motion Picture Producers and Distributors of America, later known as the Motion Picture Association of America, was formed in 1922 in response to public concern over the possible demoralizing effect of films. It fended off criticism successfully until the introduction of sound, after 1926, led to more sophisticated, and therefore more controversial, stories. The Motion Picture Production Code adopted by the industry in 1930 was at first loosely enforced. In 1932–33, as audiences diminished at the outset of the Great Depression, producers tried to lure them back with franker films. Then, under strong public pressure, strict enforcement of the Code prevailed until the 1950s.

When a serious decline in film attendance accompanied the advent of television, producers again turned to bolder screen material. In 1952 the U.S. Supreme Court overruled its 1915 decision that motion pictures were not protected by the free-speech guaranty of the First Amend-

ment to the Constitution. The Production Code was modernized in 1956 and overhauled in 1966, but its power had waned. A film-classification system was introduced in 1968, similar to one in effect in Britain, but failed to check either the production or the attendance of films exploiting sex and violence. The *X* rating, which was intended to warn that persons under 17 should not attend, seemed to become a special attraction to a large part of the public. When the Swedish film *I Am Curious (Yellow)* was finally shown in the United States in 1969 after several widely publicized trials and appeals, it quickly earned millions of dollars despite its *X* rating, its mediocre reviews, and its Swedish political orientation. Public interest in major-studio sex-oriented films, however, proved to be as unpredictable as for other types of film.

The pattern of censorship was repeated in other nations throughout the world. To protect their own film industries against the appeal of the frankness of imported films, the United Kingdom, Italy, West Germany, India, Japan, and other nations carefully re-examined their previous attitudes toward censorship in the late 1960s. Accompanying the trend toward more explicit portrayals of sex and violence was the increase in films of social and political protest. Again, there were those who found such films a significant advance in the maturity of the film art, as well as many who saw in them the potential threat of manipulating mass audiences toward questionable political ends.

With the relaxation of the Code, however, it became evident that the limitations imposed by the Code alone had not been responsible for the dullness and mediocrity of so many films. Mass production of films had induced a tendency toward the path of least resistance. It was artistically easier and economically safer to repeat previously successful situations than to create better films.

ECONOMICS OF THE MOTION-PICTURE INDUSTRY

The theatrical motion-picture industry tends to be highly speculative, with control resting in the hands of entrepreneurs. Even in nations with a high degree of state control, film making involves high risks and huge expenses; and investments usually are made only in those who have proved their ability to create popular successes or to achieve other goals of the government.

The U.S. film industry was for more than 50 years the primary supplier of the world's most popular screen entertainment, usually relegating to a distant secondary position other national film industries. By the mid-1920s, for instance, U.S. films had captured 90 percent of the British market and over two-thirds of the French and Italian markets. Thus U.S. industry practices set patterns in the primary areas of production (manufacturing), distribution (wholesaling-marketing), and exhibition (retailing).

Production. By 1915, film manufacturing became based in Hollywood, a suburb of Los Angeles. As they evolved in the 1920s and 1930s, the large Hollywood film companies such as Metro-Goldwyn-Mayer, 20th Century-Fox, Paramount, Warner Brothers, and RKO typically owned studios presided over by a corporate officer in charge of production, such men as Irving Thalberg, Louis B. Mayer, or Darryl F. Zanuck. Planning was based on selected stories, or "properties," to be filmed in the course of a year. The season's "product" was balanced to provide what seemed to be a proper mixture of subject matter, costs, and broadest possible audience appeal.

The Hollywood studio pattern

The most ambitious films, in terms of cost and creative effort, were called "A" films; "B" and "C" films represented lesser financial efforts. In addition, shorter films, of 10–20 minutes' duration, devoted to news events, sports, travel, music, animated cartoons, and serials, rounded out production schedules.

Production was carried out by the studio's pool of employees, artists, and technicians, some of whom were under exclusive personal-service contracts. Against strong studio resistance, they organized labour unions and guilds in the 1930s to establish minimum working conditions and standards.

A studio would assign its creative talent to one film after another on a basis of overall production plans and without significant allowance for an artist's personal preferences. Only after a performer or director had achieved fame might his assignments be renegotiated, and not always then. Even after she became a star, Bette Davis, for example, had long and bitter feuds with her studio over her roles. Studios also traded their contract talent with each other for corporate profit under "loan-out" agreements.

The positive effect of this system was the development of a studio's "stock company," recruited from the legitimate theatre or from the ranks of novices and perfected by repeated work. Newer talents emerged from the proving ground of B and C films to prominence in an A product. The more ambitious A films were likely to be adaptations of well-known books and stage plays. The B and C productions were the realm of original screenplays by the studio's writers and, hence, later came to be appreciated more than the A films by cinema purists. In adapting from another medium, studios often modified story lines to suit the artists assigned. Film adaptations of plays permitted greater versatility in settings and action than was physically possible on the stage. Film versions of novels, however, tended to compress characters and simplify characterizations. More educated audiences often deplored film adaptations in comparison with the original material, but mass audiences registered little disapproval.

In the silent film era, story developments not immediately apparent from gesture were made explicit through "title cards" inserted in the film; for export purposes, simple substitutions were made. After sound films were introduced, three methods were used for surmounting language barriers: multiple casts, subtitling, and voice substitution or "dubbing." The first method was least used because it was too expensive to tie up studio sets while separate lingual casts were filmed.

Language barriers

Subtitling involves the printing of dialogue translations at the bottom of the screen while performers are heard in their native language. Dubbing, on the other hand, employs trained voice specialists who synchronize translated dialogue to lip movements. A few performers are multilingual, but most stars have rarely been heard in their natural voices outside their native lands; the same voice substitutes consistently dub the same foreign stars.

Distribution and exhibition. Despite the publicity emphasis on stars, films, and studios, the seat of industry power was distribution; studios were simply manufacturing subsidiaries of worldwide distribution companies. Distribution executives viewed the annual output of a studio as a mixed product line for a single mass audience. Marketing considerations were based on the whole schedule. During periods of great attendance, such as holidays, the most expensive films would be released, and lesser products would fill out the schedule. Occasionally, one of the minor films would become a surprise hit, or, in the industry term, a "sleeper."

In determining patterns of distribution, the large companies gave their films first to their theatre-chain subsidiaries. But each large distribution company was dominant in exhibition only in certain geographical areas. Accordingly, reciprocal arrangements were made with the theatre chains of other distribution companies. Afterward, the films of smaller studios might be acquired. A system of time "clearances" fixed the number of days before the films that had been shown by the company-affiliated theatres might be exhibited by independent theatres in the same area. Distribution executives also initiated promotional campaigns for films, including personal appearances by stars at major premieres.

Films were distributed through a network of "branches" or "exchanges" located in major cities. These exchanges supervised sales to theatres, inspection and routing of prints to theatres, and collections of income from exhibitors.

In the decades of growth of the U.S. film industry, other nations instituted legal measures to protect their domestic film industries from annihilation in their competition with American films. Tariffs were placed on U.S. films by some governments, and admission prices were fixed at a higher level for U.S. films than for domestic films. France, Spain, and Italy, after World War II, attempted

Protection against U.S. films

unsuccessfully to compel reciprocity, which would have required their films to be shown in the United States in some proportion to their showings of U.S. films. The two principal methods, however, were a mandatory quota system, whereby a percentage of a theatre's operating time was reserved for domestically made films, and various types of subsidy to domestic film makers.

The exhibition arm of the industry evolved from carnival peep shows in the 1890s, through makeshift locations such as funeral parlours, to theatres originally built for stage plays. With growth, centralized booking and business methods were imposed. Luxuriously appointed new theatres were built, such as the 6,000-seat Roxy in New York City, built in 1927 at a cost of $8,000,000.

Apart from rewiring for sound films after 1926, no major innovation occurred in exhibition until the introduction in 1939 of the outdoor, drive-in theatre. Once located in remote suburban areas, these theatres flourished with population expansion and, after World War II, began to book new films. Attendance at drive-in theatres rose to about one-quarter of the U.S. total movie attendance. Portable exhibition apparatus also serves remote population areas in many countries.

Dispersal of theatre chains

Because of the tight control exerted by major film companies over exhibition, the U.S. government in 1938 brought suit against the large companies for restraint of trade. Not until the 1949–53 period, however, were they forced to dispose of their theatre subsidiaries. They were also forced to abandon the long-standing practice of "block-booking"; that is, selling an entire season's output of films on one contract. Outside the U.S., theatre ownership and block-booking remained standard practices of some U.S. companies.

Following the enforced sale of theatre chains and facing steadily declining audiences because of television, the large studios could no longer keep great numbers of artists and technicians under contract. Major talent agencies such as Music Corporation of America and some individual agents, whose clients no longer had employment guaranteed by long studio contracts, began "packaging" important film projects, which typically involved many of their clients. These packages were then sold to large film companies, which remained the source of production financing and the means of worldwide distribution. United Artists, Columbia, and Universal in this period became stronger competitors to the surviving larger companies.

Losses and gains from television

Television, still the principal competitor of theatrical films, also became an economic ally because of the demand for programming. Film companies eagerly competed to produce program series for television. Some of these corporations maintained financial liquidity by selling films that were too old to be shown in theatres to local television stations for showing in off hours. After 1965 they earned significantly greater revenue from television by renting more recent films for television network exhibition during peak audience hours. Immediately after *The Bridge on the River Kwai* (1958) attracted 60,000,000 television viewers in 1966, CBS-TV paid Metro-Goldwyn-Mayer $52,800,000 for 63 films, and other networks and studios made similar agreements.

Wide-screen processes, often with stereophonic sound, such as Cinerama, VistaVision, Todd-AO, and, especially, CinemaScope, came to be the hope of the surviving studios in the 1950s, since these processes created a product that could not be duplicated adequately by home television equipment.

European films in foreign markets

The trend away from large-studio mass-production film making was not confined to the United States but was worldwide. In Europe, major studio features declined in favour of films that were the personal expression of talented directors. In Italy, Roberto Rossellini and Vittorio de Sica created low-budget but highly artistic films such as *Open City* (1946) and *The Bicycle Thief* (1949), which became better known throughout the world than any previous Italian films had been; and later Federico Fellini and Michelangelo Antonioni achieved even greater international renown. In France, the "New Wave" directors such as Jean-Luc Godard, François Truffaut, and Alain Resnais all began as critics of the products of the French film industry and then in the later 1950s began developing their own distinctive directorial styles using determinedly anti-studio techniques.

A major development in the postwar period was the emergence of an international audience for these "art" films, often in a foreign language. Once the province of smaller independent exhibitors who could not obtain regular products, foreign films attracted growing audiences to the extent that large film companies re-evaluated both the aesthetics and the economics of their product. In the United States in the decade after World War II, imported films increased 233 percent while U.S.-made films declined by 28 percent.

The U.S.-Canadian market was always tempting to the European film industry not only because of the size of the film audience but also because of the large box-office revenues that resulted from the higher ticket prices. In the 1950s European films began to achieve unprecedented successes in this market. In 1958 the French film *And God Created Woman* earned $4,000,000 there, more than twice as much as all French films had earned there the preceding year, and the Greek film *Never on Sunday* in 1961 earned a comparable amount, representing 20 times its reported cost of production. Such films, however, were handled by major U.S. distributors, who were accused of closing the market to all foreign films except the ones they chose to exploit. The British firm of J. Arthur Rank in 1957 made a concerted effort to set up its own distributing organization in the United States, but it failed in two years.

Decline of Hollywood

An overriding economic fact of the 1950s was that because of rising expenses and diminishing audiences, a major Hollywood feature could no longer earn back its cost exclusively from U.S. audiences. Previously, films were directed to U.S. audiences; returns from foreign distribution were "extra." From the 1950s onward, non-U.S. markets had to be taken into account in the production. Similarly, the film industries of other nations also relied on foreign audiences. Disastrous aesthetic consequences sometimes resulted from the producer's fear of offending *any* potential audience and from the increase in huge, international, multilingual productions, with each leading performer from a different nation and sometimes speaking a different language, so that no version was released without some major roles dubbed.

Before World War II, U.S. film companies regularly imported recognized foreign film artists to Hollywood. After the war, however, a combination of factors—foreign currency restrictions, foreign-protectionist measures such as quotas and subsidies, increased costs of filming in Hollywood, and U.S. audience interest in foreign films—caused a major shift in policy. Through long-established foreign subsidiary companies, the U.S. film companies began producing films all over the world. By adroit use of local artists and technicians, these films qualified not only under national quotas but also received national subsidies. As a result, such a film came to that nation's theatre screens with the appeal of an imported U.S. film while enjoying the financial advantages of a domestic film. Such overseas, or "runaway," productions, in addition to the trend toward the realism of films shot on location and away from artificial studio settings, left Hollywood studios practically devoid of feature production.

By 1969–70, the large surviving U.S. companies, retrenching in production, aimed toward a level of 15–20 features per year, about 25 percent of pre-1950 levels. The cutbacks included disposal of burdensome studio ownership and other real-estate operations. Distribution was reorganized to permit fewer sales offices but greater sales impact in key areas. Financing for more independently made films became available from industrial companies and investing groups. Artists deferred high salaries of former times in return for larger film ownership, thus reducing production costs. Exhibitors sought quality films from any source rather than relying only on established companies. In place of the former single-audience production concept, films were made for various segments of the total audience, such as treatments of social and political problems for younger audiences.

The so-called independent producer (who with hardly an

The independent production

exception was financially dependent on major film companies) faced greater opportunities and challenges than before. With a decline in studio production, the independent producer had to arrange details that studio staffs previously handled. These included arranging for production financing, loan collateral, completion bonds (necessary to guarantee a finished film in a releasable form), and laboratory processing; negotiating with creative talent, technicians' unions, and potential distributors; production budgeting; selection of studio rental facilities and exterior locations; and actual supervision of filming. After production is complete, the independent producer must also be conversant with advertising and promotion, distribution practices, and censorship problems peculiar to various countries. To monitor his financial obligations he needs expert legal and auditing advice.

After World War II, the major film-producing nations of Western Europe formed agreements to facilitate international coproductions; that is, films made jointly by teams from two or more nations, which are designed to appeal to audiences of each as much as would a domestically produced work. With the expectation of vastly greater revenues than could be expected from a purely domestic production, larger budgets could be allocated. By the late 1950s, France invested far more in coproductions than in its national films. French-Italian coproductions numbered nearly 800 between 1950 and 1965. Featuring mixed Italian and French casts, some of these films, such as *L'avventura* (1959), *La dolce vita* (1960), and *Last Year at Marienbad* (1962), were masterpieces, but most were ill-coordinated productions, usually filmed in wide screen and colour, that strove only for a fast return on the investment.

Flow of film income. Box-office receipts typically are divided between the exhibitor and the distributor according to a negotiated contract, usually on a variable percentage of the box-office gross. For a popular film, the distributor may receive as much as 90 percent of the gross; thus, when the British-made James Bond series was at its peak of popularity, *Thunderball* (1965) was able to recover its $15,000,000 cost only seven weeks after its release. Advertising expenses are usually split between distributor and exhibitor on a similar basis. Exhibitors retain all income from concessions, which in drive-ins may bring as much as 40 percent of total receipts. That portion of the box-office total remitted to the distributor is known as "film rental" in the U.S. and "film hire" in other English-speaking areas. It is subject to a fixed percentage-distribution fee meant to support the existence of the distribution company. The remainder of the rental is first applied to direct charges for money advanced to the producer by the distributor for prints, advertising, etc., then is remitted to the producer or production company to pay for production costs. These typically include a percentage-overhead charge meant to compensate for intangible costs of corporate administrative and operational support not otherwise directly chargeable to production.

PERIPHERAL ACTIVITIES

Although known primarily for its theatrical films, the motion-picture industry long has been involved in interactions with other industries. Large film companies own music-publishing companies, operate phonograph-record and tape production and distribution companies, and license properties to outside firms specializing in the merchandising of novelty goods. Some of these companies also own television and radio stations.

As music publishers, film companies share with composers in royalties received from song performances. Also, their songs provide inexpensive music for their films, while increasing exposure of the songs. Outside the United States a percentage of theatrical box-office income is allocated to music publishers and composers through performing rights societies that act as collection agents for royalties. The development of phonograph and tape labels as film company subsidiaries permits cross-promotion of film music through sound-track albums and records, as well as the development and exploitation of artists and songs not directly related to any films.

Additional revenue derives from the licensing of company-owned tradenames and likenesses—such as Walt Disney's Mickey Mouse, Donald Duck, and other animated cartoon characters—to merchandising companies specializing in comic books, jewelry, and other consumer novelty items. As customers, film companies regularly acquire motion-picture rights to stage plays, novels, and other published material and properties originally produced on radio or television broadcasts. The prices paid for such rights can be very high: the screen rights for *My Fair Lady* were sold for $5,500,000 in 1962. The justification for such large expenditures is that the properties already are known to the public, hence the possibility of greater film income, which might offset the cost of rights.

Films often increase the artistic reputations and financial status of writers, authors, composers, and other original producers of properties, creating a market for the revival of plays and the republication of books. In fact, novels are sometimes written after a screenplay and published as paperbacks when the film is released. In the 1960s a market for the screenplays themselves developed.

FESTIVALS AND AWARDS

The effects of awards on the motion-picture industry have varied greatly, and the proliferation of awards further diffused their impact. Historically, the U.S. industry, with its overwhelming control of the world's screens, long had virtually no interest in film awards except for its own "Oscars," voted each year since 1929 by the Academy of Motion Picture Arts and Sciences. The Academy, which grew to about 3,000 members representing various artistic and technical disciplines, originally intended its annual awards as modest, peer-group citations within the tightly knit Hollywood industry. After media coverage created widespread interest, however, an increase in box-office revenues for winning films was noticed, and the Oscar became valuable in merchandising. The prestige of its artistic judgments, however, failed to keep pace with its economic power. Serious students of the film tended to place more credence in the awards of the New York Film Critics (founded in 1935), and the later National Society of Film Critics (1966), as well as in the oldest U.S. reviewing organization, the National Board of Review of Motion Pictures (1909). Meanwhile, foreign festivals provided a forum for promotion and recognition of artistic achievements of all other national film industries. The first of these was founded at Venice, Italy, in 1932, with the help of the Fascist regime. It was unique until after World War II when the Cannes Festival was founded. Festivals then assumed increasing importance. Struggling nations, rebuilding their shattered film industries, saw in festivals a chance for world recognition. International political alliances were being reformed, and festivals provided an environment for interaction. And the growing interest everywhere in film imports made the festivals an international marketplace for distributors. Festivals were initiated at Berlin; Moscow; Karlovy Vary, Czechoslovakia; London; San Francisco; Chicago; and New York City. Festivals for short films were held at Edinburgh; at Mannheim and Oberhausen, Germany; and at Tours, France. A Southeast Asia film festival was held annually in different places, and at Vienna, a festival for nontheatrical films was a great success. Through the early 1960s, however, the U.S. film industry remained wary of festivals; films that won festival awards did not necessarily do well at the box office, though a poor reaction at a festival might reduce the audience for a serious film. After 1965, however, distributor attitudes changed as investments mounted in films for specialized audiences, particularly in "foreign" films backed by U.S. companies. Favourable critical acclaim became a vital promotional aid. At the same time, however, the aggressive foreign activities of U.S. companies diminished the value of festivals as a marketplace. Long before festival screening, most quality foreign films had been acquired by U.S. distributors. The past role of festivals also was challenged when the Cannes festival of 1968 became the centre of revolutionary demonstrations. Thereafter, the "glamour" of these events was played down and the awards de-emphasized. The Venice festival

Academy Awards

suspended its awards entirely. Nonetheless, interest in the festivals persisted. In 1970, more than 10,000 persons attended the Cannes festival, and some 300 films were shown; and in a striking reversal of tradition, the top prize went to a U.S. film, *M*A*S*H*, and two other key awards went to films produced by U.S. companies.

BIBLIOGRAPHY. A scholarly approach to the subject may be found in T.H. GUBACK, *The International Film Industry: Western Europe and America Since 1945* (1969). This book analyzes the development of the film industries of western Europe and the U.S. since World War II and presents data on their influence on each other. The author also explores the means the U.S. industry and others have used to circumvent measures to protect domestic industries. Other comprehensive works are TERRY RAMSAYE, *A Million and One Nights* (1926), a two-volume history of the early decades of the industry; and ARTHUR KNIGHT, *The Liveliest Art* (1957), perhaps the best of the contemporary histories, providing both artistic and industrial overviews.

Accurate and readable biographies of industry pioneers by a responsible film journalist include BOSLEY CROWTHER, *The Lion's Share* (1957), a corporate history of Loew's, Inc., parent company of MGM, and *Hollywood Rajah* (1960), a biography of Louis B. Mayer, longtime MGM studio head. Valuable insight into classic studio methods is provided in the trilogy of R.J. THOMAS: *King Cohn* (1967), a biography of Harry Cohn, cofounder of Columbia Pictures Corp.; *Thalberg: Life and Legend* (1969), the story of Irving G. Thalberg, creative assistant to Mayer at MGM; and *Selznick* (1970), embracing the family and career of producer David O. Selznick. An early biography, replete with corporate infighting and financial manoeuvrings far removed from the glamour of film, is SINCLAIR'S *Upton Sinclair Presents William Fox* (1933), a subjective study, but far more credible than the reader might imagine.

The industry is covered by many trade journals, of which *Variety* and *Daily Variety* are considered the most reliable.

(A.D.M.)

Motion Pictures, Art of

A remarkably effective medium in conveying drama and especially in the evocation of emotion, motion pictures consist of the projection of luminous moving images onto a screen. The art of motion pictures is exceedingly complex, requiring contributions from nearly all of the other arts as well as countless technical skills. Nonetheless, probably no other art has proliferated as much in the 20th century nor can any other equal it in popularity or influence.

The newest fine art

The motion picture is the newest of the generally recognized "fine arts." It is the product of photographic and technological developments that culminated by 1900 in practical devices for the recording of a moving image and its projection onto a flat surface. During its early development, the motion picture was discounted by many critics for its subservience to commercial interests, for the immediacy of its appeal to the uninstructed, for its mechanical technique, and for its apparent lack of an identifiable artist as its primary creator. After the middle of the 20th century, however, increasing attention was devoted to it as a form of artistic endeavour that is as legitimate as the theatre, literature, music, or the visual arts.

Throughout its history, the motion picture steadily became more complex. Even the silent film was capable of a wide range of expression, the possibilities of which were far from exhausted when it was superseded by sound films in the late 1920s. Subsequently, the medium was enriched with complex sound tracks, combining a half-dozen separate recordings. Colours and wide screens added new breadth, and improvements in lenses, film stock, and processing techniques permitted trick effects of all kinds.

In an art in which spectacular effects were commonplace, the qualities of simplicity and spareness became virtues. Gradually, it was learned that the music of a motion picture should not demand the attention that symphonic music does, that there should be less dialogue in a film than there is in a play, that the visual composition should be less intricate than that of a painting, that the acting must be more restrained than theatrical acting. To a considerable extent, the motion picture discovered its own identity through its difficulties in adapting elements of the other arts.

This article will analyze the many elements involved in the creation and appreciation of works of art in motion pictures. The article has the following sections:

Essential characteristics of motion pictures
Expressive elements of motion pictures
Style in motion picture
Types of motion pictures
The study and appreciation of motion pictures

A detailed account of the growth and development of motion pictures as an art will be found in the article MOTION PICTURES, HISTORY OF, and its technical and scientific development is treated in MOTION PICTURES, TECHNOLOGY OF. Commercial aspects of motion picture production and distribution are discussed in MOTION PICTURE INDUSTRY.

ESSENTIAL CHARACTERISTICS OF MOTION PICTURES

In its short history, the art of motion pictures has frequently undergone changes that seemed fundamental, such as that resulting from the introduction of sound. It exists today in styles that differ significantly from country to country and in forms as diverse as the documentary created by one man with a hand-held camera and the multimillion-dollar "epic," involving hundreds of performers and technicians. Despite its diversity, however, an essential unchanging nature can be discerned in most of its manifestations.

A number of factors immediately come to mind in connection with the motion picture experience. For one thing, there is something mildly hypnotic about the illusion of movement that holds the attention and may even lower critical resistance. Also, the accuracy of the motion picture image is compelling because it is made by a non-human, scientific process. And, the motion picture gives what has been called a strong sense of being *present*; the film image always appears to be in the present tense. There is also the concrete nature of film; it appears to show actual people and things.

No less important than any of the above are the conditions under which the motion picture ideally is seen, where everything helps to dominate the spectator. He is taken from his everyday environment, partially isolated from others, and comfortably seated in an air-conditioned auditorium. There, the darkness concentrates his attention and prevents comparison of the image on the screen with the objects or people around him. For a while, he lives in the world the motion picture unfolds before him.

Still, the escape into the world of the film is not complete. Only rarely does the audience react as if the events on the screen were real; for instance, by ducking before an onrushing locomotive in a special three-dimensional effect. Moreover, such effects are considered to be a relatively low form of the art of motion pictures. Much more often, the viewer will expect a film to be true to certain unwritten conventions than to the real world. Sometimes, he will expect exact realism in details of dress or locale, but just as often he will expect the film to escape from the real world and make him exercise his imagination, a demand made by great works of art in all forms. If an indirect statement stimulates the spectator's imagination, he will complete the picture much more vividly for himself than the artist could. Fidelity in the reproduction of details is much less important than the appeal made by the story to an emotional response. If the audience senses the film maker is overstating his case merely to evoke an emotional response, the spell is broken. Between the artist and the audience, there is a constant, changing, unformulated relationship of taste. The bounds of this sort of taste depend on the particular motion picture, the audience, and the occasion, but the relationship is always present and the artist must work within it.

The unwritten conventions

The essential characteristics of the motion picture can be divided into those characteristics that pertain primarily to the motion picture image, those that pertain to motion pictures as a unique medium for works of art, and

those that derive from the experience of viewing motion pictures.

Qualities of the film image. The primary unit of expression in film is the image, or the single shot. The attribution of magical properties to images has a long history; this association is well documented among many primitive peoples, and it is even reflected in the term magic lantern as a synonym for the film projector. Any image taken out of the everyday world and projected onto a screen to some extent appears to become magically transmuted. This magical quality helps to explain the enthusiastic reception accorded such early films as *La Sortie des ouvriers de l'usine Lumière* ("Workers Leaving the Lumière Factory), which were merely photographic records of commonplace scenes in Paris in the 1890s by the French film pioneers, the Lumière brothers.

Intensity, intimacy, ubiquity. One critic singled out the qualities of intensity, intimacy, and ubiquity as the salient characteristics of the motion picture image. Its intensity derives from its power to hold the complete attention of the spectator on whatever bit of reality is being shown. Outside the theatre, a person's attention is usually dispersed in the endless reality around him except for sporadic moments of concentration on what he selects for closer scrutiny. In the cinema he is compelled to look at something that not he but the film maker has selected, for reasons that are not always immediately apparent. This quality of intensity becomes most noticeable when the camera remains fixed on something for a longer time than seems warranted, and the spectator gradually becomes acutely conscious of his loss of volition over his own attention. This technique is not often used but is most effective when used well.

The intimacy of the film image is related to the camera's ability to see things in greater detail than the eye can. This ability is demonstrated in long-distance shots through a telephoto lens as well as in closeups. At the beginning of the Japanese film *Woman of the Dunes* (1964), a pervading theme of the film was indicated by shots of grains of sand many times enlarged.

The impression of ubiquity—being everywhere at once—is achieved in part by the camera's apparent freedom to move from place to place or to approach or withdraw instantaneously. No less important to this illusion of ubiquity is the effect achieved by editing, which allows countless images representing a long, elaborate action, to be presented in a comparatively short film.

Particularity. Other equally important characteristics of the film image may be singled out. One of these is its particularity. The language of words lends itself to generalization and abstraction. In themselves, words such as man or house do not suggest a particular man or a particular house but men and houses in general; and more abstract terms such as love or dishonesty have even less precise associations with specific things. Motion pictures, on the other hand, only show particular things—a particular man or a particular house. In this way a film image may be less ambiguous than the language of words but also less evocative, less likely to be enriched by imagination, association, or recollection. Despite its particularity, however, the motion picture image may also be ambiguous in that it *shows* but does not *explain*. It does not in itself tell what it means, and people instinctively search for meanings in images. This is why commentary is so important in a documentary film. It is also why the juxtapositions of montage are so effective—the spectator compulsively searches for the reason behind that particular sequence of images.

The mind and the eye

Neutrality. Another characteristic of the film image is its neutrality. The world a person sees around him is strongly influenced by his emotions and his interests. A plumber fixing pipes in a museum may not see the masterpieces around him, while an angry man may hear an insult where none was intended. The camera, however, reproduces everything in front of it, without feeling. An important part of any artist's training is to cultivate a neutrality of vision, the ability to see everything fresh; the camera possesses this faculty naturally. When a film appears to be charged with emotion, it is usually because the director has carefully manipulated the images to give this illusion. In everyday life, the eyes follow the mind; in the cinema, the mind follows the eyes.

Characteristics of the medium. Four characteristics may be stressed as factors that differentiate the motion picture medium, either in degree or in kind, from other mediums for works of art: luminosity, realism, movement, and montage.

Luminosity. The intense brightness of the picture projected by powerful light onto a coated screen in itself transforms the most mundane element of reality. The appeal of a luminous picture is attested by efforts of advertisers to achieve luminous effects in posters and displays and in the popularity of viewing still photographs as transparencies rather than as ordinary prints. The luminosity of the motion picture image also results in a considerable range of tone, between the brightest highlight and the deepest black. Both in black and white and in colour films, the most delicate gradations in the image are therefore possible.

Realism. Another essential element of the motion picture image is that it gives an impression of reality. Whether in a drama enacted expressly for the camera or in a documentary film of an event at which the camera just happened to be present, this feeling of realism deriving from motion picture photography accounts for much of the force of motion pictures. Animated films, which lack this element of photographic realism, tend to be taken much less seriously by adults.

The attempt of the motion picture to reproduce three-dimensional reality on a flat screen presents the same problems and opportunities that are encountered in still photography and in painting. Like still photography, cinematography tends to exaggerate perspective, so that effects of great depth, with dramatic contrasts between foreground and background are possible. On the other hand, in Japanese films as in Japanese paintings, there often seems to be no perspective at all.

Movement. As a feature of the motion picture, movement is so obvious that its central importance is sometimes forgotten. As has been noted above, the motion picture has much in common with the graphic arts. The added dimension of movement transforms it, allowing it to unfold a narrative or a drama in time, in a way no other graphic art can. Both in film making and in film appreciation, movement must constantly be borne in mind: composition in the motion picture is kinetic rather than static. It is not a single colour but the cumulative effect that matters, not a single situation but a developing plot. The composition within any frame, or exposure, of a motion picture is less important than the relationship of that frame to those that precede and follow it.

Montage. Perhaps the most essential characteristic of the motion picture is montage, from the French *monter*, "to assemble." Montage refers to the editing of the film, the cutting and piecing together of exposed film in a manner that best conveys the intent of the work. Montage is what distinguished motion pictures from the performing arts, which exist only within a performance. The motion picture, by contrast, uses the performances as the raw material, which is built up as a novel or an essay or a painting, studiously put together piece by piece, with an allowance for trial and error, second thoughts, and, if necessary, reshooting. The order in which the segments of film are presented can have drastically different dramatic effects.

Several major contributions to the theory of the montage were made by Soviet directors. After the Russian Revolution of 1917, Soviet films were encouraged for their propaganda value, but film stocks were scarce. Soviet directors carefully studied the films of D.W. Griffith and other masters to make the most effective use of their own meagre resources. One of these early Russian directors, Lev Kuleshov (1899–), conducted an experiment involving identical shots of an actor's expressionless face. He inserted it in a film before a shot of a bowl of soup, again before a shot of a child playing, and still again before one of a dead old woman. An unsuspecting audience, which was asked to evaluate the actor's per-

Kuleshov's experiment

Through skillful use of montage in this sequence from *Potemkin* (1925) Sergey Eisenstein created a feeling of horror and outrage. The film deals with a mutiny aboard the Russian cruiser "Potemkin" during an abortive antitsarist uprising in 1905. (A) The people of Odessa assembled on the harbour steps to cheer the victorious crew of the "Potemkin." (B) The tsar's Cossacks firing into the crowd. (C) The carnage after the first volleys. (D) The shock and fury of a student onlooker (played by Eisenstein). (E) The relentless advance of the Cossacks. (F) The defiance of a mother holding her slaughtered child.

By courtesy of the Rosa Madell Film Library; photographs, from the Museum of Modern Art Film Stills Archive, New York

formance, praised his ability to express respectively hunger, tenderness, and grief.

Sergey Eisenstein (1898–1948), who excelled both as a director and as a teacher, based much of his theory of film on montage, which he compared to the compounding of characters in Japanese writing. The character for "dog" added to the character for "mouth," he noted, results not merely in "dog's mouth" but in the new concept of "bark"; similarly, film montage results in more than the sum of its parts. Still another great Russian director, Vsevolod I. Pudovkin (1893–1953), also stressed the importance of the carry-over in the spectator's mind. Only if an object is presented as part of a synthesis, he said, is it endowed with filmic life.

Two types of montage have been distinguished—narrative montage, which is intended primarily to tell the story, and expressive montage, which is intended to produce an immediate effect, much like that of a metaphor in literature. It was in expressive montage that Eisenstein was chiefly interested both in his writing and in his films. In his *Strike* (1924) there is a famous example in which a shot of workers being cut down by cavalry is followed by one of cattle being slaughtered.

The motion picture experience. Other distinguishing characteristics of the motion picture are more closely related to the conditions under which they are normally seen. These are: collectivity, impersonality, and accessibility.

(Left) The academy screen tightly frames individual figures and expresses their powerful emotions in this scene from Carl Dreyer's *Passion of Joan of Arc* (1929). (Right) Broad, spectacular crowd scenes are presented well on the CinemaScope screen, as in this frame from *The Robe* (1953), the first film made in the CinemaScope process.

By courtesy of (left) Societe Nouvelle des Etablissements Gaumont, Paris, (right) 20th Century-Fox Film Corp.; photographs, from the National Film Archive, London

Collectivity. The kinetoscope, a peep-show motion picture that could be seen by only one person at a time, became obsolescent soon after it was introduced by Thomas A. Edison in 1894 because it could not compete with projected films, which could accommodate audiences. Since that time, an atmosphere of group entertainment has been associated with films. Even for films being seen on television, this aspect of community viewing remains significant; though the audience before any particular set may be small compared with a theatre audience, each spectator brings to it his past associations of films seen in the theatre and the realization that as he watches, millions of others do so too.

Impersonality. Although the motion picture experience is a group experience, it is exceptional for its impersonal nature. The cinema is quite different from the theatre, in which the audience feels the presence of the actors. The cinema is less inhibiting: people come and go as they please, unnoticed, in an easy, relaxing atmosphere. There is comparatively little feeling of association with the other members of the cinema audience. By contrast, at a sporting event, rivalry and intense partisanship are shared openly by the spectators; at a popular music festival, audience participation may reach a level of hysteria; at a concert of classical music, the demand for quiet and serious listening is deeply felt. The informality of the cinema has probably contributed to the tendency to regard it less seriously than other performing arts.

Accessibility. Another basic feature of motion pictures is their accessibility to an audience of vast size. A drama presented in a theatre could be attended by no more than a few thousand persons at a time. In a film, however, the same drama and the same performers may be seen throughout the world by millions of people almost simultaneously. Accessibility has been a determining characteristic of films. From the start, they were a sidewalk attraction, and with the advent of television in addition to cinemas, motion pictures are available almost anywhere at almost any time. This mass audience allowed the motion picture industry to develop vast production and distribution organizations, but it also limited the artistic development of films, since it was commercially safer to present the simplest and least demanding entertainment that would appeal to the widest audience. After television took over much of this mass audience in the 1950s, the motion picture was taken much more seriously as an art.

Effect of the mass audience

EXPRESSIVE ELEMENTS OF MOTION PICTURES

Many observers have seen in films a means of expression comparable to a language. The French poet and film maker Jean Cocteau (1889–1963), for example, called the cinema "picture writing." The language of motion pictures, however, is not the language of words, even though spoken dialogue has been an integral part of motion pictures since the late 1920s and written captions were usually required to explain the action before that. It is primarily in its images that the language of the cinema must be sought.

Cinematographic expression. The film maker has a number of ways of modifying the camera's neutrality and thereby the "reality" that is conveyed to the audience. It is largely by means of these devices that the motion picture becomes such an expressive medium. Several of these expressive techniques should be emphasized. First, there is framing; that is, carefully selecting what will be included within each frame of the film and what will be excluded. Second, there is scale, the size and placement of a particular object or a part of a scene in relation to the rest; a relationship that is determined by the placement of the camera. Third is camera movement, or the lack of it, during shooting. Fourth, there are the peculiar advantages of either colour or black and white photography that can be exploited. Finally, through the skill of the cinematographer and his knowledge of laboratory processes, other highly expressive techniques can be achieved. Each of these means of expression will be discussed below.

Framing. The process of framing is intended to eliminate what is unessential in the motion picture, to direct the spectator's attention to what is important, and to give it special meaning and force. Each frame of film, which corresponds in shape to the image projected on the screen, forms the basis for a graphic composition in the same way as the frame of a painting encloses the area in which the painting must be organized. Several different ratios of frame height to frame width (called "aspect ratios") have been used in motion pictures. The most common is 1 to 1.33, or 3 to 4, a ratio corresponding to the dimensions of the frame of 35-mm film. By using 70-mm film or a special CinemaScope lens, an image with wider horizontal and shorter vertical dimensions is achieved—a proportion of about 2 to 5, or between 1 to 2.25 and 1 to 2.65. A similar effect, called wide screen, was sometimes achieved without the expensive equipment required for CinemaScope by using 35-mm film and masking the top or bottom, or both, giving a ratio of 1 to 1.75, or 4 to 7. Though most cinemas installed wide screens in the 1950s, the normal or academy screen again became more popular in the next decade.

The aspect ratio

The academy screen offers several advantages in composition over those of greater width. An expansive effect is more easily achieved on the academy screen by photographing small-scale figures in the foreground or background with the sky or hills towering above them. Feats of mountain climbing, the monumentality of trees in a forest, or the oppressiveness of a city's skyscrapers are effectively rendered on the academy screen. Wider screens may be more vivid in expressing the loneliness of a prairie, such as that of the Texas panhandle in which *Hud* (1963) was set, or the vastness of the desert, as in *Lawrence of Arabia* (1962). On the other hand, they cause distinct disadvantages for tête-à-tête, or love, scenes.

Regardless of its ratio, the frame may be divided to

show two or more scenes at the same time. In *The Thomas Crown Affair* (1968), for example, the mastermind of a bank robbery and his accomplices are seen talking to each other by telephone, all at the same time, by means of multiple images.

A most effective use of framing consists of temporarily excluding a vital part of the action to stimulate the spectator's interest. For example, the camera may remain fixed on the hero while the villain is perceived only by a voice saying, "Hands up!" or in a science fiction film, the camera may linger on the horror expressed by the victim before revealing the monster that causes it.

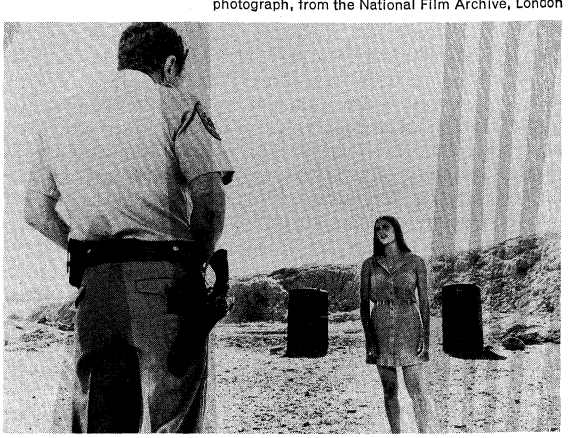
Finally, very strong dramatic effects may be obtained by oblique framing; that is, by turning the camera sideways so that the image on the screen appears askew. This was done in an early Russian film *The Ghost That Never Returns* (1929), in which a prison riot shown by oblique framing gives the impression that the building is being pushed over. A more lyrical effect is seen in *Women in Love* (1969), in which the image of two young lovers walking naked among flowers and trees, is turned sideways, as if they were floating.

Scale. Since scale in the cinema constantly changes from shot to shot, the spectator can easily be deceived about the size of objects. When appearing next to enormous tables and chairs, for instance, the actors can be made to look like midgets or children, as in the Stan Laurel and Oliver Hardy comedy *Brats* (1930). By contrast, in *King Kong* (1933) a small-scale model of New York City gave the illusion of the actual city under attack by a giant gorilla. Scale may have a marked effect on the emotional tone of a scene. In the distance an actor may seem lonely, remote, helpless, pathetic; close up, he may appear powerful, threatening, bestial. The range of shots for artistic purposes is from an extreme long shot (the widest view on the smallest scale), with houses or ships appearing as tiny dots on the horizon, through medium shots, two shots (*i.e.*, a shot of convenient size to include two actors) and others, to the close-up, with part of a face, an eye, or a fist, filling the screen (the most restricted view on the largest scale). Telescopic or microscopic shots beyond these extremes are usually of scientific rather than of artistic interest.

Range of shots

Different scales may be evident in the same shot, as in a scene in *The Train* (1964) in which a dead Nazi looks like a huge but now-harmless monster lying in the foreground while the much smaller figures of the French Resistance fighters escape into the background. In *Zabriskie Point* (1968) there is a shot in which a policeman bulks large in the foreground, a girl stands in the middleground, and the fugitive she is protecting is hidden by a tiny comfort station in the background. As has been noted, the camera exaggerates perspective, and this exaggeration adds to the dramatic effect. It is most striking in an ordinary still photograph of an enormous hand or a sunbather's giant feet that were close to the camera lens. In cinematography, the director ordinarily will minimize the effect of this distortion, but occasionally he will use it in an extreme form. In *Easy Rider* (1969), for instance, it was used to give an atmosphere of hallucination and nightmare to a drug-taking session.

By courtesy of Metro-Goldwyn-Mayer Inc. © 1970; photograph, from the National Film Archive, London



A still from Michelangelo Antonioni's *Zabriskie Point* (1968), in which perspective makes the policeman appear overwhelming in relation to the girl while, at the same time, he looms as an easy target for the fugitive hiding behind one of the structures in the background.

Finally, scale will be affected by what precedes and follows. The close-up will have its most dramatic impact coming after long or medium shots, and after many close ups, it is a relief to escape to the middle or far distance.

Shooting angle. Another element in motion picture language is shooting angle. In common language, the phrases "to look up to" and "to look down on" have connotations of admiration and condescension in addition to their obvious reference to physical viewpoint. In one sense or another, children, dogs, and beggars are often looked down upon; while the preacher in his pulpit, the judge on the bench, and the policeman on his horse are looked up to. Even a slight upward or downward angle of a camera is enough to express a mood of inferiority or superiority. A downward angle gives a more expansive, a freer view; an upward angle gives a shut-in, imprisoned feeling. In both cases, the emotional effect of the shot is increased.

Upward or downward shooting angles lead to questions of objectivity and subjectivity. In most motion pictures, both for variety and for breadth of treatment, the camera's viewpoint will switch from one character to another, and sometimes will be associated with none of the characters but will merely look on. The camera might take the viewpoint of the heroine, looking with dismay at the villain as he breaks into her room; in this case, an upward camera angle will give a subjective impression of her fear. Similar subjectivity may be seen in a shot of buildings reeling in the way they might appear to a drunken man, as in the German classic *The Last Laugh* (1924), or in a rapid camera movement from a window to the pavement below to express a thought of suicide, as in the Italian neorealist film *Umberto D.* (1952). Occasionally, an entire motion picture may be shot from one person's point of view, usually with a "stream of consciousness" commentary, speaking the person's thoughts. A well-known picture in this style is the first American film of the British director Alfred Hitchcock, *Rebecca* (1940), in which the heroine, an innocent young girl who marries a wealthy widower, dramatically uncovers his turbulent past.

The camera's "point of view"

Extreme upward or downward angles are too far removed from ordinary experience to have many applications in motion pictures, but they may express exceptional situations—a sick man on his back, a baby's or a dog's point of view, a man in a pit or in a coffin, a spy covertly looking down on an enemy meeting. As with scale, the shots that precede and follow will alter the effect of the shooting angle. Upward angles will be stronger following a level or downward-looking camera, and vice versa.

Camera movement. Another factor in motion picture language is camera movement. Film makers experimented with camera movement very early in film history (from the 1910s onward), but it was considered artificial, and equipment had not been developed for moving the camera evenly and holding it steady. After the 1930s, however, camera movement was much more frequently used.

The most common movement is to turn (or pan, from the word panorama) the camera horizontally so that it sweeps around the scene. It can also be tilted up or down in a vertical panning shot, or in a diagonal pan, as when it follows an actor up a stairway. Almost as common today is the tracking (or trucking) shot, in which the camera is moved bodily forward toward an object, backward away from it, or sideways along a street, a wall, a beach, a table. The camera may pan while it is tracking, a complex combined movement. A camera movement that is frequently used imitates the gaze of a traveller in a car or train turning to focus on a point of interest as he moves past it. Tracking shots may be made from a camera mounted on a dolly (or trolley), sometimes riding on rails, or from a crane, or boom. The aerial movements

Panning and tracking

Innovative use of multiple images was made by Abel Gance in *Napoléon* (1926). Here the main vista is projected across three screens with three projectors, while on the centre screen additional images of a map and a woman's face are superimposed.
From the National Film Archive, London

of a crane are freer, but the trolley can travel more easily. Cameras may be mounted on automobiles, trains, planes, or balloons. Although a machine-mounted camera gives a steadier effect on the screen, the director may sometimes want the unsteadiness or jerkiness of a hand-held camera. Hand-held cameras are used when standard equipment is too bulky or too expensive. Occasionally, extreme movements are used: the innovative French director Abel Gance in *Napoléon* (1926) had his cameras thrown as snowballs, fired as cannonballs, dropped from a cliff into the sea, fixed to the saddle of a cantering horse, and mounted on a swinging platform.

The effect of camera movement will depend on the context and on the pace of movement. At a deliberate pace the camera can explore a scene and reveal significant details. If it is raised well above the ground the movement has a dreamlike power and, when combined with slow motion, may give a somnolent impression or express recollection or hallucination. The camera movement may end dramatically on a dagger, on a gun half-hidden in an assailant's hand, or on a suspicious bulge in a pocket. It may link the hero walking in the garden and the heroine watching him with loving eyes from a window. It may bring a dramatic surprise as in the American western *Stagecoach* (1939), when director John Ford had the camera, mounted high above a rocky defile, move slowly from the stagecoach below to reveal a band of Indians waiting in ambush. On the other hand, the camera may simply turn away from a scene to leave the remainder to the spectator's imagination, as when it withdraws from a torture scene or from a love scene. In filming a conversation the camera may turn from one speaker to the other, animating the scene with its own movement and showing the expression of the speaker, or listener, more closely than would be possible with a two shot, which would encompass both at once. Camera movement may be used to change the scene to a distant place, to a different period of time, or to an imaginary world. Very rapid camera movements (which are often affected by a zoom lens—one that rapidly brings the image forward to a close-up or makes it recede into the background) may express a sudden surge of emotion or a contemplated action, such as in the suicide example from *Umberto D.* already mentioned. In *The Rains Came* (1939), as the heroine realizes with horror that she has drunk from a glass that may be contaminated with typhus, the camera rushes forward to a close-up on the fatal glass, shining in the darkness.

Zooming

The introduction of camera movement significantly affected the style of motion pictures, enabling the action to be followed more smoothly and continuously than was possible when it was broken up into brief shots.

In extreme cutting styles, the audience may lose the dramatic relationship between shots of brief duration. Shots that last too long, on the other hand, may cause the spectator's attention to wander.

Colour and black and white. Early attempts to obtain colour effects in the cinema were made by using film stock tinted all one shade. Though sepia, for instance, was effective for occasional use, it was in the long run too artificial. In some motion pictures every frame was hand coloured, but that method was too laborious for continued use. More important, black and white proved to be an entirely acceptable medium, with an artistic quality of its own.

A few films were made in the 1920s and earlier, by various colour processes, but a practical commercial system was not perfected until Technicolor was introduced in Walt Disney's animated short *Flowers and Trees* (1932) and in the feature film *Becky Sharp* (1935). The introduction of colour was less revolutionary than the introduction of sound; though the silent film soon disappeared, the black and white film has continued to coexist with colour.

A black and white motion picture is not merely a picture that lacks colour but an artistic creation with positive qualities of its own. An ample range of effects can be obtained—from precise images in which every hair, every grain can be clearly seen, to a smudged charcoal effect. Oddly, black and white often gives a stronger impression of realism than colour does, perhaps because of associations with newsreels and newspaper photographs. Black and white may suggest the hard news of the daily paper, while colour may connote the fantasies of glossy fashion magazines, advertisements, and comic strips. In the cinema, black and white composition is often stronger and more dramatic than colour, cutting can be more forceful, and the coloured image tends to lose depth.

Use of black and white for realism

Nevertheless, colour introduced a new world into the cinema and steadily grew more effective. Colour can have a powerful dramatic effect. In *The Damned* (1969), the Italian director Luchino Visconti used the flaming red of a blast furnace to represent the hellish power of the Nazis, and garish whites and reds in a drunken homosexual orgy ending in a massacre. Colour can be used symbolically—in Eisenstein's *Ivan the Terrible,* Part II (1946), red turning to a bluish shade represents the fear of a pretender about to be assassinated.

Role of the cinematographer. Cameramen remain virtually unknown outside the industry even though the cameraman's contribution to a movie sometimes matches that of the director in importance. Although the director has ultimate control over the visual image, the cameraman actually records it on film, translating the director's ideas and creating the atmosphere and the look of the film. The association between cameraman and the processing laboratory is also of highest importance because he often spends hours there after shooting, checking the negative. On feature films there is a camera team and responsibility is shared.

The cameramen are responsible for exact framing, sometimes for screens of more than one type. They also must decide upon the use of masking, the choice of lens, the camera angle, and the control of camera movement. They must either keep the focus sharp or put all or part of the picture out of focus if this effect is required. The cameraman will also control slow motion or accelerated motion; with early hand-cranked cameras, he simply slowed down or cranked faster, but later special controls and cameras were developed. Trick photography was once effected by simple manipulation of the camera: magical transformations were made simply by stopping it and changing the scene; and the impression of backward

motion was achieved by turning the camera upside down and reversing the film. More elaborate processes are now at the photographer's command involving laboratory technicians as much as the camera crew. Many effects require the actors to perform against a background of previously prepared film. The cameraman must be in command of all these processes. The best of them will give the picture a visual style that is peculiarly their own.

Editing. The process of trimming and piecing together lengths of film in order to make an artistically concise and complete motion picture is probably the most important technique of film language. The terms editing, cutting, and montage are often applied interchangeably to the process. In montage the emphasis is on the juxtaposition of ideas resulting from this process, cutting stresses the physical work with the actual strips of film, and editing encompasses both.

Editing, cutting, and montage

In a single shot (*i.e.*, the film exposed at one time, without interruption, by one camera), a motion picture contains a copy of some part of the real world; by effective editing, this world may be taken apart and put together again in the form of a work of art. Editing permits sudden changes in viewpoint from outside to inside, from one character to another or one place to another, from far away to close at hand. Most editing within a scene is made unobtrusive by correlating it with the action or the dialogue. From scene to scene, editing must establish the change in time or place as the story requires; it resembles the changing of sets in the theatre, though in the cinema changes are far more frequent, and are more like the chapters or paragraphs of a novel. The cut may be made startling rather than unobtrusive—for instance, by a sudden close-up of a strange face. Or the picture may fade out and fade in, the screen being left dark for a moment. Or it may dissolve, or mix, to a new scene, one image showing on top of the other for a moment. The film makers may use other devices, such as a wipe (*i.e.*, a line moving across the screen that wipes out the preceding image while introducing the next); or irising, gradually reducing the old image from the edges to a pinpoint size and then expanding the new one in the reverse way; or a turnover (the entire screen seems to turn over, with the new image seeming to appear on what was the reverse side). The director may introduce his own creative touches in cutting. The German-born director Max Ophüls, for example, connected the separate episodes of *La Ronde* (1950) by means of the musical leitmotiv of a hurdy-gurdy tune. In *Vivre sa vie* (1962; *My Life to Live*), Jean-Luc Godard, one of the outstanding French New Wave directors of the late 1950s, introduced chapter headings marking the heroine's step-by-step involvement in prostitution and ultimately, her murder, as if it were a didactic 19th-century novel. And Alfred Hitchcock, probably the greatest director of suspense films, in his British film *The Thirty-Nine Steps* (1935) cut from a woman's scream to the similar sound of a train whistle, an effect so dramatic that it was frequently imitated thereafter. By skillful editing highly dramatic effects may be achieved more easily than if they were staged. In the American western film *Butch Cassidy and the Sundance Kid* (1969), for example, the title characters are seen cornered by lawmen on a high cliff overlooking a river, into which they make an almost suicidal leap to their escape. Actually, the scenes involving the two leading actors on the cliff, and those of the dives were shot weeks apart, and they involved different crews and even different rivers, yet the audience readily accepts the illusion created by the editing.

The editor's function. Like camerawork, editing is a function that is ordinarily hidden from the audience, but it is vital to the finished picture. It is the editor's job to judge the length of each shot, choosing the exact moment to cut. The length of a shot may depend upon the amount of detail it contains, its scale, its dramatic impact, or its context in relation to the shots that precede and follow it. Though the audience is unconscious of these judgments, the impact of the finished film will depend on how well they are made.

The director will generally view the day's rushes (*i.e.*, advance prints of the film shot that day) and, in consultation, select what is to be used. Thereafter, he may spend a good deal of time in the cutting room or none at all. In some cases, the editor may be influential in rearranging shots, discarding them, or even ordering reshooting or additional shooting. An important factor in the work of the editor is the cutting ratio—the proportion of film shot to that used in the final film. Some directors shoot as little as two or three times as much as is required, while others may shoot six times as much or even more. In its widest sense, editing includes mixing the sounds and correlating them with the visual film, a process that is even more complex than cutting the visual film.

Cutting ratio

Cinema time. The motion picture has been defined as a series of images of space that are arranged in time. The time of film language is quite different from that of reality and that conveyed by the other arts, such as drama and literature. Movement on the screen is produced by showing the spectator every second 24 frames, or still photographs, with dark intervals between them. The movement seems to be at the same rate as that of ordinary life only if the pictures are taken and shown at the same speed. Slow motion may be achieved either by speeding up the camera or slowing down the projector, and accelerated motion is obtained in the opposite way. In common practice, the speed of the projector is constant, and the speed of the camera is varied to achieve these effects. Like extremes of scale, extremes of speed, such as accelerated motion films of plant growth, or slow-motion films of bullets, explosions, or materials being broken, are of less interest to the art of motion pictures than to science. But moderate slow motion can express dreams, or ecstasy, while accelerated motion is often very effective in comedy. The cinema can give the illusion of reversing time by showing events happening backward, or of holding time still by showing the same image again and again.

Time conventions. Despite the possibilities for manipulating it, the time presented in a single shot of film is ordinarily the time of the real world. From shot to shot, however, the time is presented according to certain conventions. In most motion pictures, the story may be assumed to be presented in chronological order and in real time except when certain conventions are invoked, such as: (1) ellipsis, (2) repetition for emphasis, (3) flashbacks, or (4) thought, or dream, sequences.

The narrative may be advanced with immense speed and economy simply by the omission, or ellipsis, of what is not essential. Generally, if a great span of time is elided between shots, some emphasis will be required on the cutting itself. A straight cut may be used between a shot of a girl dressing for a ball and a shot of her at the ball itself. To show a lapse of years, however, it might be necessary to fade one shot slowly from the screen and fade the next in, or to use a dissolve, or mix, which shows both shots superimposed as one supersedes the other.

To emphasize important scenes of short duration, repetition is an effective device. Such a scene may be shown from different angles, from a distance and then close up, and it may occupy much more time on the screen than the actual event would. By emphasizing what is important and eliminating the rest a motion picture can give the illusion of covering a lifetime in only 90 minutes.

A flashback is an interruption of the actual chronology of a story to relate a significant event of an earlier time. The flash forward, a much less used device, interjects future events in the same way. These devices require special optical effects, such as fades, dissolves, or irising, to stress the break in continuity. The break can also be stressed by the use of a melody associated with the past, or by an unusual camera movement, as well as by the more obvious devices of using noticeably different period styles in the settings or having the actors made up to look much older or younger.

Flashback and flash forward

A thought or dream sequence requires similar emphasis on the departure from chronology of real time. Nearly all of *La Rivière de l'Hibou* (1962), a prize-winning French short film adapted from Ambrose Bierce's 1891 short story, *An Occurrence at Owl Creek Bridge*, consists of the fleeting last thoughts of a man about to be hanged; by

not indicating a break between the actual events of the hanging and the fantasy of the condemned man, the film deceives the audience, until the very end, into thinking he is making an escape.

Tempo. The tempo or pace that an audience senses in a film may be influenced in three ways: by the actual speed and rhythm of movement and cuts within the film, by the accompanying music, and by the content of the story. For most people, time seems to pass quickly during moments of happiness, excitement, or exhilaration, and slowly during sadness or boredom. In films, it is possible to reverse this apparent cause-and-effect relationship and to induce a feeling of happiness, excitement, or exhilaration by making the picture seem to move quickly. Means of accomplishing this include lively music, quick cutting, or fast action. Conversely, a sense of sadness or boredom can be induced by solemn music or immobility of the images.

A feeling of suspense is unusual in combining excitement with a sense that time is passing slowly. Much of the suspense depends upon the audience's awareness of a danger unknown to the characters in the film. The height of the excitement that may be achieved in this way is sometimes seen in children and unsophisticated spectators who shout out to warn the actors.

Tempo is not necessarily related to the actual length of a motion picture. A poorly made short film may drag, while *Intolerance* (1916), D.W. Griffith's 3½-hour masterpiece, or *Greed* (1924), which its German–American director Erich von Stroheim cut from ten hours to four hours and then to two, still commands the full attention of connoisseurs. Regrettably, the length of a film has more often been determined by commercial than by artistic considerations.

Sound. Mechanical reproduction of sound was developed as early as the first motion pictures, but the problems of amplifying sound sufficiently for an audience and synchronizing it with the film image were not solved until the late 1920s. Although sound attracted crowds to the cinema to hear the new miracle, the artistic levels of the best silent pictures were not reached immediately. The new "talkies" were mostly poor imitations of theatrical plays, with dialogue and sound used indiscriminately. Sound equipment was cumbersome and imperfect. The mobile camera of the silent film lost its freedom, and the editing of film tied to a sound track became stodgy and slow.

Only gradually did film makers realize that dialogue is a much more important element in a play than in a film, and that much that the play must express in words can be shown in pictures in films.

When film makers learned to use it, of course, sound resulted in great advantages. The cumbersome captions of the silent film could be dropped; certain strained methods of showing sound in pictures, such as shots of factory whistles, guns firing, or rows of clapping hands, became unnecessary. Good music could be composed for a film and enjoyed in the humblest as well as the grandest cinema. Just as the visual image in the frame of a motion picture was elevated from the profusion of nature and could be seen fresh, so could sound be isolated for artistic purposes—the screech of automobile tires, the ticking of a watch, the baying of hounds, the whinny of a horse.

Dramatic effect of sound

The dramatic effect of sound could be tremendous. The rushing, crackling sound of a great fire in the last scene of Robert Bresson's *Trial of Joan of Arc* (1962) is as terrifying as any visual effect could be. In Hitchcock's *Torn Curtain* (1966) there is a desperate struggle in the kitchen of a lonely farmhouse; as the doomed man's head is held in an oven, and his hands (the only thing in the picture) convulsively twitch, the sound of hissing gas dominates the scene. It should be stressed that the introduction of sound made it possible to use *silence* with a dramatic effect that is more telling than either words or music.

Like images, sounds represent subjective thoughts, indicating not what the character is saying but what is in his mind. A famous example is in Alfred Hitchcock's *Blackmail* (1929), the first English sound film, in which the words "knife, knife, knife" are repeated in the thoughts of a frightened girl who thinks she has committed a murder.

In terms of montage, the sound, dialogue, and music are used in combination not only with each other but also with the visual image. They can overlap and vary in intensity in a flexible and complex pattern.

Music. The live music that accompanied silent films varied from a full orchestra to a honky-tonk piano, according to the size of the cinema. Music was also effectively used on the film set to improve an actor's performance. With sound, music became an integral part of the picture on the screen. Early mood music was so expressive that it now seems overblown. Later, film makers learned the virtue of restraint, using better music but less frequently and with more effect. Music has an important function in emotional climaxes to relieve or sublimate intolerable intensity—of grief, pain, or ecstasy. *The Gospel According to Saint Matthew* (1964) by the Italian director Pier Paolo Pasolini reveals how effective periods of silence can be, and how great music can ennoble the scenes of Christ's persecution and agony on the cross.

The sound track is generally more complex than the visual film. It may involve mixing together tracks of dialogue, background noises, and music recorded at different times; they all must be matched to each other and to the visual film. Though the audience may hear it simply as an accompaniment to what they see, the sound may be the most expensive and difficult part of a motion picture. There is an increasing use of electronic music for astringent effects, as in *2001* (1968). Music may be used symbolically: in *Léon Morin, Prêtre* (1961; *Leon Morin, Priest*) a sequence of harsh chords represents the German occupation forces, and a dancing bugle motif represents the Italian troops. Organ music is used in scenes showing the heroine with the priest in church, and piano music when they are in his flat. Hurdy-gurdy music represents two gossiping spinsters, and in a climactic scene louder and louder electronic music represents the heroine's obsessive sexual feeling for the priest, until she reaches out to take his hand.

Sound engineering. It is the function of the sound engineer to select and modify sound as the cameraman selects his images. Since the noise of crockery, cutlery, or paper or the chirping of crickets would be intolerable transferred in full volume to the screen, the sound engineer must tone them down. Treble and bass must be balanced. In other cases, in order to get the effect needed, sound has to be built up and orchestrated as if it were music. Again, sound need not correspond exactly with the visual images. Artistic use can be made of asynchronism; that is, contrasting the sound to the visual image rather than reflecting it. Motion picture sound is capable of remarkable delicacy, richness, and variety. Sound libraries put most conceivable sounds readily at the disposal of film makers. Instruments and voices can be modified, overlapped, echoed, or given a resonance and volume that transform them. Dialogue can be crystal clear, bringing the audience far closer to an actor than in the theatre, or it may deliberately reproduce the careless enunciation of everyday speech.

Dubbing

Dubbing (*i.e.*, recording the voices after the film has been shot, carefully matching the spoken or sung words to the lip movements in the film) allows the film to be shown in many languages, though many of the most serious students of the art prefer to see a film in its original language with subtitles. In musicals the sound of a song is ordinarily recorded after the film has been shot, sometimes by a singer other than the actor.

The script. Although conventions vary from one country to another, the script usually develops over a number of distinct stages from a synopsis of the original idea, through a "treatment" that contains an outline and considerably more detail, to a shooting script. Although the terms are used ambiguously, script, or screenplay, usually refers to the dialogue and the annotations necessary to understand the action; and a script reads much like other printed forms of dramatic literature, while "shooting script" or "scenario" more often implies not only all of the dialogue but also extensive technical details regarding

the setting, the camera work, and other factors. Moreover, a shooting script may have the scenes arranged in the order in which they will be shot, a radically different arrangement from that of the film itself since all the scenes involving the same actors and sets are ordinarily shot at the same time, for economy. Some scripts are subsequently modified into novels and distributed in book form, such as the U.S. bestseller *Love Story* (1970) by Erich Segal, and, in the instance of Dylan Thomas's *The Doctor and the Devils* (1953), a script became a literary work without ever having been made into a motion picture. Generally, more elaborate productions require more elaborate shooting scripts, while more personal films may be made without any form of written script.

Different directors follow different methods in working with scripts: D.W. Griffith and other early directors worked virtually without a script; Hitchcock plans the script thoroughly before shooting any film. Exponents of *cinéma vérité* ("cinema truth"), a French movement of the 1960s, produced movies by first tape-recording interviews, conversation, and opinions; they then shot visual images to correspond with the words and finally created the film in the cutting room.

Adaptation from other art forms must take into account differences of complexity and scale in film. A film often must omit characters and incidents that were in the novel from which it is adapted, for example, and the pace usually must be accelerated. Ordinarily, only a fraction of a novel's dialogue can be included. In an adaptation of a play, the curtailment is less severe, but much dialogue still must be cut or expressed visually. On the other hand, movement and cutting can be used to vary the scene and express more in a given time. For a successful adaptation, film makers must recreate a work as a motion picture.

Although many eminent literary authors, including F. Scott Fitzgerald and Aldous Huxley, have worked on film scripts, the ability to write a good one requires a stronger consciousness of visual possibilities than many novelists develop. The American novelist William Faulkner, who worked on a number of Hollywood motion pictures directed by Howard Hawks, attested to the superior ability of the relatively unknown scriptwriters to express in a concise visual shot what Faulkner himself would have conveyed less effectively by means of dialogue.

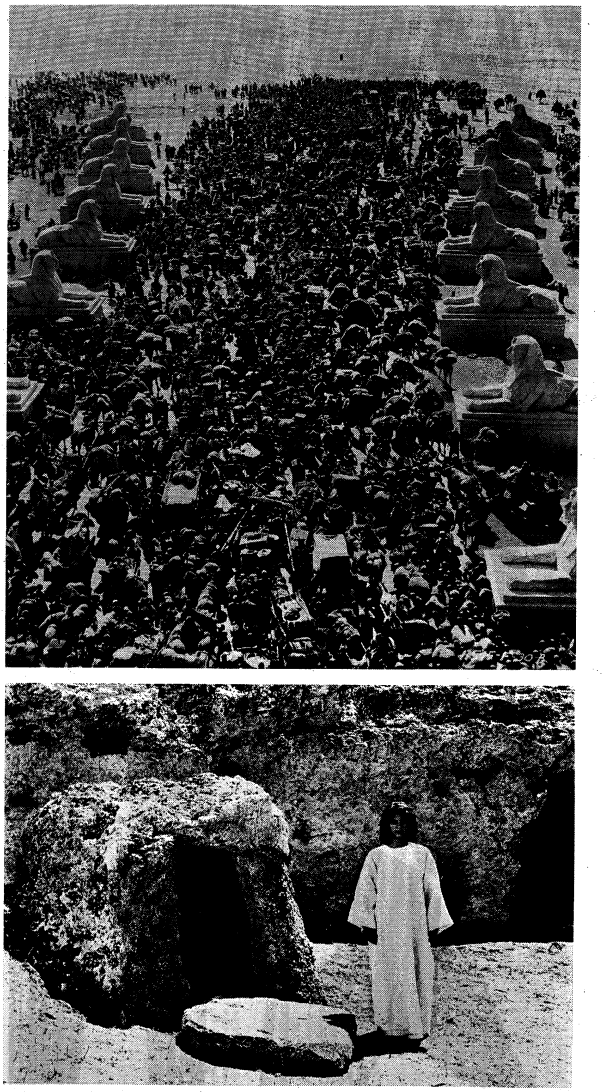
Motion picture acting. Of all the artists involved in films, the actors and actresses are closest to the audience. The public more often goes to see a motion picture for its stars than for any other single reason. The divergent techniques of stage and film acting are well understood, and there are many leading players who excel in both. But the greatest film stars have a talent peculiar to the screen alone. This talent in many instances seems to be related not to how well they act, but to the sort of person they appear to be.

Being oneself

Film acting requires less training and skill than stage acting, but much more restraint. "Don't act, think," was the advice of the eminent German director of the 1920s, Friedrich W. Murnau. While the stage actor may be praised for a performance that is highly wrought, the film star usually must appear to be himself. Few, however, seem able to achieve the greatest success. Not only is it difficult to appear to be oneself but also the self that is revealed must be photogenic and able to sustain its appeal under the camera's relentless scrutiny. Cinema audiences generally feel closer to the personality on the screen; the stage performance remains a performance. Many stars remain the same from picture to picture, but some film actors such as the U.S. actor of the 1920s Lon Chaney, the "man of a thousand faces," drastically change their appearance from one film to another. A motion picture performance can be synthesized bit by bit, by the joint efforts of the actor, the director, the cameraman, and others. The conditions of film production are such that some actors find them trying; they may have to put up with long hours on the set and endless repetition. If they require less talent than theatrical performers, they need other qualities, which are perhaps equally rare.

Motion picture design. Under the heading of design, all of the elements of a picture's setting may be included —art direction, scenic composition, set design, costume, and makeup. At its simplest and most naturalistic, the camera can choose and frame ordinary people in a real location. At its most elaborate, motion picture production may involve the expenditure of vast sums to put up gigantic sets that require building houses, ships, churches,

By courtesy of (bottom) Arco Films; photographs, from the Penguin Photo Collection



Treatment of Biblical stories in film has ranged from the colossal epics of Cecil B. DeMille to the earthy simplicity of Pier Paolo Pasolini. (Top) The Exodus scene from DeMille's *Ten Commandments* (1956, Paramount). (Bottom) The angel before the empty tomb in Pasolini's *Gospel According to Saint Matthew* (1964).

and monuments, or re-creating lost cities, by-gone landscapes, or ancient battles, and dressing thousands of extras in period costumes. Thunderstorms, tornadoes, snowstorms, earthquakes, volcanic eruptions, or tidal waves may be simulated, or monsters, spaceships, and cities of the future created. Film sets are constructed more solidly than stage scenery, but of the lightest, cheapest materials that will both look solid and photograph well.

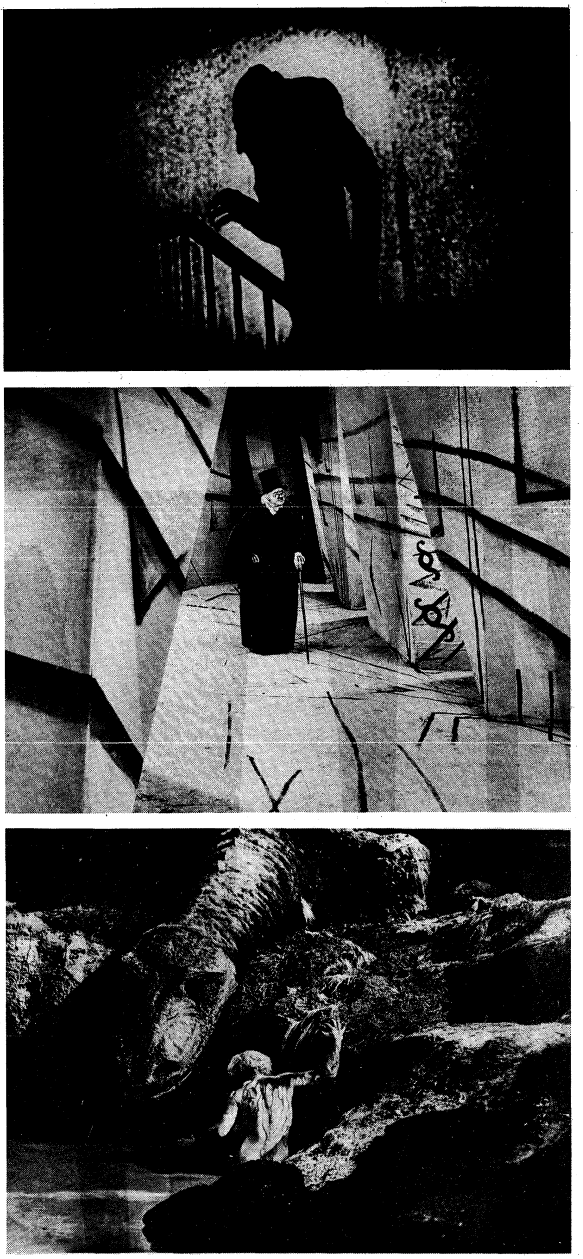
Whether for black and white or for colour, their shades, proportions and shapes are chosen for the camera rather than the eye. The camera can deceive the viewer about scale, however, and the model makers and special effects men can reproduce virtually anything in miniature, from the aurora borealis shining on an igloo to a tempest destroying the Spanish Armada. A whole series of film monsters have attracted goggle-eyed audiences since the French pioneer Georges Méliès' formidable man-in-the-moon of 1902. Miracles can be achieved in

film, either in the colossal form of the crossing of the Red Sea by the children of Israel in Cecil B. DeMille's epic *The Ten Commandments* (1956) or in the comparatively simple treatment of angels and miracles in Pier Paolo Pasolini's *Gospel According to Saint Matthew* (1964). Film publicity makes much of the creations of the epics such as *Moby Dick* (1956) or *Cleopatra* (1963) but, even on a more modest scale, sets can contribute enormously to a picture, as in Ken Adam's sets for the James Bond series such as *Dr. No* (1962), *Goldfinger* (1964).

The Expressionist influence on set design

The strongest single artistic influence in motion picture set design was the German Expressionist cinema of the 1920s, which combined elements of Max Reinhardt's theatre, a Wagnerian philosophy of doom, and Expressionist painting and graphics. Expressionist films such as *The Cabinet of Dr. Caligari* (1919), *The Golem* (1920), *Nosferatu* (1922), and *Metropolis* (1926) created a world of fantasy and horror peopled with menacing, shadowy figures. In other German Expressionist motion pictures, such as *The Student of Prague* (1926) or *The Nibelungen* (1924), there was a Baroque beauty of architectural, woodland, and floral settings. The influence of Expressionism can be seen in later cinema in the work of the directors Orson Welles and Carl Dreyer, and in many gangster movies.

By courtesy of (top) TRANSIT-FILM GMBH, (bottom) Friedrich-Wilhelm-Murnau-Stiftung, Wiesbaden; photographs, from (top, bottom) the National Film Archive, London, (centre) the Museum of Modern Art Film Stills Archive, New York



Fantastic settings and lighting effects were used by German Expressionist film makers in the 1920s. (Top) *Nosferatu* (1922) directed by F. W. Murnau. (Centre) *The Cabinet of Dr. Caligari* (1919) directed by Robert Wiene. (Bottom) *The Nibelungen* (1924) directed by Fritz Lang.

The subject and how it is treated will be reflected in all the elements of design in a motion picture. In the hectic atmosphere of a horror movie or a Hitchcock thriller, a studio set may pass unnoticed, even though it would seem disconcertingly artificial in another film. In a musical everything may look artificial: the chambermaid's bedroom may be a model of daintiness and taste. But Luis Buñuel in *Los Olvidados* (1950; *The Young and the Damned*), Satyajit Ray in *Pather Panchali* (1952), and Luchino Visconti in *La terra trema* (1948), render the peasant interiors, as well as the people who inhabit them, disturbingly real.

Lighting. In a medium consisting of the projection of impressions on a light-sensitive material, lighting is of very special importance. Daylight is the readiest, cheapest, and strongest source of lighting. Hollywood was said to owe its pre-eminence as a motion picture production centre to its sunny climate. Even in daylight shooting, however, artificial aids may be necessary to reduce the highlights or to lighten the shadows. In circumstances not involving daylight, shooting is entirely dependent on high-powered spotlights and floodlights.

Lighting, a part of camerawork, is under the control of the chief cameraman or lighting cameraman. In black and white films, lighting is of paramount importance in giving the overall light or dark tone of the scene and in providing for dramatic contrasts or emphasis within the scene. In color, the lighting works indirectly to bring out or modify the colour but is of equal importance. Lighting in the cinema is a method of composition, used to complement and reinforce the dramatic situation. Generally dark lighting is associated with tragedy and bright light with romance or comedy, but overlighting gives ghastly or unearthly effects, as in the memory sequence at the beginning of the Swedish director Ingmar Bergman's *Sawdust and Tinsel* (1953; also called *Naked Night*). Lighting an actor from above gives his face a spiritual effect; from below, an uncanny or evil appearance. Front lighting blurs faults but takes away character; side lighting gives relief and solidity but may show wrinkles and defects. Lighting from behind tends to idealize a subject and give it an ethereal quality.

Lighting as a method of composition

Fashions in film lighting have fluctuated through the years. Early films tended to be flat and uniform. Later, especially in German Expressionist films of the 1920s, strong contrasts of light and shade were favoured for their highly dramatic effect. Still later, this tendency was modified toward more balanced lighting. Even in dramatic scenes, the centre of interest now tends to be more discretely emphasized.

Related to lighting is the development of the print in the laboratory, where sections of film shot under different conditions can be modified to avoid a violent contrast where none is desired. Finally, the lighting of the projector that shows the picture and the brilliance of the screen are important. As light sources for the projector, both the traditional carbon arc and the modern Xenon lamp have advocates; what matters most is that the light should be powerful, brilliant, and unvarying throughout the projection of the picture.

Settings. A salient feature of the cinema is its ability to reproduce natural scenery. The appeal of some motion pictures lies chiefly in their natural settings. In the Swedish film *Elvira Madigan* (1967), the lushness of the greenery afforded a visual metaphor for a storybook romance between an aristocratic officer and a blonde showgirl that was doomed by the conventions of the time.

With the invention of lighter, more portable equipment, filming on location—that is, in an actual setting like the one in which the story takes place rather than in a studio set—became less difficult and expensive and more often used. As a result, many earlier motion pictures now look

artificial, since no studio set could equal the authenticity of a real location. Nonetheless, films must be shot in the studio when natural settings are unavailable, as in historical films, or too remote. In some cases, a film will be taken of a setting, such as a mountain range, and brought back to the studio, where, by means of "back projection," it will be used as a background against which the actors perform.

Costume. In Hollywood motion pictures, roughly from the 1920s through the 1940s, the costumes worn by the stars were a major force in setting fashions throughout the world. A motion picture was often an opportunity for an actress to wear one gorgeous costume after another. As the influence of motion picture costume on the world of fashion declined, the aesthetic and expressive functions of costume became more important in films, though costume was seldom completely realistic.

Costume and the actor's identity

Costume also once played a more important role in an actor's identity. Charlie Chaplin, Buster Keaton, W.C. Fields, Mae West, and other stars of the 1920s and 1930s all created characters in which costume was an integral part of the total identity. Later film comedians, such as Peter Sellers or Jerry Lewis, cannot be identified in the same way with any style of dress.

Makeup. Film makeup differs significantly from that of the stage, where heavy lines are required to convey a characteristic expression to the audience. By means of the camera, the motion picture audience can study the actor's face quite closely. The makeup must be flawless to stand up under such scrutiny, but since it need not be applied night after night, as in theatre, elaborate preparations are feasible. Many of the greatest film makers, such as Eisenstein, Carl Dreyer and Robert Bresson, have favoured simple naturalism in makeup—the unadorned lines of old age or a face caked with dust, running with perspiration. Whatever the style of makeup, its purpose is to idealize the face, to make it a more photogenic object, whether monster or ingenue. The efforts of wig makers, dentists, and plastic surgeons, as well as cosmeticians and others, are aimed at a heightened reality.

STYLE IN MOTION PICTURES

Since the motion picture is a group art, the importance of the contributions of those involved in making it may be difficult to assess. Cases have been put forward on behalf of the author of the story, the star, the studio, and the national industry as the primary artistic influence on the motion picture. Since the 1950s there has been an increasing tendency to credit the director with the merging of the myriad elements into an artistic style.

The auteur theory. Discussion of this point was stimulated by the "auteur theory" first propounded in 1954 by François Truffaut, who wrote for the French film journal *Cahiers du Cinema* and later became a major director. He attacked the French cinema for giving a dominant role to script writers, who wrote scenarios that allowed the director to do no more than frame them. Some directors, however, such as Bresson, Jean Renoir, Hitchcock, and others, were what Truffaut called *auteurs* (the French word for "authors") in that they wrote dialogue, invented their own stories, and made motion pictures that were an artistic whole, expressing their own style. The idea was taken up in France and elsewhere and led to a search for directors who were *auteurs*. It turned out that the chosen *auteur* was often a director not of high-budget popular films, which are more vulnerable to financial pressures, but of fast-moving, strong, commercial entertainment, such as Samuel Fuller, whose best works include such unpretentious Hollywood films as *I Shot Jesse James* (1949) and *House of Bamboo* (1955). This theory, like others, was useful insofar as it corresponded to the facts and offered an approach to an interpretation of a complex art. Also, by focussing attention on the professional director, the *auteur* critics helped correct a literary bias common in much film criticism. Carried to extremes, however, it resulted in a tendency to see no wrong in the work of an *auteur*. Most directors control some elements of a film and lack control over others and make films that reflect their best work only in part. In a group art with many elements interacting in shifting patterns of authority, generalization is difficult.

Directors and their styles. Of all the persons involved in making the motion picture, the director is usually the one in the best position to influence the total work with his own creative personality. Even under the studio system that prevailed in Hollywood during the 1930s and 1940s, when the directors were supposed to be the same as other corporate employees, their position was strong. Executive control was often uncertain, uninformed, or ineffectual, and authority tended to devolve upon the director. Later, when individual protection units supplanted the large studios, the director became still stronger.

The director's impact

The director lives with the motion picture. If he chooses, he usually can control every detail. Though the producer has financial control of the film and, with it, the authority to overrule, or even fire, the director, he seldom influences the artistic development of the picture as much as the director. The best directors tend to be men of strong personality who wield almost dictatorial power. Their genius lies in inspiring others, in either persuading or bullying a variety of creative persons to excel themselves.

A director's command of technical elements such as visual rhythm and the handling of the camera corresponds to a writers' command of words or a painter's control of the brush. Technical mastery is a prime virtue. Moreover, just as plain prose may represent a higher level of literature than more decorative writing, so can an unpretentious thriller directed by Alfred Hitchcock achieve a higher level of artistic expression than the blown-up epics of Cecil B. DeMille, or the grandiloquence of John Huston's *The Bible* (1966). The best director and the best stylist is not the one with the grandest manner but the one who can bring the varied elements of a picture together in a unified and near-perfect whole. Eisenstein's epic *Alexander Nevsky* (1938) is a great picture but so is a gossamer comedy like René Clair's *Sous Les Toits de Paris* (1930; *Under the Roofs of Paris*) or a miniature satire like *Simon of the Desert*(1965) by the Spanish director Luis Buñuel. The director's influence on the style of a film is far less direct than, for example, that of a writer on a novel. Charlie Chaplin, Erich von Stroheim, and Buster Keaton are recognized as master stylists, but they were more than directors. They dominated all their films, as writers, producers, actors—whatever was required. Their personality is evident in every film they made. In the work of lesser artists, some trick of style, some superficial characteristic, is sometimes taken for true style. The true style of a motion picture, however, is in its overall shape and handling. Hitchcock's style lies in his command of the medium to tell a gripping story at a spanking pace; to combine innocence and villainy in a macabre juxtaposition; to show the ordinary man caught in a web of danger; to depict continuously interesting situations and characters; to keep the audience in suspense; and at the end to provide a fast-moving climax. He does this, with verve and colour, by careful planning, by control of the actors, by a lifetime's knowledge of film making. The Indian director Satyajit Ray from the start developed a personal style—one that is gentle, unhurried, and deeply humane; he was able to do this because he made his own pictures, with modest means, outside the powerful Bombay film industry. In most cases, however, directors must cope with a system, an industry that, inevitably, will limit their freedom to express themselves. Only the exceptional or the veteran director will be free, and then only occasionally. For this reason, the style seen in a motion picture is more often not that of an individual but that of a country, a school, a group, or a movement.

The true style of a film

National styles and schools of motion pictures. In film making the international exchange of leading actors, actresses, directors, cameramen, and other film makers is common. Jean-Luc Godard's satirical motion picture, *Le Mépris* (1963; *Contempt*), shows an American producer and a German director making a picture of *The Odyssey* in France from an Italian script; there are more complex real examples. Many such pictures, by attempting to transcend national frontiers, lose their character altogether.

As a popular art, the motion picture has roots deep in

native soil; it is possible to discern national schools of film making. American silent comedy, for example, is reminiscent of an irreverent Yankee verve that was expressed in literature by Mark Twain, Artemus Ward, and Bret Harte. The western and the gangster picture were true native American products; that Hollywood was at its best dealing with an environment near at hand may be seen by comparing examples of those genres with the reverent but wooden *Story of Louis Pasteur* (1936). Two American masterpieces of the epic, *Gone With the Wind* (1939) and *The Birth of a Nation* (1915), deal with a central American theme, the U.S. Civil War and its aftermath.

Studio styles

In Hollywood at its peak, each motion picture company had its own "house" style, and whoever the director or stars, a picture from MGM, Warner Brothers, Paramount, or Columbia was recognizable. This was partly due to company policy in choosing subjects, to the influence of top executive producers, the work of the same scriptwriters, cameramen, lab workers, cutters, art directors, and others, and to the influence of a closely knit organization

Other countries also fostered distinctive native styles in film. Among the best British motion pictures were the cheerful comedies made by the Ealing studio after World War II, such as *Kind Hearts and Coronets* (1949) and *The Ladykillers* (1955). The settings, the actors, and the stories were all typically British. Across the English channel, the films of René Clair, Marcel Pagnol, and Sacha Guitry were distinctively French—light, elegant, poignant, and sophisticated. A later generation of film makers, the *nouvelle vague* or New Wave directors of the late 1950s, revealed a very different aspect of the French character—iconoclastic, idosyncratic, personal, hostile.

The Soviet directors of the silent era probably created the most consistently powerful films ever seen, singularly lacking in humour or in tender graces but expressing emotion on a grand scale; relationships can easily be drawn to Russian music and literature. The characteristics of each of the national schools are treated in detail in MOTION PICTURES, HISTORY OF. It will suffice here only to indicate the fact of their existence.

TYPES OF MOTION PICTURES

Popular and art films. The cinema audience of any nation can be divided roughly into a mass audience that seeks entertainment in films and a smaller group that is consciously concerned with artistic values in the motion picture. The films that appeal to these two groups, however, will vary from one nation to another and from one period to another. The Hollywood comedies of the 1920s by Charlie Chaplin, Buster Keaton, and others, for example, were originally popular entertainment but they were later taken up wholeheartedly by the art film audience. The comedies of Jerry Lewis received little serious critical attention in their native United States but a greal deal in France. The low-budget thrillers of directors who were adopted by *auteur* critics met a similar experience. And several films, such as *Easy Rider* (1969), which represented attempts to break away from the commercialism of the industry to do something of artistic merit, achieved a popular success far beyond that of patently commercial efforts. The distinction between popular entertainment and high art seems indisputable in most instances, but confusions of fashion and conflicts in artistic standards resulting from experiment and change make it difficult to generalize about them.

Origin of genres

Genres. Motion picture genres developed spontaneously for the most part. Since film making is so expensive, there has always been a strong tendency to imitate the most recent popular successes. If motion pictures involving the same type of subject matter continue to be successful over the years and are flexible enough to adapt to changing tastes, a genre is born.

The Western. The genre most often associated with films is the western. It began with *The Great Train Robbery* (1903), an immensely successful film that indicated the narrative power of the film more effectively than any of its predecessors. For the next half-century, the genre flourished in countless examples, featuring one cowboy star after another cut in the same heroic mold. By mid-20th century, it proved its ability to change with the times in *The Gunfighter* (1950) and *Shane* (1953), in both of which the protagonist attempted to live down his reputation as a fast-drawing gunman of the sort that had hitherto been idealized. In *Midnight Cowboy* (1969), which might be called an antiwestern, the cowboy image was reduced to a pathetic figure.

Gangster films. The next most popular genre is probably the gangster film. Examples appeared as early as *Musketeers of Pig Alley* (1912) by D.W. Griffith and *The Gangsters and the Girl* (1914) by Thomas Ince. The genre was greatly invigorated, however, in the late 1920s, when the prohibition of alcoholic beverages in the United States led to the burgeoning of racketeering. Outstanding films of this period were *Underworld* (1927), *Little Caesar* (1930), *The Public Enemy* (1931), and *Scarface* (1932). They became so popular that considerable public concern was expressed over the possible antisocial effects of glamourizing criminals, and some of the films were suppressed. Such films continued to be made, however, and set in contemporary situations, *e.g.*, *Rififi* (1954) set in France. The 1960s saw a revival of the Prohibition era gangster film in *The St. Valentine's Day Massacre* (1967) and the tremendously popular *Bonnie and Clyde* (1967). Gangster films are notable exceptions to the popular demand for happy endings; the gangster hero nearly always must die.

Musicals. The musical became popular with the introduction of sound in *The Jazz Singer* (1927), a silent film with several musical numbers featuring Al Jolson. Its immense success led to the immediate creation of the musical genre. In 1929 and 1930 a number of Broadway musicals, such as *Gold Diggers of Broadway*, were filmed. At the same time, the great French director René Clair proved his mastery of the element of sound in films such as *Sous Les Toits de Paris* (1930; *Under the Roofs of Paris*), which, unlike the American musicals, featured only a few songs in a primarily dramatic story. The musical remained predominantly a U.S. genre; no other nation produced dancers as popular as Fred Astaire and Ginger Rogers, or production numbers as spectacular and inventive as those of the choreographer–director Busby Berkeley. Among the masterpieces of this genre are *On the Town* (1949), *West Side Story* (1961), and *The Sound of Music* (1965), the last of which broke all previous motion picture records for earnings.

The rise and decline of genres. Not all genres enjoy such consistent popularity in motion pictures as those mentioned previously. "Mother" films, for example, were popular through the 1930s. This genre was epitomized by *Stella Dallas*, the story of a mother who sacrifices all for her daughter; it was an outstanding success in both a silent version in 1926 and a sound version in 1937. Similar works appeared regularly until the 1950s, when television became the major medium for these stories.

"Mother" films

Another example of the fluctuations of a particular genre may be seen in films about spies or secret agents, a recurring subject in films from the earliest days, but the James Bond films of the 1960s achieved unprecedented popularity. *Goldfinger* (1964) and *Thunderball* (1965) represented two of the most successful financial investments in motion picture history. That fact encouraged a spate of imitations, both in motion pictures and on television, that virtually exhausted the genre by the end of the decade, a relatively short life for a genre.

The serial. Also called the chapter-play or cliff-hanger, the serial flourished in the days of silent films, when cinema going was a weekly habit. Perhaps the most famous were Louis Feuillade's *Fantômas* (1913–14) and *Judex* (1916) in France and the U.S. series of the same period with Pearl White, such as *The Perils of Pauline*. In the 1960s, old serials were revived as period pieces of popular art, and their improbable plots, exaggerated acting, and old-fashioned decor were rediscovered by modern, sophisticated audiences. There are occasional art-house showings of complete series. The French director Georges Franju made a modern pastiche *Judex* (1963) based on the early serial style.

Interpretations. For want of a better term, "interpretation" may be used to describe the type of motion picture in which a play, a ballet, an opera, or some other work of another art form is kept virtually intact and recorded by the camera and microphone. Adaptations of novels or plays re-create the work in motion picture form, but interpretations merely give the performance a wider audience. The English director Tony Richardson's version of *Hamlet* (1970) is an example of such a filmed record of a theatrical performance. More camera movement was added in *Long Day's Journey Into Night* (1962) and *The Caretaker* (1963), but they remain essentially stage plays. The straightforward transcript of Laurence Olivier's stage play *Othello* (1966) affords a striking contrast to Orson Welles's *Othello* (1952), in which the play was very successfully transformed into a motion picture. Most motion pictures of operas and ballets may be classed as interpretations.

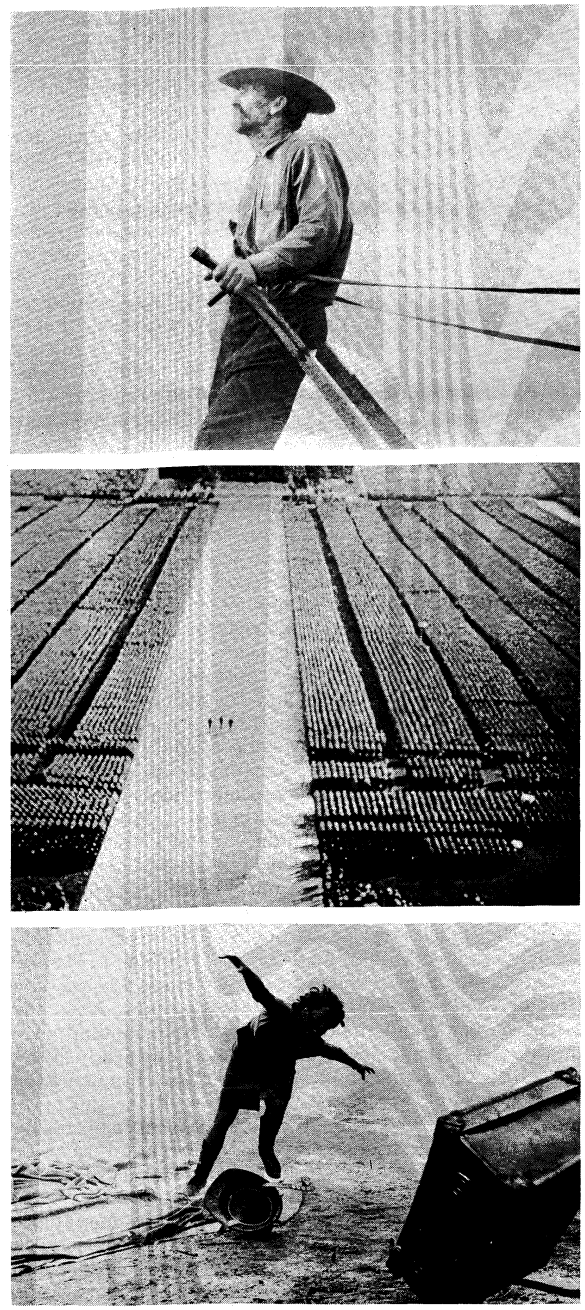
Children's pictures. Motion pictures made specifically for children are rare compared with, for example, the number of books made for children. In the 1930s, the effect of motion pictures on young people was hotly debated. The Children's Film Foundation was established in England, and in America a National Children's Film Library Committee was formed, not, like the English body, to make pictures but to select from normal commercial production those suitable for children. Few outstanding children's pictures have ever been produced and, with the advent of television, the number diminished. There has been more production of children's pictures in east European countries than in the West. While pictures produced especially for children seem deserving of support, the best children's pictures, like the best children's books, often are simply the best pictures, which appeal to audiences of all ages—such as Robert Flaherty's classic *Nanook of the North* (1922), Harold Lloyd's *Safety Last* (1923), *The Wizard of Oz* (1939), or *Le Ballon Rouge* (1956).

News and documentary films. For all of its ability to show the actual world, the motion picture failed to provide genuine news until it did so by means of television. Too stale and infrequent for day-to-day coverage, newsreels showed not news but parades, ceremonies, sporting events, bridge building, launching of ships, and similar events. *The March of Time*, inspired by Time magazine and produced by Louis de Rochemont from 1935 to 1951, was a series in which a topic of political or social importance was discussed in depth in a 30-minute film. The series was an immediate and continued success. It is television, however, that has developed the screen presentation of news, comment, and discussion beyond anything known before.

It is not in the straight presentation of reality but in its creative interpretation in the form of the documentary that the motion picture has produced works of lasting value. Among the pioneers of the documentary were the American Robert Flaherty, creator of *Nanook of the North* (1922), and the Russian theorist Dziga Vertov, whose films include *The Man with the Movie Camera* (1929). It was the British director John Grierson, however, who had the greatest influence. Although he is not to be compared to Flaherty as a film maker, he had the gift of organizing others and negotiating with officials, and he was a persuasive writer. Beginning with his own *Drifters* (1929) a school of fine directors followed his lead in producing a succession of memorable documentaries through the 1930s. With the outbreak of World War II, Humphrey Jennings' *Fires Were Started* (1943), and Harry Watt's *Target for Tonight* (1941), two among many outstanding British wartime documentaries, dramatized Britain's war effort better than fictional films could.

The British documentarists

In America, Pare Lorentz made dramatic documentaries about soil erosion and the dust bowl, such as *The Plough that Broke the Plains* (1936) and *The River* (1937), during the era of the Great Depression, and, during World War II, Frank Capra, who had been an outstanding director of Hollywood comedies, made a series of documentaries under the title *Why We Fight*. The later French movement, *cinéma vérité* which has already been mentioned, made films that are much closer to journalism than to the careful compositions of the English documentary school. Though often untidy, they are fresh and realistic. Television deeply affected the development of the documentary film in two major ways: by providing a training ground for documentary directors, and by building a supply of news film that could be adapted to documentary form. *Point of Order* (1964), a U.S. documentary film that ran successfully in motion picture theatres, was made from television films of the U.S. Senate hearings on the charges and countercharges made by Sen. Joseph MacCarthy and the U.S. Army. *The War Game* (1966), a British television documentary, enjoyed a similar cinematic success.



Documentary film techniques have often been adopted for political or ideological ends. (Top) *The Plough that Broke the Plains* (1936), a U.S. government-sponsored film directed by Pare Lorentz, publicized New Deal measures to combat soil erosion. (Centre) *Triumph of the Will* (1936), directed by Leni Riefenstahl, was sponsored by the German government to convey the power of Naziism. (Bottom) *The War Game* (1966), a BBC television production directed by Peter Watkins, used a documentary technique to show what might have happened if Britain suffered nuclear devastation.

By courtesy of (top) Pare Lorentz, (bottom) the British Film Institute on behalf of BBC-TV; photographs, from (top, centre) the Museum of Modern Art Film Stills Archive, New York, (bottom) the National Film Archive, London

Propaganda. In presenting a background, an environment, and characters who behave in a certain way, every motion picture may be said to be propaganda of a kind. The term may be restricted, however, to pictures made deliberately to influence opinion or to argue a point. The most powerful and most consistent use of the cinema for propaganda has been by the U.S.S.R. After the 1917 Russian Revolution, Soviet films exploded on the screen with fervent conviction. Gradually, however, the pictures became lifeless, and in the 1930s and 1940s, during the Stalin regime, the great directors such as Eisenstein and Aleksandr Dovzhenko worked under severe restraints. A curious feature of Russian pictures is the selection of actors who exactly resemble revolutionary heros—Lenin and Stalin especially. The Stalin pictures were frequently seen by him and were even said to have influenced his conduct. Nazi Germany produced its own brand of propaganda pictures in the 1930s, the most striking being Leni Riefenstahl's *Triumph of the Will* (1936), a terrifying spectacle of a tremendous Nazi rally.

THE STUDY AND APPRECIATION OF MOTION PICTURES

Writings about film. Criticism of motion pictures generally follows the pattern of the other arts, though it has some peculiar features. In the criticism of painting, music, theatre, and literature, some technical knowledge of the medium has generally been required. Motion picture reviewers, however, rarely have any experience of film making. A lack of experience in film making is not necessarily a handicap. Film is such a complex medium that a technician is unlikely to be expert in every aspect of it and may well be overly conscious of his specialty. Also, if the technique is complex, the results are plain to view, and an intelligent critic may well appreciate such technicalities as fine camerawork, tight editing, authentic location shooting, or competent direction without knowing in detail how the result was achieved.

Film criticism

Besides reviews in the daily and weekly press, there are for motion picture enthusiasts specialized monthlies or quarterlies, each with its particular style and reputation —such as the U.S. journals *Film Quarterly* and *Films in Review*, the British quarterly *Sight and Sound*, the French *Cahiers du Cinéma*, the Italian *Bianco e Nero*, the German *Kino*, and many others. At the popular level there are fan magazines, many of which are short-lived, though *Picture Show* and *Photoplay* have continued from 1919 and 1921, respectively. Trade journals appear in most countries, and, though primarily of specialized interest, the files are useful sources of historical information.

Writing about the cinema ranges from serious studies of its economic, sociological, and technical aspects to light memoirs, ephemeral gossip, and catchpenny biographies of actresses, actors, producers, or directors who happen to be in the news. There has been less writing about film history and aesthetics because of the youth of the art, the fact that it has not been taken seriously, and the difficulty of access to research materials. Although documentation is improving, leading to more and better historical writing, the application of scholarly attitudes to motion pictures is still restricted.

The largest book library specializing in motion pictures is at the Czech Film Archive in Prague, 70,000 volumes. The British Film Institute has 18,000 books and pamphlets besides newspaper cuttings and 100,000 index references, now on microfilm. Other libraries are at the Museum of Modern Art in New York, the University of California at Los Angeles, Institut des Hautes Études Cinématographiques in Paris, and the Centro Sperimentale in Rome.

Preservation of film. The permanence of the motion picture medium—the fact that films can be stored and reproduced indefinitely—not only makes it an enduring theatrical art but also a vivid record of past life. Though motion pictures may last forever, however, relatively few have been preserved and many of these are in poor condition. One reason is that inflammable nitrate film stock, which was generally used until the 1940s, when it was replaced by an acetate film stock, is chemically unstable. Also, as a projector runs and jerks the film, it is eventually worn, scratched, or damaged. Still another factor is that commercial conditions of film making discouraged preservation. In the trade the stress was on the present and the future. It is easier to find copies of early pictures made independently, because only the big companies had systems in force for junking (*i.e.*, destroying) old films. In the past movies have been best preserved when a film maker such as Chaplin or Disney has had control over his own work and a personal interest in preserving and representing it. Recently, renters have begun to appreciate the possibility of revivals and reissues, and attitudes are changing.

Film preservation that allows access to old pictures is costly, requiring careful scientific control of storage conditions. Few private organizations or governments have been willing to spend the necessary funds. The earliest film archive was the Swedish Film History Collection begun in 1933. Archives in Paris, London, and New York followed shortly afterward. An international federation (FIAF; Fédération Internationale des Archives du Film) was founded with headquarters in Paris in 1938. The archivists' problems are many: first, selecting the motion pictures to be preserved; second, acquiring copies (a negative or a fine-grain positive if possible) in good condition; third, storing them under the best possible conditions of temperature and humidity; fourth, cataloguing them and keeping some record of their contents; and fifth, letting them be viewed or letting stills or extracts be taken, without damaging the copies. The ideal solution to the problem of choice would be to preserve everything, but the cost would be prohibitive. Even with a limited selection, acquisition is expensive and difficult. The storage of nitrate film necessitates regular testing to see whether it has deteriorated enough to require copying and destruction. Safety film, which was introduced in the 1940s, has an acetate base and an indefinite life and is considered the best means of preserving motion pictures, better than the later magnetic or videotape. Access remains a problem because a projector damages a film. Either an expendable viewing copy is needed or a table viewer has to be used. Motion picture companies and television companies have their own film collections but only for currently popular films.

Problems of the archivists

Film societies. These groups represent an intellectual fringe interested in studying the art of the cinema, discussing motion pictures among themselves or with visiting film makers or lecturers, and making films individually or together. In England and America *ciné-club* is the term used for a group more interested in film making, but on the Continent it corresponds to a film society, formed primarily to see out-of-the-way motion pictures.

The first film society was The Film Society of London, founded in 1925 by H.G. Wells, George Bernard Shaw, Augustus John, John Maynard Keynes, and other dignitaries who wanted to see French, German, and Russian pictures that commercial exhibitors did not handle. The movement spread rapidly. To a great extent the original interest in foreign films was met later by the spread of art cinemas. Nonetheless, film societies proliferated, mostly through the use of 16-mm film and portable, easily run projectors, and lending libraries. Many groups were centred in schools and universities, clubs, hospitals, businesses, and neighbourhoods. By 1961 film societies existed in 46 countries with a total membership estimated at more than 4,000,000. By the 1970s there were several thousand such groups in the United States. In contrast to the casual, impersonal nature of commercial exhibition, they provided the enthusiast with historical perspective to film viewing and analyses and discussions of aesthetic qualities. They also fostered experiments in film making.

BIBLIOGRAPHY. RUDOLPH ARNHEIM, *Film as Art* (1958), an original approach, though not fully developed, comparing the motion picture to the real world; BELA BALAZS, *Filmkultúra* (1948; Eng. trans., *Theory of the Film*, 1952), a somewhat dated but interesting account, especially on the close-up and the human face; ANDRE BAZIN, *Qu'est que le cinéma?*, 4 vol. (1958–62; Eng. trans. of selections, *What Is Cinema?*, 1967), essays by a French critic with wide knowledge and penetrating insight; GEORGE BLUESTONE, *Novels into Film* (1961), a specialized aspect of motion pictures, but the

analysis of particular cases leads to valuable comment on the aesthetics of the cinema; KENNETH BROWNLOW, *The Parade's Gone By* (1968), a nostalgic book, mostly about the early days of Hollywood, with information and perceptive comment about every aspect of motion pictures; SERGEY EISENSTEIN, *The Film Sense* (Eng. trans. 1948), a classic of motion picture theory, but difficult reading; ARTHUR KNIGHT, *The Liveliest Art* (1957), the best and most readable short general history, with valuable comment on many aspects of cinema; SIEGFRIED KRACAUER, *The Nature of Film* (1961), an exhaustive examination of motion picture aesthetics, though sometimes slow reading; ERNEST LINDGREN, *The Art of the Film*, 2nd ed. (1963), a readable and excellent introduction; EDGAR MORIN, *Les Stars* (1957; Eng. trans., 1960), a fascinating account of the star system; V.I. PUDOVKIN, *Film Technique and Film Acting* (Eng. trans. 1958), a classic of film aesthetics by a great Russian silent director, who is easier to read than his colleague Eisenstein; KAREL REISZ, *The Technique of Film Editing* (1953), a readable account of a central technique of motion pictures; R. STEPHENSON and J.R. DEBRIX, *The Cinema as Art* (1966), an aesthetic analysis of the whole range of motion picture techniques in terms of the picture on the screen.

(R.St.)

Motion Pictures, History of

The motion picture's prehistory is almost as long as its history. The evolution of the first cinema machines was a slow and gradual process; in the first years after their invention, discovering the applications of those machines also came very slowly and tentatively. Their possibilities began to be understood in the second decade of the 20th century and an astonishing explosion of creativity followed. The essential problem facing the earliest film makers was to free the motion picture from the tyranny of time and space. Once the film artist had discovered that his film could leap forward or backward in time, could mirror subjective time as well as its literal passing, could dart anywhere in the world at will (including into the thoughts of its characters), could simulate the feeling of an experience as well as its appearance, could use the rhythms of an event to depict it as well as a literal dramatization, the essential problem of visual-film aesthetics had been solved.

The introduction of synchronized sound threatened the motion picture with a second tyranny of time and space —for the spoken word can restrict the location and duration of a scene as concretely as the earlier assumption that the motion-picture camera was analogous to a single spectator in the theatre. As they did 15 years before, the film artists of the early '30s learned to master the new machine and then learned to make it work for them rather than against them. The artist integrated dialogue and picture, just as his predecessor of the silent era integrated composition and cutting, space and time, texture and rhythm. In the years preceding World War II, most film artists sought effective and evocative visual accompaniment for dialogue. The dependence on sound made the film world less flexible than in the silent era, more allied to space and time in the natural world. Like speech itself, the film world became more logical, more literal, and more linear than it had been in the late silent period.

Since World War II the film artist has worked on destroying that logic. The introduction of the wide screen and the almost total domination of colour, the reappearance of obtrusive editing and subjective camerawork, the reliance on music, sound, and silence in addition to dialogue have all created as flexible a motion picture as in the most fertile years of the silents. The visual techniques of films made after 1960 are closer in many ways to silent methods than to those of the intervening 30 years. This fact may partially explain the recent renaissance of silent films and their revivals in theatres and on television.

Commercial factors

Audience demands and the perceptions of businessmen about those demands determined the history of the motion picture as forcefully as the discovery of the aesthetic principles of the art. For example, although both film producers and film critics believed (or hoped) that silent and sound films would coexist, the public, pursuing novelty, refused to buy tickets to silent motion pictures. Commerce inevitably determines the direction of the film art; the film artist uncovers and develops all the aesthetic possibilities once that general direction has been charted. In the history of motion pictures, therefore, commercial realities, aesthetic principles, mechanical innovation, and audience taste are inextricably linked.

This article traces the historical development of the motion picture. Other aspects of the medium are treated in MOTION PICTURES, ART OF; MOTION PICTURES, TECHNOLOGY OF; MOTION-PICTURE INDUSTRY; and ANIMATION.

This article is divided into the following sections:

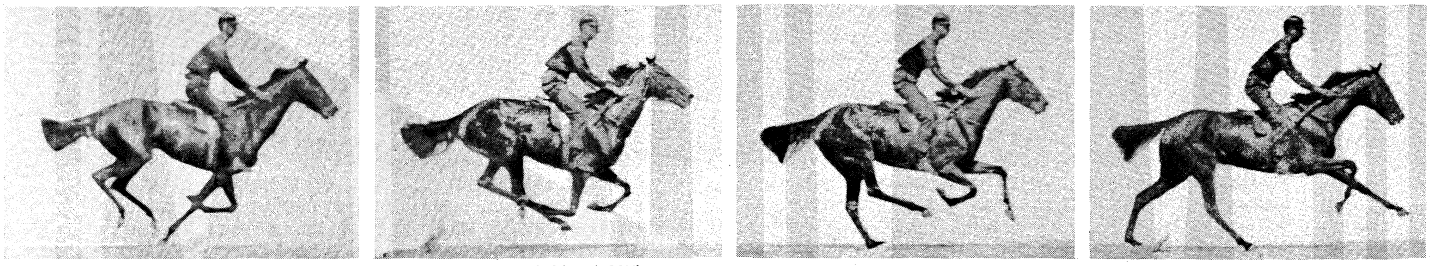
I. Origins: 1815–95
  Edison and his laboratory
  The Lumière brothers
II. Filmstrips and "one reelers": 1896–1907
  Europe
    France
    England
  United States
III. The silent feature: 1908–28
  United States
    The age of D.W. Griffith
    The early age of comedy
    Maturity of the U.S. film industry
    Comedies of the 1920s
    The early documentary
  Europe
    Germany
    Soviet Union
    France
    Scandinavia
IV. The sound era
  1928 to 1945
    Experimentation and the first "talkies"
    France
    United States
    England
  1945 to 1965
    United States
    Italy
    France
    Sweden
    Britain
    Eastern Europe
    Japan
    India
V. Recent trends

I. Origins: 1815–95

The motion picture was born in the final decade of the 19th century from the fusion of three separate mechanical interests of that mechanically minded century: the persistence-of-vision toys, still photography, and mechanized audience entertainments. The persistence-of-vision toys capitalized on the physiological fact that the brain retains the images of the eye for an instant longer than the eye actually records them. This retention allows a series of separate pictures to blend into one another and appear as a fluid and continuous action. This optical illusion, essential to motion pictures (a motion picture is merely a synthesis of individual still photographs or frames), was the basis of many 19th-century toys. Called by curious, ornate names, such as Phenakistiscope, Zoetrope, Choreutoscope, Praxinoscope, and Stroboscope, these toys were merely some kind of circular, revolving machine with drawn figures upon it, each in a slightly different phase of movement. The circle could be a flat board, a revolving drum, or even a complex rotating machine with different planes and mirrors. But the principle of the machines was identical—to bring the inanimate to life and the apparently still into motion by spinning the disc or drum.

Early still cameras

But without photographs of the natural world, there could be no motion pictures of the natural world. The camera can be traced back to Leonardo da Vinci's diagram of the *camera obscura*, the little box that captured an inverted image of the world opposite the lens (originally, simply a hole) on the rear wall of the box. Early in the 19th century, scientists began looking for the means to replace that wall with a photographic plate that could retain the image opposite the lens clearly and permanent-



An early series of photographs by Eadweard Muybridge (1887).
By courtesy of the British Film Institute, London

ly. Those experiments culminated in France in 1839 with the collaborative efforts of Louis-Jacques-Mandé Daguerre and Joseph-Nicéphore Niepce. Daguerre's first photographs required an exposure time of 15 minutes; the motion picture would require at least 16 exposures per second. There could be no motion picture until the photographic equipment and material became faster and more sensitive. By 1870 the still camera had acquired a shutter, and shutter speeds of 1/1000 of a second were being used.

The work of Eadweard Muybridge

The first man to apply still photographs to motion photography was Eadweard Muybridge, an English adventurer who emigrated to California in the years following the 1849 Gold Rush. In 1872 the governor of California, Leland Stanford, employed Muybridge to help him win a bet. Stanford, an avid horse breeder and racer, bet one of his rivals that, at some point in the galloping horse's stride, all four hooves leave the earth. In 1877 Muybridge's discovery won Stanford the $25,000 bet. Muybridge placed 12 still cameras alongside the race track; he stretched strings across the track, each attached to the shutter of one camera. As the horse galloped down the track, his hooves broke a string and tripped a shutter, exposing one phase of his movement. Muybridge realized that the original movement could be resynthesized by mounting these still photographs on a revolving wheel and projecting them with a magic lantern. For the next 20 years, Muybridge continued shooting and projecting these "pictures in motion." But they were not really motion pictures, for the photographs had been taken with 12 or 20 or, eventually, 40 cameras, not with one.

A Frenchman, Étienne-Jules Marey, took the first motion pictures with a single camera in 1882. Inspired by Muybridge's work, Marey constructed a photographic rifle that exposed 12 images in one second of "pulling the trigger." The images were recorded on a single glass plate in the chamber of the "rifle," which rotated 12 times during the second of shooting. But this early camera and Marey's subsequent improvements on it still did not permit more than a few seconds of action. Further development of the motion picture required photographic material other than the single photographic plate.

EDISON AND HIS LABORATORY

Claims to the invention of the motion picture

Because this material—celluloid roll film—was invented by George Eastman in the United States, another American, Thomas A. Edison, often receives credit for inventing the motion picture, for the Edison laboratory applied Eastman's celluloid film to motion picture photography. Edison's claim, however, is quite fuzzy. First, Edison did not invent any motion-picture machines, although he took credit for others' inventions. Invented as early as 1888, the Edison motion-picture camera, called a Kinetograph, was the work of William Kennedy Laurie Dickson, an Englishman who migrated to the U.S. and worked in Edison's laboratory. The Edison projector, the 1896 Vitascope, was the invention of Thomas Armat, an independent American inventor who, like hundreds of similiar Americans in the 1890s, had been tinkering with various motion-photography machines. Second, although the Edison laboratory indisputably produced the first motion-picture camera, Edison opposed projecting motion pictures on a screen. As a result, England (William Friese-Greene), France (Louis-Aimé-Auguste Le Prince), and Germany (Max and Emil Skladanowsky) all claim invention of the motion picture because inventors of these countries either patented machines for projecting motion pictures or actually projected pictures in public before Edison did.

Edison cared very little about motion photography. After inventing his phonograph, Edison merely wanted a machine that might supplement the music with pictures. But Dickson, certainly the most important pioneer in the earliest years of U.S. cinema, realized the power of the motion picture. It was Dickson who decided to use George Eastman's new celluloid film rather than etch the tiny pictures on wax cylinders, a way analogous to that used to make phonograph recordings. In 1889 Dickson shot the first film in the U.S. on celluloid, *Fred Ott's Sneeze*, a medium close-up of one of the workers in the Edison factory sneezing for and at the camera. Dickson also invented the Kinetoscope, a little peep-show viewer that displayed *Fred Ott's Sneeze* and other filmed items to the eye of a single spectator. Edison felt a motion picture could only make money if each curious customer looked into the viewer alone—and paid his coin each time. Finally, it was Dickson who helped perfect one of the first U.S.-made motion-picture projectors.

Edison patented Dickson's Kinetograph and Kinetoscope in 1891, but slowness of manufacture kept them from making a real impact until 1894. The Kinetograph ran at a speed of 48 frames per second (a speed that Edison believed produced the most realistic impression of movement). The body of the camera could hold only 50 feet of film on a series of loops and spindles; therefore, it contained only a few seconds of unedited action. To extend the spectacle, the film ran continuously, its ending leading directly into its beginning, as in the persistence-of-vision toys. Among the most significant and lasting details of this film machine were the little perforations, four pairs per frame, that Dickson added to the sides of the celluloid strip to keep the film moving both in the Kinetograph and in the Kinetoscope.

Among the early Edison filmstrips (1893–96) were dancing girls (*e.g., Annabella* and *Carmencita*), gymnastic knock-abouts (*e.g., Fun in a Chinese Laundry,* in which a policeman chases a limber Chinese man around a painted set), historical pageantry (*e.g., The Execution of Mary Queen of Scots,* in which Mary's head bounces off like a basketball), and re-creations of Broadway plays (*e.g., The John Rice–May Irwin Kiss,* the first close-up screen embrace, which predictably produced the first wave of moral reaction against the motion pictures). The Edison films were shot on an indoor set, with an inert camera that masqueraded as a spectator in the theatre. The Edison movie studio was patterned after a theatre—the camera was a stationary audience and the performer acted on a specially erected stage.

Edison's decision against projection

Edison's decision against projection was a serious mistake. It eventually cost him prestige, millions of dollars, and over a decade in fruitless lawsuits. In assessing the commercial possibilities of the motion picture, Edison should have realized that 19th-century audiences delighted in mechanical shows combining movement, figures, and light. Magic-lantern shows and their variants (the Phantasmagoria, the Projecting Phenakistascope, the Panorama, and the Diorama) were popular before there was ever a motion picture. The photoplay of Alexander Black, an American novelist, lecturer, and photography "fiend," produced (1894) an evening in the theatre that combined a melodramatic plot, live actors, and pro-

Scene from *The John Rice–May Irwin Kiss* (1896), a filmed scene from a current play, *The Widow Jones.*
By courtesy of the Museum of Modern Art Film Stills Archive, New York

jected slides. This type of entertainment was the immediate ancestor of the fictional motion-picture story. The projecting of motion pictures for paying audiences was an inevitable culmination of the urge of the entire 19th century.

THE LUMIERE BROTHERS

Because Edison refused to extend his U.S. patent on the Kinetograph and Kinetoscope to Europe, the most significant motion-picture inventions between 1895 and 1900 came from Europe. The Europeans took Edison's early machines, copied them, and improved on them. The most important of these early European inventors were the French brothers Auguste and Louis Lumière, who extended the family business of manufacturing still cameras and photographic plate into motion-picture equipment. The Lumières improved on the Edison machines in three ways. First, they invented a light, portable, hand-cranked camera that could shoot anywhere. (Edison insisted on running his bulky Kinetograph with electricity.) Second, the Lumières slowed the camera speed to 16 frames per second, which not only required less film but also allowed a projector to run more smoothly and quietly. And, third, the Lumières used the same machine to shoot, project, and print the pictures. Because the machine necessarily used the same principle of intermittent motion (each frame stopping momentarily before the lens and then moving on) in both shooting and projecting, the primary technical obstacle to clear, smooth projection had been hurdled. Edison's early experiments with projection, as well as the other early U.S. attempts, erroneously used continuous motion of the film past the light source, not intermittent motion.

The Lumières shot their first film early in 1895. Unlike *Fred Ott's Sneeze*, the Lumière film, *La Sortie des ouvriers de l'usine Lumière* ("Workers Leaving the Lumière Factory"), showing workers going about their daily business, was spontaneous and unstaged. This difference between the Lumière "aesthetic" and the Edison "aesthetic" was consistent through most of their films—Edison was the father of the fictional, "theatrical" film, Lumière of the natural, "documentary" film. The Lumière films were usually shot outdoors. They tended to capture some life process actually occurring in front of the lens rather than something especially prepared for the camera. Among the most famous early Lumière films are *L'Arrivée d' un train en gare* ("The Arrival of a Train in the Station), parodied 70 years later in Jean-Luc Godard's *Les Carabiniers*; *Le Repas de bébé* (a sequence showing parents feeding their infant) and *L'Arroseur arrosée* (a comic film about a gardener who gets his face doused by a mischievous boy). The last film, one of the first bits of slapstick comedy on film, is the ancestor of the later great silent comedies. It contains many of the essential comic elements in miniature: the butt (the gardener), the comic antagonist (the mischievous boy), the comic physical object (a hose), the comic discovery, and justice (the gardener's catching the boy and then administering a spanking).

Early slapstick film

Lumière films were fresher, freer, and more active and alive than Edison's, although the Lumières still considered themselves manufacturers and scientists, not artists. Their firm made thousands of films that were listed in a catalog—the filmstrips being sold outright. In over five years of making these filmstrips, the Lumières did not vary their aims or methods. They took "views" (a view of a place, a process, an event) and did not edit them. The novelty of seeing the natural world mechanically recorded was exploited, and, therefore, they did not use the camera to any effect other than capturing the view. When the novelty died, the Lumière business also died, eclipsed by shrewder businessmen and more imaginative artists. The Lumières first displayed their films to a paying public in the basement salon of Paris' Grande Cafe on December 28, 1895—the date that marks the end in the first era of cinema history, for the mechanical toy passed out of the hands of the inventors and into the hands of artists, financiers, and the public.

II. Filmstrips and "one reelers": 1896–1907

EUROPE

Artistic and commercial development progressed more quickly in Europe than in the U.S. In the first decade of motion-picture history, the best films came from France and England.

France. In 1896 the Pathé brothers and Leon Gaumont in France began their commercial conquest of the European industry. In the same year, the magician Georges Méliès began shooting films of magical tricks. Méliès saw that by stopping and then starting the camera again not only could rabbits be pulled out of hats, but hats could be made to appear where there previously were none or made to disappear from where they previously had been or made to turn into rabbits themselves. Méliès' use of stop-action, as this technique came to be called, was his single contribution to developing the unique powers of the cinema. In other respects, he was like the other film directors of his era in that he saw the motion picture as the stepchild of the stage. Like Edison, Méliès used the stationary camera, the especially erected stage for the action, and the two-dimensional scenographic conventions of the theatre.

Méliès' showmanship

Despite this subservience to the theatre, Méliès' films are clever, surprising, and funny. He reveals his showmanship by filling the screen with elaborately moving scenery, with chorus lines of scantily attired ladies, and with the quirky, jerky gymnastics of acrobats, who cavort in his films as gremlins, demons, devils, elves, or moon creatures. Typical Méliès plots include the Gypsy's curse (a hex on some object causes it to erupt into fantastic activity), the drunken dream (too much wine leads to fantastic nightmare incidents and creatures), and the journey to a bizarre region (for example, his most famous, *Le Voyage dans la lune* [*A Trip to the Moon*], of 1902, or *La Conquête du pôle* [*Conquest of the Pole*], of 1912). In addition to miraculous tricks, Méliès exhibited a whimsical, anti-intellectual humor that perpetually debunked the pretensions of scientists, doctors, and professors. Méliès' company, Star Films, achieved worldwide distribution for its films and influenced film ideas in every country where the films were shown. In the U.S. many of Méliès' films were pirated and duplicated. With the collapse of the company on the eve of World War I, Méliès disappeared; however, he was redeemed from obscurity more than a decade later by enthusiastic French filmgoers.

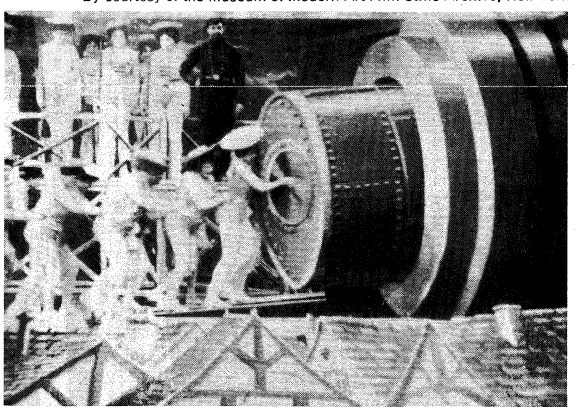
Another remarkable director of early French films was Ferdinand Zecca, head of production at Pathé. Zecca's most notable contribution to the motion picture was the chase sequence, demonstrating the vitality of motion. By taking his camera outdoors, Zecca freed the films from many of the spatial limitations they experienced in the studios of Edison, Méliès, and other contemporaries. The speed and energy of Zecca's running bodies compensated for his passive use of the camera and lack of editing. One of the striking facts of early cinema history is that, for the

La Sortie des ouvriers de l'usine Lumière (1895); "Workers Leaving the Lumière Factory."
By courtesy of the Museum of Modern Art Film Stills Archive, New York

first 15 years of motion pictures, the films that have life were those shot outdoors, where life was already available to the lens.

The final significant trend of the prewar French film demonstrated this principle very clearly, if unintentionally. In 1907 a film company calling itself the Film d'Art decided to bring "seriousness" into the motion picture. This self-conscious enterprise (the French were the first to become conscious of the film as an art rather than a novelty) intended to introduce the great writers, musicians, and actors of the French stage to the humble audiences of the motion picture. The company made several films based on the melodramas of Alexandre Dumas (*père*), Eugène Scribe, and Victorien Sardou, featuring the Comédie-Française (the national theatre of France) and the internationally renowned actress Sarah Bernhardt. The most famous of their efforts, *Queen Elizabeth* (1912), was a plodding, overacted, pretentious work that ironically showed less intelligence than Méliès and less vitality than Zecca.

By courtesy of the Museum of Modern Art Film Stills Archive, New York



A Trip to the Moon (1902), directed by Georges Méliès.

England. In England, inventors such as Robert William Paul and Birt Acres also copied the Edison machines to improve on them. A group of innovative artists used those machines to make some of the most imaginative films of the first decade of the 20th century. Often called the school of Brighton, these British film makers included G. Albert Smith, James Williamson, Esme Collings, Cecil Hepworth, and Charles Urban. Their films were notable for experiments with the moving camera, the close-up, and the rhythms of editing. Cecil Hepworth's *Rescued by Rover* (1905), a melodramatic story of an infant who is stolen away by a Gypsy and then rescued by a faithful collie, is probably the most fluidly and coherently edited film in the first 15 years of cinema history. The British

School of Brighton

also began developing the motion picture as a social document (for example, *Tatters, A Child of the Slums*, made by Collings in 1905). Members of the school of Brighton also elaborated on many of the filming tricks and surprises they had seen in the Méliès films.

UNITED STATES

Both artistic and commercial progress was slowed in the U.S. by what could be called either fertile opportunism or lawless chaos. The first official public screening of a motion picture in the U.S. was on April 23, 1896, in a New York City music hall, Koster and Bial's (at 34th Street and Broadway). For almost ten years, the motion picture did not have a U.S. home of its own but merely appeared as one act on a vaudeville bill. Eventually, the novelty of seeing the mechanical marvel wore off. U.S. films did not innovate but copied the same items—one-joke pranks modelled after the Lumières' *L'Arroseur arrosée*, inevitable "views" of people and places, speeding trains (the favourite film subject of all), historical pageants, and dancing girls. The U.S. film pioneers seemed to occupy themselves more with piracy—stealing each other's machines, machine parts, ideas, and even the film-strips themselves—than with invention.

By 1897 Edison's two major film producing rivals, the American Mutoscope and Biograph Company and the Vitagraph Company, had pirated themselves into existence. Biograph, as the former company came to be called, manufactured a peep-show viewer (the Mutoscope) that outperformed Edison's Kinetoscope and is still found at carnivals, amusement parks, and penny arcades. The Mutoscope used large cardboard photographs flipped by a hand crank rather than the Kinetoscope's whirring strip of film driven by electricity. The inventive intelligence behind Biograph was William Kennedy Laurie Dickson, who took his Edison secrets with him and put them to his own use.

Vitagraph Company

Vitagraph's founder and director of production was J. Stuart Blackton, a former newspaper cartoonist and a former friend of Edison's. The friendship cooled when Blackton received a Vitascope franchise, copied Edison's machine, and formed his own company with his partners, William "Pop" Rock and Albert E. Smith, choosing a name as close to Edison's as the law would allow. Among the other U.S. pioneers busily engaged in invention, imitation, and piracy were George Kleine, George K. Spoor, Sigmund Lubin, and William Selig.

There could be no progress of U.S. film art without some stabilization of the U.S. film industry. This stabilization process worked itself out slowly in the years preceding World War I. First, motion-picture equipment and, more important, the finished films required legal protection. Edison began suing all his competitors in 1897 for piracy and theft. Thomas Armat's projector

(which Edison bought, claimed, and named the Vitascope) required the film to be threaded with a small loop to relax the tension on the film (all projectors today use such a loop). This Armat loop ironically became known as the Latham loop because an earlier U.S. inventor, Woodville Latham, used such a loop in his camera. Because Edison owned both the Armat and Latham patents, he sued everyone who used the Latham loop. Later, Armat sued Edison; Biograph bought Latham's patent so that they too could sue Edison; and suit and countersuit continued in the courts for ten years.

This "war of the patents" eventually produced an uneasy stability and unanimity. In 1908 the major U.S. producing companies decided to call a truce. They would share all the patent rights among themselves, and they would keep all competitors out of the film business permanently. Those companies—Edison, Biograph, Vitagraph, Essanay (S and A—Spoor and Anderson), Lubin, Selig, Kalem (KLM), Méliès, and Pathé (both of the latter had begun producing in the United States)—combined as the Motion Picture Patents Company. This "company" had sufficient power to monopolize the industry because the members made the best pictures and because those members would cancel their valuable contract for raw film stock with George Eastman if he sold to anyone else. Furthermore, no theatres could rent Patents Company films if they rented from any other company.

Theatres that showed films exclusively were the real source of stability in the U.S. film industry and became a sign of the industry's growth. The first nickelodeon opened in Pittsburgh in 1905. By 1908, the year of the patent settlement, there were thousands of these theatres in the U.S. No longer did the film producer sell the filmstrip to a projectionist who toured from city to city, going from vaudeville house to vaudeville house. The permanent movie theatre developed a regular clientele. The owner of such a theatre needed many films, however, to keep his regular customers. Because a film program lasted one hour and each film only ten minutes (one reel of 1,000 feet or less), it took at least six one-reel films to make a full program. And some theatre owners changed programs as often as twice a day. For this system of exhibition, the theatre owner needed to rent films, not buy them. Between the film-producing companies and the theatre owner, or "exhibitor," stepped a middleman, the "distributor," or "distribution exchange," who bought or leased films from the producer and then rented prints to theatres. When the Motion Picture Patents Company was organized in 1908, it also created a monopolistic exchange system for distributing its films, the General Film Company, thereby increasing its power and control over the industry.

Monopoly challenged

Ironically, the solution of 1908 was a temporary one. No sooner had the Patents Company been formed than certain theatre owners rebelled against the monopolistic, anti-competitive control of what they considered an illegal trust. They sued the Motion Picture Patents Company and its General Film Company; the film industry went back into the courts. Even more devastating, the rebels formed independent film-producing companies, independent distribution exchanges, and independent circuits of theatres to show the films. Among the early independents were Carl Laemmle (the founder of Universal Pictures Company) and William Fox (the founder of the Fox Film Corporation), who were to become far more rich and famous than the founders of the Motion Picture Patents Company. In its creation of stability, the Patents Company unknowingly eliminated itself by supporting the one-reel film as opposed to the lengthy feature film and by supporting the use of anonymous screen players as opposed to the star system. Within ten years, the U.S. film industry would be based on stars in feature films. The Motion Picture Patents Company summed up the past; the independent film makers carried the art and industry into the future.

The only significant U.S. film artist of the motion picture's first decade was Edwin S. Porter. Director of production at Edison, Porter made *The Great Train Robbery*, his most famous and important film, in 1903. Often

The Great Train Robbery (1903), directed by Edwin S. Porter.
By courtesy of the Museum of Modern Art Film Stills Archive, New York

mistakenly called the first motion picture to tell a story, this film was important not because it told a story (motion pictures had told stories for at least seven years) but because of the way it told it. In its story of four bandits who rob a train, shoot one of its passengers, and eventually die in a shoot-out with the posse, Porter cut backward in time (switching focus from the bandits to the telegraph operator who summons the posse), and he cut back and forth in space between the formation of the posse and the flight of the bandits. Porter's most distinctive technique in the film was that he cut; these cuts produced an ellipsis that allowed the audience to infer a perfectly clear continuity despite the gaps in time and space.

The Great Train Robbery was the biggest commercial success of its time and held the screen for at least ten years. Most of Porter's other films, however, lack its fluidity and freeness. The vitality of Porter's films is most evident in the scenes shot outdoors. *The Life of an American Cowboy* (1903), for instance, contains outdoor scenes that are fresh and vital, while the indoor scenes seem posed and staged. Porter's use of editing as the basis of motion-picture construction was certainly inconsistent and probably unconscious. Whereas Porter's *Life of an American Fireman* (1903) contains rhythmic editing and even narrative crosscutting, his *Uncle Tom's Cabin* (1903) is a pure stage piece, and *The Dream of a Rarebit Fiend* (1906), though occasionally quite amusing, is a mere copy of Méliès.

Porter's use of natural scenery

III. The silent feature: 1908–28

UNITED STATES

The age of D.W. Griffith. The conscious discovery of the aesthetics of the motion picture begins with D.W. Griffith. To some extent, it also ends there. There was not one device of the visual, black-and-white cinema that Griffith did not master or point the way for others to control. Griffith's two general discoveries were the effects of composition and cutting—the manipulation of space and time. Griffith developed the texture, the look, the feel, and the tone of the individual shot, as well as the rhythms, the meaning, the kinetic sensations, and the intellectual commentary of joining the shots. His central realization was that the basis of film construction was the shot, not the scene (as Méliès, Zecca, and usually Porter had defined it), and that the motion picture was a new kind of language with a "grammar and rhetoric" of its own. The combining of shots into scenes, scenes into sequences, and sequences into whole plots was parallel to the combining of words into sentences, sentences into paragraphs, and paragraphs into whole arguments. Griffith made these discoveries not in theory but empirically by making hundreds of films.

Early films. Born in Kentucky in 1880, Griffith considered himself a playwright, not a film actor, but he debased himself (such was the early prejudice against the "flickers") to earn a steady $5 a day. He began acting in motion pictures in 1907, playing a lumberjack in Edwin

S. Porter's *Rescued from an Eagle's Nest*, a melodramatic one reeler about an infant stolen away from its family by a large black bird. Only a year later, Griffith directed his first film, for the Biograph Company, *The Adventures of Dollie*, another melodramatic one reeler about a stolen infant—this time by Gypsies. In the five years between 1908 and 1913, Griffith directed over 150 films. This experience gave him the opportunity to experiment weekly, to note the results, and then to polish and develop the most effective elements in his next week's film.

Griffith rejected the notion of the "standard shot" then supported by the Motion Picture Patents Company. According to the Patents Company's reasoning, why should the public pay to see half an actor when it could see the whole figure for the same price? Griffith developed a whole series of shots from extreme close-up (of a single object, face, or feature) to extreme far shot (a vast panorama). Griffith discovered that the emotional content of the shot, its structural role in the plot, and the human and thematic information that it conveyed must determine the composition of the frame, not any maxims based on an erroneous analogy with the stage. Griffith also developed the rhythms of editing to link the shots: when to cut languidly to give the impression of quietness and ease, when to use quick cuts to give the impression of tension and speed, when to cut subjectively to reveal a character's thoughts or intentions.

Griffith developed the technique of masking, in which a part of the screen rectangle is blacked out to focus attention on some circle, oval, horizontal plane, or other shape of light and image within the frame. He developed the irising effect, in which a scene began with a small circle of light within the frame and then expanded to fill the screen; and he ended a scene in the reverse manner, shrinking the full image into a tiny point of light. Griffith began integrating the pan shot (a horizontal or vertical movement of the camera on its fixed tripod) and the travelling shot or tracking shot (the whole camera moving to accompany the action) into his emotional and narrative screen language. The significance of Griffith's contributions was not that he invented these shots (many had been used before) but that he synthesized them all into a coherent and powerful way of telling a story.

Griffith's influence on acting

In addition, Griffith improved screen acting. No longer would actors demonstrate their emotions with grandiose gestures borrowed from the stage. Actors became people, subtly responding rather than gesticulating. Many of the most important and accomplished silent actors worked first for Griffith—Mary Pickford, Lionel Barrymore, Lillian and Dorothy Gish, Robert Harron, Henry B. Walthall, Erich von Stroheim, Richard Barthelmess, Mae Marsh, Blanche Sweet, Donald Crisp. Griffith also began solving the problem of making the indoor world look as real, as three-dimensional, and as fully textured as the outdoors. His sets gradually became more detailed and developed. His lighting gradually overcame the flatness of earlier indoor scenes that were really filmed outdoors (in most early films sunlight was diffused with muslin to light an indoor scene). The tone and texture of Griffith's lighting began to capture the feeling of indoor life.

Griffith's one reelers were of various types. He excelled in melodramas—the attack of vicious, irrational forces on the good and innocent. In *The Lonely Villa* (1909) and *The Lonedale Operator* (1911), Griffith developed the climactic chase, in which the husband or suitor arrives just in time to save his wife, children, or sweetheart. The switching from attacker to victim to rescuer—and the tense, nervous rhythms of Griffith's editing—played important parts in these exciting climaxes. Griffith also made domestic comedies such as *Mr. Jones Has a Card Party* (1908) and *Effecting a Cure* (1910), which laughed at the marital conflict between husbands who liked to drink, play cards, and woo the ladies and their moralistic wives who were following the principles of temperance advocate Carrie Nation. He made rural, folksy comedies such as *Examination Day at School* (1910) and *A Country Cupid* (1911), which examined the gentle tenderness of the country life that Griffith knew and loved so well. And he made social dramas such as *A Corner in Wheat* (1909) and *The Musketeers of Pig Alley* (1912), which examined such social and moral problems as the selfish gluttony of the rich, the pathetic poverty of the farmer, and the vicious corruption in city slums.

Despite the temptation to treat Griffith as simply a master technician, the motion picture's first, he was also its first intellectual and poet. Griffith was the first to realize the motion picture's potential for presenting a personal interpretation of human values and social assumptions. If Griffith's moral system was uncomfortably sentimental and romantically naïve, it was undeniably consistent throughout his career, and he was passionately, sincerely committed to it. Griffith believed in the family, the home, virginity, fidelity, social peace, the preservation of tradition, and the equitable distribution of material wealth and human necessities. His models of virtue were consistently women, or rather girls—blonde, thin, angelic, frail, usually in their teens. The Griffith ideal was gentleness, and the Griffith moral centre was the heart. If people treated each other gently, lovingly, there would be no wars, no poverty, and no vice.

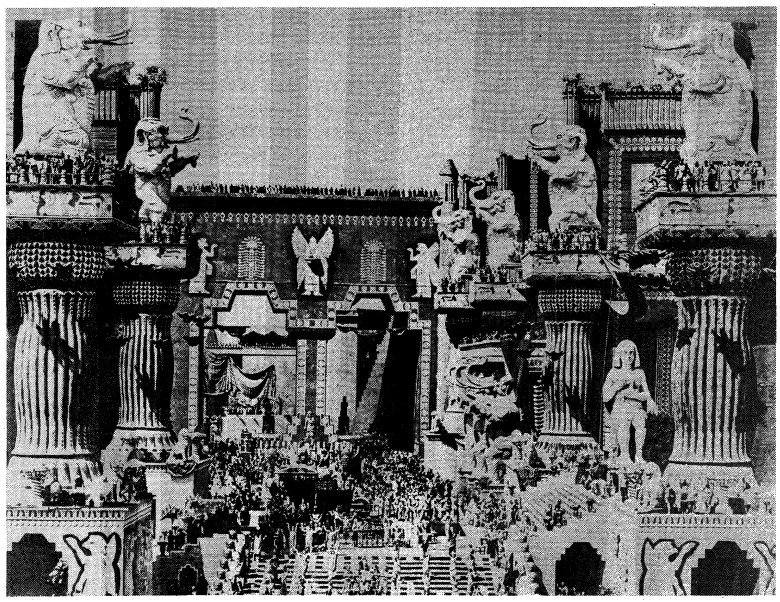
Griffith's moral system

But the Griffith gentle ideal inevitably collided in his films with the brutal realities of human existence. This reality could take the form of either a specific human menace or a general social evil. Griffith treated moral villainy as psychological depravity. The man who wishes to molest an innocent girl, disrupt a happy home, or grab as much money for himself as possible is a psychopathic beast. He is violating the essential human laws that make life livable. And general social problems, such as wars, poverty, slums, drunkenness, and strikes, were merely the consequences of this selfish, demonic lust for more power or money than any man has the right to have.

In his late one reelers, Griffith was able to give his sentimental moral system vitality and texture. First, there is a lot of sarcastic comedy in the early Griffith films, a sly delight in debunking moralistic, prudish ladies who are so ugly that they can have nothing to do with men except make them miserable. Second, there is a genuine tenderness for simple, rural people, a feeling for gentle, smiling faces that shows itself in *The New York Hat* (1912) and a feeling for the warmth and richness of the soil that shows itself in the loving, languid shots of the fields and planting in *A Corner in Wheat*. Third, Griffith's use of objects, such as hats, wrenches, telephones, pistols, and locomotives, communicated more and more information in a believable, economical, effective way. This use of objects in the one reelers would later develop into Griffith's concrete symbolism—the use of animals to convey emotional states and human attitudes, of flowers to convey the sense of beauty, fragility, and mortality.

Birth of a Nation and Intolerance. Griffith reached the peak of his career with two mammoth feature films, *The Birth of a Nation* (1915) and *Intolerance* (1916). Griffith left Biograph in 1914 when the company refused to make feature films. Following the lead of *Queen Elizabeth* and the Italian spectacles of 1912–14 (*Quo Vadis?, Cabiria, Anthony and Cleopatra*), Griffith wanted to make long films. He took his talented cameraman (G.W. "Billy" Bitzer) and actors with him, leaving Biograph so desolate that this greatest of all the Patents Company studios soon collapsed. Biograph's second most important director, Mack Sennett, had "gone independent" the year before.

The Birth of a Nation was a story of the gentle pre-Civil War South (Griffith's ideal of a serene, sensitive social system) and its violent demise in the brutal war and horrifying postwar Reconstruction (Griffith's symbols of social chaos and disruption). Out of the ashes of this struggle, a new nation is born, symbolized by the marriage of the Southern soldier Ben Cameron (Henry B. Walthall) and the Northern girl Elsie Stoneman (Lillian Gish). Within this overall symbolic framework, Griffith fills the screen for over three hours with scenes of spectacle, tender human details, careful reconstruction of the era's architecture and costume, exciting melodramatic suspense, and a grandiose allegorical image of the future of humanity.



The temple of Babylon sequence from *Intolerance* (1916), directed by D.W. Griffith.
By courtesy of the Museum of Modern Art Film Stills Archive, New York

Impressiveness of *The Birth of a Nation*

It is difficult to say whether *The Birth of a Nation* is most impressive for its huge spectacle or its tiny details, its exciting suspense or its intellectual commitment. Despite the oversimplifications in the film's theme, it was the most massive piece of cinematic thought that had yet been attempted. Among the huge scenes are the sweeping shots of Civil War battles and the detailed historical facsimiles in which Griffith carefully re-creates Ford's Theatre, the Appomatox Court House, and the South Carolina legislature. The film's gentle, touching moments include the early scenes in the prewar South (Griffith uses playful puppies and kittens as symbols), the wooing scenes between Ben and Elsie (symbols include a cooing dove and a remarkably Freudian bedpost), and the painful, sorrowful return of Ben from the war, when he discovers that his once thriving town is desolate and ruined.

Griffith was especially careful in his handling of the scenes of suspense. There are four of them in the film, each one more developed and fuller than the one preceding it. In the first of them, the Cameron home (and women) is attacked by a renegade band of Northern soldiers—all black except for their white leader. In the second the youngest Cameron girl is pursued through the woods by Gus, a vicious black man whose threats of marriage drive the girl to her death. And the final two occur simultaneously at the film's climax, revealing Griffith's complete awareness of rhythmic construction. While the evil mulatto politician Lynch tries to force Elsie into marriage, the other members of the Cameron family are trapped inside a little cabin, mercilessly attacked by Northern soldiers. Both last-minute rescues take place within the same last minute, giving the film one of the most exciting, breathless climaxes in film history. Griffith underscored the film's excitement with two additional contributions to film "rhetoric"—music (Griffith released a musical score cued to the film's effects for the theatre's piano or orchestra) and colour tinting (Griffith tinted sequences in blue, orange, yellow, green, or red to intensify their emotional tone).

Throughout *The Birth of a Nation* the theme of miscegenation and racial conflict runs strong. The film's villains are mulattoes, who want to seize power (and white womanhood), and a vain, white Northern liberal, Congressman Stoneman, who, out of misguided ideas of social justice (and because of pressure from his mulatto mistress), wants to give power to the black man. In typical Griffith fashion, erring political ideas can be traced to sexual corruption. The valiant heroes in the film's last-minute rescue, the forces of good in the film, are members of the Ku Klux Klan, led by Ben Cameron, the film's protagonist. Although Griffith's attitude was unquestionably racist, his primary target was not the black man himself but the mixture of the races and the destruction of established political power. According to Griffith's interpretation of the Civil War, a group of selfish, lustful, power-hungry men would simply not let a very beautiful, peaceful social order alone. Griffith realized that the very importation of black slaves planted the poisonous seed that would eventually destroy the South, but he could not help looking at the social system that was built on slavery with warm affection. This interpretation of history predictably caused a stormy social reaction, protests and riots in some cities, censorship and banning of the film in others. The reaction to the film was not only noisy but profitable, its receipts soaring into unheard-of millions of entertainment dollars.

Negative reaction

Griffith was more stung by the social controversy than pleased with the profits. He thought he had painted a balanced picture of the black man (*i.e.*, he had sketched a few "faithful souls" who were really "Uncle Toms," or blacks sympathetic to the whites) and had softened the racial bias of his source, Thomas Dixon's novel *The Clansman*. This anger led him to write a pamphlet on "the freedom of the screen." It also fired him for his next major project, *Intolerance*, of which he felt himself the victim. This film wove together four stories, unified by the theme of intolerance through the ages and a recurring metaphoric image, a mother rocking a cradle, who symbolizes eternal humanity fulfilling its evolutionary destiny through the ages. In each of the four ages, there is intolerance (or injustice). An intolerant religious leader of Babylon envies Belshazzar, the country's ruler, and betrays his own country to the Persian invaders. The intolerant Pharisees drive Christ to the cross. The intolerant Catholics slaughter the Huguenots. And intolerant modern bigots drive sympathetic people to the brutal city, drive women into prostitution, drive mother and child apart, and almost drive an innocent man to the gallows. In all four stories, differences of religious belief and moral opinion produce the catastrophe.

The Babylon and modern sequences are the most developed. Most significant about Griffith's Babylon episode were the immensity of its sets, the detail and size of its battle sequences, the exuberant, jaunty comedy of the tomboyish "Mountain Girl" (Constance Talmadge), who loyally tries to protect Belshazzar, and the tender, ideal love of Belshazzar and his queen, symbolized by a tiny cart of roses pulled by delicate white doves. For Griffith, Babylon paralleled the prewar South—an idyllic land of peace and comfort—destroyed by the forces of rapine and death that sprung from the envy and ambition of a few men. The film's modern sequence was far more

complex. Griffith condemns social-reforming ladies who turn to reform when they can no longer turn a man's fancy. But Griffith equally condemns the rich owner of a factory who is so insensitive to human needs that he starves his workers into a strike and then mows them down with his own private army of detectives. Griffith also shows the poverty in cities, the illicit means of making money when there are no legitimate jobs, and the resulting disruption of the essential human unit—mother, father, and child.

Interwoven climaxes of *Intolerance*

Griffith culminates all four stories with four interwoven climaxes. The Mountain Girl races home to warn Belshazzar; Christ drags the cross toward Calvary; a Huguenot races home to save his family from slaughter; the governor and the mother race to save her innocent husband from execution. The breathtaking effect of leaping from century to century, chase to chase is overwhelming. But, after the breathtaking excitement, Griffith ends the film, as he did *The Birth of a Nation*, with an allegorical pageant. Angels appear in the heavens; prisons dissolve into fields of flowers; roses bloom in the rocking cradle. As in the earlier epic, this pageant is irrelevant to the film's story, a sure sign of the simplicity of Griffith's ideological system. Whereas Griffith was very able with human details and melodramatic excitement, he was very weak at abstract speculation about the nature of man and the direction of history. Beneath the immense structural framework of *Intolerance* was the trite homily that nasty people are not nice and may do nasty things.

Later films. The narrowness of Griffith's vision was at least partly responsible for his decline in the years following *Intolerance*. The attitudes and films of other directors became far more complex and popular once Griffith had shown them the way to incorporate ideas into exciting motion-picture entertainment. Movie audiences of the 1920s became too sophisticated and too materialistic for Griffith's sentimental simplicities. The later Griffith films were not, however, without their virtues: *Hearts of the World* (1918) was a tender study of war-torn Europe and an attempt to redress Griffith's injury to the black man; *Broken Blossoms* (1919) was a poetic story of the London slums in which a gentle Chinese man, a delicate girl, and a brutal boxer all tangle and die; *Way Down East* (1920) was an affectionate look at the rural United States, culminating in a spectacular rescue sequence when Richard Barthelmess saves Lillian Gish as she floats on an ice cake toward the murderous falls; *Orphans of the Storm* (1921) contained careful Dickensian evocations of the French Revolution (Dickens was Griffith's favourite author); *America* (1924) used luminous, soft-focus photography. But the later Griffith films lacked the spark and the opportune historical timing of his two epics. The motion-picture industry caught up with Griffith's pioneering leadership very quickly. After two disastrous attempts with synchronized sound, Griffith was forced into involuntary retirement in 1930.

The early age of comedy. The dawn of the age of Griffith was also the dawn of the age of comedy.

Mack Sennett. In 1907 Mack Sennett began working in films. He worked for Griffith at Biograph between 1908 and 1912. And out of Sennett's films came, in turn, such pupils as Charlie Chaplin, Fatty Arbuckle, Marie Dressler, Mabel Normand, Ben Turpin, Billy Bevan, Harold Lloyd, Harry Langdon, and Charlie Chase. Sennett left Biograph in 1913 to direct comic films for an independent company, Keystone. Sennett's Keystone comedies were improvisational, spontaneous, and very, very fast. Sennett packed his cameras, performers, and crew off to some nearby location, shot some footage, then returned to his little studio to finish the interlinking continuity. The dominant technique of the films was speed; the dominant activity was hitting, kicking, bashing, and crashing; and the dominant intellectual attitude was that nothing should be taken seriously, especially something that took itself seriously.

Sennett's comedies reduced men to machines. His distant shots reduced people into things, making them close cousins to the other machines that zip through Sennett films—automobiles, airplanes, bicycles, boats. Sennett increased the impression of mechanization with his mechanical camera tricks—undercranking, reversal, combining animated drawings with live action. The Sennett world at Keystone was one of pure mechanical fun, total irrationality, the surprises of incongruity, impossibility, and the *non sequitur*. Motion for the sake of motion and speed for the sake of speed were the essential cinematic delights. The Keystones reduced serious social questions to absurd comic stereotypes—the impolite, but often funny, jokes about Jews, Germans, blacks, and the like implied an honesty and easiness about racial and class differences in U.S. society that later disappeared. Among Sennett's other comic targets were rich snobs, moralistic ladies, preachers, professors, and, ultimately, the police—who were the dizziest, daffiest, most incompetent toy soldiers on the Sennett lot. Sennett claimed that his primary target was pretention. His special delight was applying a kick in the pants, a brick in the head, or a pie in the face as an antidote to pretentiousness. But Sennett delighted in the kick, brick, or pie for its own sake as well.

Mack Sennett's Keystone Cops in a typical scene from *On Edge* (*c.* 1915).
From the Penguin Photo Collection

Sennett's parodies of other films

Sennett also enjoyed parodying other films, which he often found absurdly pretentious. Those Keystone films that had a plot at all (many were simply fighting, chasing, and running around) usually made fun of other plots. He parodied the story of the pure country girl engulfed in the wiles of the big city in *Tillie's Punctured Romance* (1914); he parodied westerns in *A Versatile Villain* (1915); he parodied Griffith's last-minute rescues in *Barney Oldfield's Race for Life* (1914) and *Teddy at the Throttle* (1917); he parodied romance and young love in the "Fatty and Mabel" series, and romance and domestic bliss in the "Charlie and Mabel" series. In his later films, Sennett produced hilarious parodies in which the cross-eyed comedian Ben Turpin reduced the great stars of the era (*e.g.*, Rudolph Valentino, Erich von Stroheim, and William S. Hart) to the absurd.

Sennett's style of comedy was a kind of summary of what had preceded it—a synthesis of Georges Méliès' mechanical tricks with the camera, of the jests and pranks stemming from *L'Arroseur arrosée*, and of the frantic chases conceived by Zecca and copied by numerous imitators. Sennett, in turn, passed on this delight in irrational surprise and physical movement to his many descendants.

Early Chaplin. The man who was to take comedy in a new direction, Charlie Chaplin, began working as one of Sennett's mechanical toys. It was Sennett's definition of comedy as speed and machine that helped define Chaplin's sense of comedy as human and supple. Although Chaplin consistently felt that Sennett did everything too quickly, that his own style of comedy had no place in Sennett's world of clockwork puppets, Chaplin's year at Keystone was extremely valuable. Knowing nothing about film acting, editing, and construction, Chaplin became so proficient within a year that he directed some of

the best Keystone films that Sennett produced—*Dough and Dynamite, The Rounders, The New Janitor* (1914). In his second Keystone film, *Kid Auto Races at Venice* (1914), a delightful little improvisation about a troublesome tramp who interferes with a camera crew that is trying to take pictures of the races, Chaplin first wore his tramp costume of baggy trousers, enormous shoes, tight vest, derby hat, and a cane. Throughout his Keystone year, Chaplin refined both the costume and the character of the tramp.

Genesis of Chaplin's tramp

Another valuable Keystone lesson was Chaplin's discovery of objects. If Sennett turned people into inanimate objects, Chaplin turned objects into living beings. In *His Favorite Pastime* Charlie fights with a swinging saloon door; in *Mabel's Married Life* he fights with a boxing dummy; in *Dough and Dynamite* he turns baking dough into bracelets, boxing gloves, and bricks. The Chaplin comic technique was built around this transposition of objects—turning one kind of thing into something else. This transposition would dominate such later films as *The Pawnshop* (he turns an alarm clock into a can of tuna, a sick patient, an infected tooth, a precious jewel), *One A.M.* (he turns a collapsing bed into a dancing and sparring partner), *The Gold Rush* (he turns a boiled shoe into a gourmet feast), and *The Great Dictator* (he turns the globe of the Earth into an ethereal balloon and a performing seal's toy ball).

At Keystone, Chaplin also began to perceive that the world was not as simple and as silly as Sennett made it. The police were not merely incompetent clowns; they enforced the law. Further, they enforced laws that had been written by people who owned property to protect property. Chaplin began to realize that the essential human need was not how to get one's hands on a pie to hurl in someone's face but how to get one's hands on a loaf of bread if one is hungry. Even in the Keystones, Chaplin began building his films around more serious social themes, such as the need for material comfort and for human companionship or the contrast between the man of nature and the regulations of society. Chaplin could not develop many of these ideas with Sennett. At Keystone he remained a tough, gritty, jaunty tramp who kicked anybody in the pants who kicked him—and often even if they had not. To develop his own ideas, Chaplin had to go his own way.

Use of serious social themes

In 1915 Chaplin received a one-year contract to make primarily two-reel comedies for Essanay (most of his Keystone films were one reelers) at a total salary of $75,000. In 1916 he received a contract to make 12 two reelers for the Mutual Film Corporation at a total salary of $670,000. And in 1918 he signed a contract with the First National Company to make nine pictures (of two reels or more) for $1,000,000, plus a percentage of the profits. These astounding figures indicate three things: Chaplin's enormous popularity, the conversion of the motion picture from cheap novelty to big business, and Chaplin's success at going his own way. Many of Chaplin's early Essanay films were in the Sennett mold, such as *His New Job* (a parody of pretentious films and film stars) and *A Night Out* (a physical film of drunken stunts with Ben Turpin). But in *The Tramp* (1915) Charlie played a man of nature who sided with the pretty girl against his fellow tramps and who lost her when his rival, the handsome city slicker, returned to her farm. In *The Bank* (1915) he played a poor janitor in the institution dedicated to money who performs a heroic action only to wake to discover that his heroism was a dream. He demonstrated in *Police* (1915) that moralistic preachers are hypocrites pursuing their own gain and that policemen are more interested in their vanity and self-respect than in helping anyone in trouble. Despite the growing social commentary in his films, Chaplin never failed to be funny, combining serious implications with hilarious uses of props, mime, and human conflict. His mastery of the cinematic medium also became more controlled. He realized that his comic style was best suited to the medium shot with a very spare use of cutting, since cutting would detract from his delicate mime.

At Mutual the Chaplin films took a further leap in cinematic control, structural unity, and the integration of comic action with serious content. He reached perfection in the short film. Chaplin in *The Rink* (1916), for example, glides magnificently on roller skates, an act repeated in *Modern Times*, using his physical grace to avoid his rich enemy and the police. In *The Floorwalker* (1916) he miraculously captures the store's thief, who also turns out to be its owner. A frequent motif of the Chaplin films is that the boss steals from his own business, but Charlie is constantly accused of such crimes since he appears so shabby and unrespectable. If the rich man does not steal, he is often guilty of callousness and brutality. Whereas the actress Edna Purviance—blonde, sweet, mild—plays the consistent ideal of gentleness and beauty in the Mutual films, Eric Campbell—huge, brutish, mean—plays Charlie's consistent foe. Eric, the physical heavy, also becomes the societal heavy—a rich man, foreman, or boss. He was the embodiment of social success, as Charlie was of social failure. Chaplin's comic reversal is that the tramp's worthlessness is more valuable in human terms than the success of the characters acted by Campbell.

Comedies produced for Mutual

Some of the final films of the Mutual group are the most pointedly serious. In *The Immigrant* (1917) Chaplin comically recounts the problems of immigrants coming to the United States. *Easy Street* (1917) is the most bitter and mature of the group, a satiric deflation of utopias and idealistic clichés of social progress and human nature. In this film Chaplin comically demonstrates that the problems of living—poverty, hunger, drug addiction, animal brutality—are larger than any conceivable social solutions; the police who try to enforce the law fail, and the preachers who urge men to be better than they are, or can be, are mere dreamers. The irony inherent in the title of *Easy Street*, which is really a slum where life is very hard, is the kind that propelled Chaplin into his next creative state (see below *Comedies of the 1920s*).

Maturity of the U.S. film industry. Griffith, Sennett, and Chaplin all reached artistic maturity in about 1915. The same is true of the U.S. film industry as a whole. While the Europeans were more absorbed with World War I than motion-picture entertainment, the U.S. film industry pulled itself out of its infancy of commercial chaos. The Patents Company collapse began in 1915, defeated both in the courts and in the marketplace. And 1915, the year of *The Birth of a Nation*, was also the year when feature films seized control of the film industry and film audience. It was the year when Samuel Rothafel ("Roxy") first opened a plush "movie palace" on Broadway in New York City to show the new mammoth feature films. It was the period when the star system captured the public's imagination—as Chaplin's soaring commercial value and also that of Mary Pickford (she, too, signed a $1,000,000 contract) revealed. Other early stars included John Bunny (a fat comic whose distaste for propriety and the moralistic nastiness of his skinny wife, Flora Finch, foreshadowed the work of W.C. Fields), G.M. "Broncho Billy" Anderson (half owner of Essanay and the United States' first cowboy star), Theda Bara (the seductive sexual siren who added the word vamp to the English language), and Douglas Fairbanks (the energetic American go-getter who used his gymnastic talents to achieve the classic goals of the American Dream—love, money, and success).

The star system

The first fan magazines had appeared by 1915; so had the first motion-picture serials (*e.g., The Adventures of Kathlyn, The Perils of Pauline*), and the first reviews of films were published in the popular press. This was also the era of the first loud moralistic reaction against the sinfulness of screen stories and screen actors. It was the decade when Hollywood became the movie-production capital, when the first movie moguls became millionaires, and when the studio system became the standard means of film production. Thomas H. Ince was one of the first producers to organize a motion-picture factory, establishing shooting schedules, careful budgets, and approved shooting scripts before filming could begin. In 1916 Adolph Zukor assumed the presidency of the powerful organization that eventually called itself Paramount Pictures. In 1918 an outraged group of theatre-chain owners,

Rudolph Valentino as a romantic swashbuckling hero in *The Four Horsemen of the Apocalypse* (1921).
By courtesy of ©–1921, Metro-Goldwyn-Mayer Inc.

who found themselves pressured by the block-booking tactics of Zukor and the other powerful producers, formed the huge First National Exhibitors Circuit. In 1919, Chaplin, Griffith, Fairbanks, and Pickford combined their assets and talents to form the United Artists Corporation. And, by 1925, Goldwyn Pictures Corporation, Universal Pictures Company, Columbia Pictures, Warner Brothers, the Fox Company, and Metro-Goldwyn-Mayer, the biggest of Hollywood's studios, were all in business.

The aim of the studio system was to regularize production procedures, distribution methods, and the product itself. Films became less individual in style and theme; the Griffith days of improvising a whole film out of the creator's head were gone except in the work of Chaplin and a few others. Hollywood films tended to become increasingly categorized into recognizable genres: westerns (either with stars such as William S. Hart and Tom Mix or mammoth productions such as James Cruze's *Covered Wagon*, 1923, and John Ford's *Iron Horse*, 1924); romantic swashbuckling in faraway lands and times (the films of Valentino, Ramon Novarro, the later Douglas Fairbanks); idyllic, gentle rural romances (Mary Pickford's motion pictures, Henry King's *Tol'able David*, 1922); domestic stories of sexual seduction and marital temptation, either comic or melodramatic (von Stroheim's *Blind Husbands*, 1919, *The Devil's Passkey*, 1920, and *Foolish Wives*, 1922; deMille's *Male and Female*, 1919, and *Why Change Your Wife?*, 1922; Lubitsch's *Marriage Circle*, 1924, *Three Women*, 1924, and *So This Is Paris*, 1926); deMille's biblical spectacles produced after the sexual scandals of the 1920s when he was converted from the bedroom to the cross; war films with pacifist–humanist messages (King Vidor's *Big Parade*, 1925; Raoul Walsh's *What Price Glory?*, 1926); and comedies (the features of Chaplin, Keaton, Langdon, and Lloyd; two reelers from Sennett with Billy Bevan, Ben Turpin, and Harry Langdon; two reelers from Hal Roach with Charlie Chase, Will Rogers, "Our Gang," Snub Pollard, Stan Laurel and Oliver Hardy).

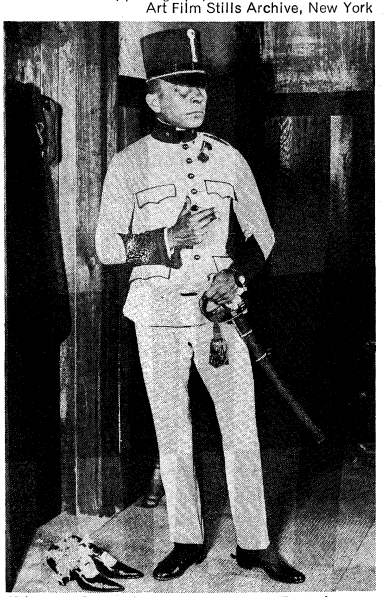
DeMille. Cecil B. deMille was one of the most successful of Hollywood directors at making hack formulas look very special. His bedroom comedies contained an air of naughtiness (many passionate stares) but very little bedroom activity and less comedy. His historical romances featured grandiosely unhistorical sets, costumes, and plot points. His religious spectacles featured spectacular sinfulness mixed with a dash of righteousness and a lot of self-righteousness. DeMille was a descendant of the Griffith tradition, and many of his films, like Griffith's *Intolerance*, began with the warning (on title cards and, later, from the booming voice of a narrator) that the film to follow was to be taken very seriously. The leaden pace, the huge sets, and the stilted seriousness of the acting styles in deMille's films continued to reassure the audience that what it was watching was "highly significant." In his blend of dramatic emptiness and overstated moralizing, deMille set a pattern for successful commercial film making that survived as long as he did. Significantly, his two periods of greatest popularity and influence, the 1920s and the 1950s, were the two decades most devoted to surface splash and bombastically ornate and ponderous production values.

Griffith's influence on deMille

Von Stroheim. Erich von Stroheim, on the other hand, was Hollywood's *enfant terrible*—a rebel in production values (decor, sets, costumes, lighting), moral assumptions, sexual attitudes, and personal temperament. Whereas deMille used passionate stares, von Stroheim captured genuine sexuality with shots of the details of the human body, and, where deMille turned desire into breast heaving and panting, von Stroheim consistently described human desires in animal terms using animal imagery. Unlike deMille, he insisted on absolute historical accuracy and minutely perfect atmospheric detail (much to the chagrin of his financial producers). Furthermore, while deMille's lighting was flat and even, von Stroheim's, echoing the character's passions, was grotesquely shadowed, with eye-shattering contrasts of dark and light. Von Stroheim acted in many of his own films, playing the central character from which the film's perversion and grotesqueness emanated. He played the elegant Prussian—fastidiously dressed, complete with military uniform, monocle, and white gloves, elegantly, suavely mannered, with complete knowledge of every social form and grace. But beneath the external elegance the man was a lecherous, sensual animal who assumed that other human beings were merely a means to his pleasure. For the film director to cast himself in such a role was central to the spirit of the films he directed.

Despite his bizarre style and themes, von Stroheim's sexual films for Universal made money. So did his strange version of *The Merry Widow* (1925) for Metro-Goldwyn-Mayer, which turned an operetta into a film of perversion suggesting the Marquis de Sade rather than Franz Lehár. But *Greed* (1924), von Stroheim's attempt to shoot a literal translation of Frank Norris' *McTeague*, a brutal novel of human corruption and the animal drive for wealth, was a catastrophe. Originally, von Stroheim wanted the film to run 42 reels (about eight hours); he agreed to cut it in half. Irving Thalberg, the new head of production at Metro-Goldwyn (Louis B. Mayer's name had not yet been added to the studio's), took the film out of his hands, ordered the film be cut to 10 reels and

By courtesy of Universal Pictures, © Universal Pictures; photograph, The Museum of Modern Art Film Stills Archive, New York



Eric von Stroheim as the elegant Prussian in *Blind Husbands* (1919).

the remainder of the negative destroyed. The result is a film with powerful images, symbols, and episodes but one that is incomprehensible as a whole. After the silent era von Stroheim only acted in films, most memorably as von Rauffenstein in Jean Renoir's *Grand Illusion* (1938) and Max in Billy Wilder's *Sunset Boulevard* (1950).

The Motion Picture Producers and Distributors of America. The sensuality in the films of von Stroheim (and the pretense of sensuality in deMille's) was an indication of the mood and temperament of the audiences of the 1920s. Gone was the sentimental sweetness of Griffith's world of purity. The bizarre doings of the stars off screen—the Pickford–Fairbanks marriage three weeks after her divorce, the mysterious deaths of the actor Wallace Reid and the director William Desmond Taylor, Fatty Arbuckle's involvement in a rape and homicide—began to provoke as much public attention as the stars' activities on screen. In an attempt to avoid federal censorship and to placate moralistic critics, Hollywood created the Motion Picture Producers and Distributors of America in 1922, a public relations organization that sought to repair Hollywood's image in the press and Hollywood's activities both on screen and off screen. The man selected to head the organization, Will H. Hays, former postmaster general of the United States, served as Hollywood's moral leader for over 20 years. The "Hays Office" (as the MPPDA became colloquially known) took only informal and advisory censorship actions in the late 1920s; it would reach the height of its power a decade later.

Changes in the 1920s

Comedies of the 1920s. Hollywood's most vital and enduring product of the 1920s was comedy. Chaplin had been joined by three rivals—Buster Keaton, Harold Lloyd, and Harry Langdon—who surpassed him in productivity and equalled him in popularity.

Chaplin. Chaplin's longer films became much more conscious of exactly who the tramp figure was and how he contrasted with the social institutions and assumptions of the day. In the First National films of 1918–23 (varying in length from two to six reels), Chaplin plays several social roles (soldier, father, preacher) that are inconsistent with the character of his outsider tramp, in order to examine the differences between the tramp and the role. He compares the tramp's existence to that of a little mongrel named Scraps in *A Dog's Life*, showing that, for both Charlie and the mongrel, existence is an animal struggle for the necessities of survival—food, shelter, and companionship. In *Shoulder Arms*, the tramp attempts, without success, to adapt himself to the rigours of military routine and discipline. He eventually performs heroic military feats, though in his own unroutine, improvisatory way. He also awakens from a dream to discover that his heroism has been merely a romantic fantasy (a typical Chaplin motif). Charlie exposes in *Sunnyside* the dirt, selfishness, and dishonesty beneath the rural-U.S. ideal of small-town life. In *The Kid*, a rich film combining comedy and delicate pathos, Chaplin defines genuine familial love as opposed to society's definition of it with the tramp's relationship to his adopted child (Jackie Coogan). Chaplin burlesques both religious and legal definitions of righteousness in *The Pilgrim*, revealing that the tramp's personal moral system is superior to the assumptions of both church and state.

Chaplin's full-length features

When he began making full-length features for United Artists, Chaplin's control of cinematic distance (his awareness of the quiet intimacy of his style that required an equally intimate use of the camera), his knowledge of who he was, his awareness of societal assumptions, and his ability to make serious inferences with hilariously comic routines reached perfection. *A Woman of Paris* (1923), his first film for United Artists, was a daring departure. Chaplin directed it, and there was no tramp character; he played only a tiny walk-on part. The film starred Edna Purviance, the actress who had played Charlie's female foil for eight years, in the role of a woman who enjoyed the luxuries of the high life as the mistress of a sophisticated Parisian sensualist, acted by Adolphe Menjou. The film contrasted the purity of provincial life with the temptations of the city, the lure of materialism as opposed to the need for love. Remarkable about this much-repeated Chaplin theme of the conflict between love and prostitution was his refusal to moralize about the sinfulness of adultery; instead, he asks what other chance a poor woman has to enjoy the delights and comforts of opulence. Also remarkable was his building the film around concrete objects that implied so much more than they said—pipes, handkerchiefs, necklaces, napkins, a saxophone. The film's ironic restraint and clever comment became a major influence on the future direction of polite comedy—particularly on the films of Ernst Lubitsch.

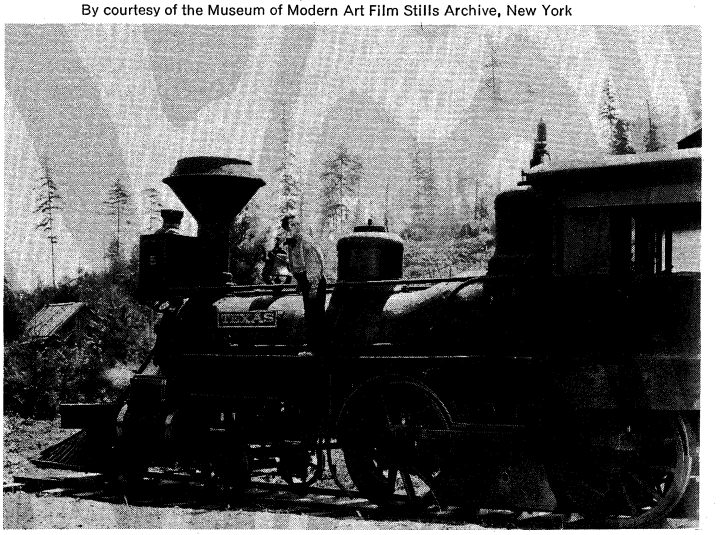
In *The Gold Rush* (1925), Chaplin's favourite film and perhaps his best, Charlie prospects for gold in the frozen north. Despite the many hilarious scenes, every comic sequence is built around the question of human survival. The search for gold (a cold, dead metal) in a frozen desert (equally cold and dead) is a murderous, inhuman occupation. This coldness produces one of the great moments of pathos in all of Chaplin's films, when the woman of his dreams, who works in the dance hall, stands him up on New Years' Eve, leaving the sympathetic, gentle tramp completely desolate and alone. The tramp's gentleness and warmth are incongruous in the film's icy world. Similarly, *The Circus* (1928) contrasts Charlie's gentleness with the callousness of the circus owner and the romantic suaveness of the trapeze artist who steals the girl away from the tramp. The film also contrasts the spontaneous hilarity of the tramp's style of comedy to the usual circus-clown routines, which the tramp performs badly—as he does all routines.

Buster Keaton. Chaplin's opposite as a comic actor and director was Buster Keaton. He performed stunts with his whole gymnastic body rather than supple pantomime with individual parts of the anatomy. His little body was contrasted against the huge world of nature (forests, seas, rivers, rain, fire), unlike Chaplin's contrast of man and society, and Keaton played against huge objects that dwarfed him (locomotives, ocean liners, herds of cattle, an entire army, the entire police force) rather than the small objects that Chaplin manipulated and the individual opponents that Charlie fought. Keaton's films depended on driving, suspenseful plots, as opposed to the circular thematic structures in Chaplin's films. Keaton delighted in parodying other films, and he delighted in far-fetched gags, crazy machines, and bizarre tricks with the camera itself. Keaton's style of comedy was faster, zanier, and trickier than Chaplin's slow, subtle, intimate style. Charlie smiled and winked and sighed, all of which the camera recorded slowly and compassionately. Buster's face was a classically frozen mask of inexpression, producing its meaning from contrapuntal contrast with the huge, complex, bizarre world that surrounded the face and body.

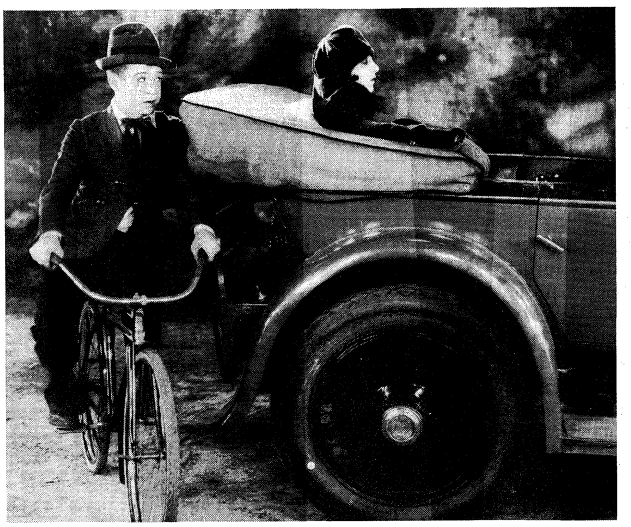
Keaton's comic style

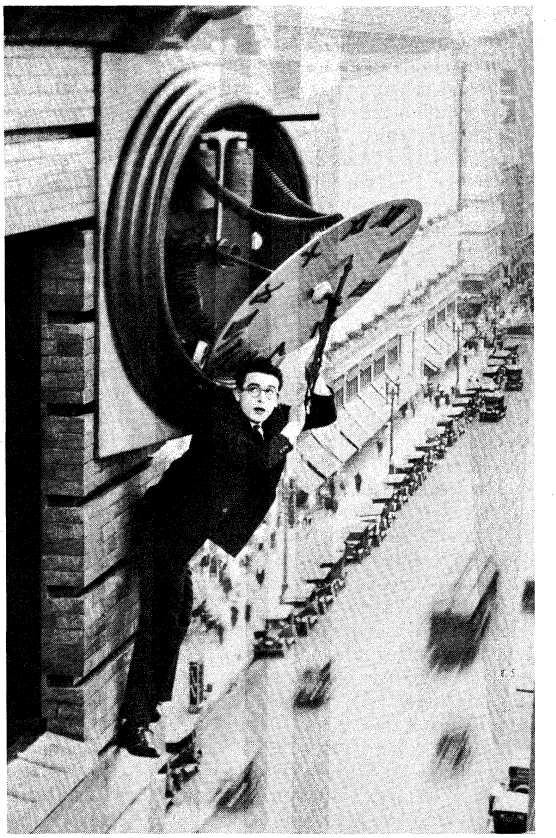
After a brilliant career in vaudeville (beginning at the

By courtesy of the Museum of Modern Art Film Stills Archive, New York



Buster Keaton in *The General* (1926).





Two American comics of the 1920s.
(Left) Harry Langdon in his typical role as the innocent clown in a scene from *Long Pants* (1927). (Right) Harold Lloyd in a "cliff-hanger" scene from *Safety Last* (1923).
By courtesy of (left) United Artists Television, Inc., (right) Harold Lloyd Foundation; photographs, from the Museum of Modern Art Film Stills Archive, New York

age of three), Keaton went to work as a supporting player for Fatty Arbuckle in 1917, much as Chaplin had served an apprenticeship with Mack Sennett. Between 1920 and 1923 Keaton made two reelers, among the most interesting of which are *The Playhouse* (a recollection of vaudeville days that exhibits Keaton's brilliant control of camera trickery), *The Paleface* (a parody of cowboy-and-Indian films), *The Electric House* and *The Scarecrow* (revealing Keaton's love of bizarre machines), and *Cops* (a driving film of cumulative mistakes in which Buster winds up the foe of the entire New York City Police Department).

From 1923 to 1929 Keaton made silent features that combined his sense of parody, his gymnastic stunts, his camera trickery, his crazy machines, and his understated character with a very strong sense of narrative drive and rhythm. Among his parodies of this period are *Our Hospitality* (1923), in which Keaton ridicules family feuds; *Sherlock Jr.* (1924), in which he mimics the famous detective; and *Go West* (1925), a parody of cowboy films, in which Keaton eventually succeeds in driving an entire herd of cattle to market, his sole aid in this achievement coming from "Brown Eyes," his lover in the film, who happens to be a cow.

Keaton plays a character who succeeds—despite his size, apparent insufficiency, and apparent effeminateness. He succeeds because his shortcomings are merely apparent. In *The Navigator* (1924), for instance, he succeeds in manipulating a huge ocean liner by himself; in *Battling Butler* (1926) he succeeds as a prizefighter; and in *Steamboat Bill Jr.* (1928) he succeeds in saving his tough father and soft sweetheart from a mammoth storm. His most striking success—as well as his best film—is *The General* (1926), in which Keaton singlehandedly defeats what seems like the entire Federal Army. He plays a Southern railroad engineer whose train is stolen by the enemy. He dutifully chases after his train because he loves it, and in so doing he accidentally performs romantically heroic deeds.

In the film's narrative suspense and its cynical deflation of heroism, war, and romance, Keaton creates his most unified comic epic. Keaton's apparent smallness and weakness are merely camouflage for his human potential, just as his apparently frozen face is merely camouflage for a pragmatic and active mind.

Harold Lloyd. Harold Lloyd is not a comic figure of camouflage at all. His internal personality shows on his face—in his smiles, his energy, and his horn-rimmed glasses. Lloyd plays the eager American who is energetic, anxious to succeed, and willing to compensate for his shortcomings with dash and speed. Lloyd's comedies put him in more familiar domestic situations than Chaplin's outsider looking at society or Keaton's elemental man fighting against the immensity of nature. Lloyd's goals are material success and romance, the typical goals of domestic comedy. His films are of various types. He is perhaps most famous for his comedies of thrills, his excitingly funny battles with gravity on the ledges and roofs of buildings, such as in *High and Dizzy* (1920), *Never Weaken* (1921), and *Safety Last* (1923). But Lloyd was also a master of a comedy of situations that built many laughs from either a reversal of character (in *Grandma's Boy*, 1922, he plays a coward who discovers courageousness) or a comic culmination of the character's supreme zeal and drive (in *The Freshman*, 1925, he pushes his way to eventual success in college). Although not as haunting and subtle as Chaplin, not as imaginative and surprising as Keaton, Lloyd comedies generate many laughs from lengthy, brilliantly constructed and developed gag sequences.

Harry Langdon. Harry Langdon, the fourth of the great U.S. comedians of the 1920s, was the opposite of Lloyd. Langdon was slow rather than fast, absent rather than aware, passive rather than active—a baby, not a man. Of the four famous clowns, Langdon was the stupidest, the least competent, the most effeminate, and the gentlest.

The plots of his comedies invariably plunge the male-fe-

male child-man into a dangerous or difficult environment where such foolish angels should fear to tread. He battles with desperadoes, vamps, tornadoes, gun molls, athletes, and gangsters, and he manages to win these impossible battles not because of his competence, like Keaton, or his push, like Lloyd, but simply because chance smiles on him. In *Tramp, Tramp, Tramp* (1926), for example, little Harry wins a cross-country walking race (despite such obstacles as virile male competitors, the vast countryside, the police, a chain gang, and a cyclone). In *The Strong Man* (1926) Harry rids a small town of its bootleggers and marries the girl of his dreams, who is blind (a suitable Langdon partner). He discovers in *Long Pants* (1927) that the girl of his dreams is a vicious, brutal gun moll, and little Harry crawls back home sheepishly to the comfortable reality of the small town (where this innocent really belongs).

The early documentary. An important development of the U.S. silent film was the birth of the documentary film. Robert Flaherty's *Nanook of the North* (1922), the first documentary to become a U.S. commercial success, was a close study of the life of an Eskimo. In its depiction of the way Nanook struggled daily to survive, the film also provided an awareness of the way it felt to be Nanook and the way life appeared to Nanook. The film's success allowed Flaherty to make other films for Hollywood, some alone (*Moana*, 1926), some with collaborators (*White Shadows of the South Seas*, 1928, and *Tabu*, 1931). Flaherty eventually found Hollywood antagonistic to his aims of observing and re-creating the experience of a unique way of life. He migrated to England, where he helped to steer the direction of that country's great documentary-film movement of the 1930s.

Success of *Nanook of the North*

By courtesy of International Film Seminars, Inc.; photograph, the Museum of Modern Art Film Stills Archive, New York

Scene from *Nanook of the North* (1922), documentary film directed by Robert Flaherty.

EUROPE

While Hollywood and the U.S. film dominated the commerce of the international film industry in the final 15 years of the silent-film era, the film industries of Europe began to recover after the devastation of World War I. European films during this period, rather than working for a conventional, standardized product, carried on many of the U.S. experiments of the Griffith era. Influenced by Griffith's composition and cutting, the Europeans compounded these cinematic techniques more complexly, personally, and innovatively than most U.S. producers and directors were willing or able to do.

Germany. The German film between 1918 and 1933 developed the effects of composition—the look, feel, and texture of the individual shot—so communicatively that many of the greatest German films contained the barest minimum of subtitles or none at all. German films demonstrated the emotional power of shadowy lighting, of architectural shapes and configurations, and of the subjective camera, often using the lens to mirror a character's inner vision of reality. Significantly, the German film was a studio film, shot indoors on a stage rather than outdoors in nature. Even outdoor scenes took place indoors, so technicians could perfectly control the effects of light, shadow, tone, and texture. This deliberate artificiality worked in favour of the most successful films of the era, contributing a feeling of stifling claustrophobia to stories of trapped people. These studio films necessarily put a heavy emphasis on decor, giving the designer great creative freedom in constructing his immense indoor cities—the sets of Hermann Warm, Robert Herlth, Walter Röhrig, and Walter Reimann were particularly notable. The emphasis on composition, however, put an equal aesthetic burden on the cinematographer, and Fritz Arno Wagner and Karl Freund consistently met the challenge by finding new and surprising ways of looking at the artificial universe they captured on film.

German studio system

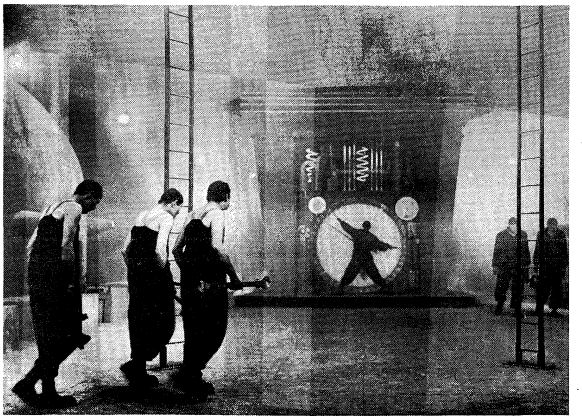
The German studio system was born in the late months of World War I, when the government sought a means of producing propaganda. After the Kaiser's fall, the huge studio Ufa (Universum-Film-Aktiengesellschaft) still stood. Just as there was a great burst of expression and creativity in German painting, poetry, fiction, and the drama after the fall of the monarchy and the birth of the Weimar Republic, creative expression in the German motion picture flourished. The German films thrived on the newfound freedom, but, when this freedom died in 1933 with the rise of Hitler and Fascism, the German film died as well, seemingly forever. The new excitement of the now mature motion-picture art attracted many talented artists to Ufa and its subordinate affiliates: Erich Pommer, a sensitive and intelligent producer; insightful and imaginative directors, including Ernst Lubitsch, F.W. Murnau, Fritz Lang, and G.W. Pabst; and a skillful group of repertory actors, such as Emil Jannings, Conrad Veidt, Asta Nielsen, Werner Krauss, Pola Negri, Lya de Putti, Henny Porten, and Lil Dagover. As early as 1918 Lubitsch had begun to direct his romantic historical spectacles with their clever and human premise that the great public moments of history were the products of intimate private moments. These films were especially distinguished for their careful control of decor and visual texture.

The beginning of the great German era of film making is marked by *The Cabinet of Dr. Caligari* (1919). Directed by Robert Wiene, the script was by Carl Mayer (who would become the greatest of the German scenarists) and Hans Janowitz, and decor was designed by Warm, Reimann, and Röhrig. *Caligari* is a multilevel story. On its first level, it is the story of a maniac, Dr. Caligari, who has hypnotized another human being, Cesare, subjugating the man totally to his own will and power. Though Caligari pretends to use his power merely for human amusement by displaying this human mannequin at fairs and circuses, he really uses his slave as an evil tool, sending him out to murder unsuspecting citizens at night. The film's hero follows Caligari and Cesare and discovers that the mad murderer is ironically the head of a mental hospital. The keeper of the insane is himself insane. The horror story of the film's first level clearly condemns bureaucratic social institutions (the incompetent police and the state asylum), official authorities, official definitions of sanity and madness, and the inhuman control of one man's mind by another.

But the film's second level gives it a completely different twist. Only at the end of the film does the viewer discover that the film's hero and narrator, who has been telling a listener as well as the spectator the whole horrifying story, is himself a patient in an insane asylum. And, further, the insane Caligari, whom he so viciously describes, is indeed the head of that asylum. Is the viewer, therefore, to believe the narrator's story of Caligari's murderousness or to interpret his story as the narrator–hero's paranoid fantasy?

Visual style of *Dr. Caligari*

The entire film has been designed in a visual style that defies objective reality: the walls are too white, the shadows too black; the houses, windows, and doors are not perpendicular with the Earth; there is no sunlight, merely painted black shadows on the walls. There are no streets, for the film seems to take place entirely indoors. The film raises many questions—is one to interpret the bizarre,



Metropolis (1926), directed by Fritz Lang.
By courtesy of the Museum of Modern Art Film Stills Archive, New York

psychotic unreality of the film's visual style as a mirror of the diseased brain of the narrator or as a metaphoric projection of the way an evil mind such as Caligari's casts its spell on reality? Is the film told in the first person or in a metaphoric third person? These two possibilities are as difficult to unsort as is the priority of the film's two plot levels (in fact, the uncertainty of the intention about the film's visual style is a perfect mirror of the uncertainty about its social and intellectual intentions). But, in the film's two levels—one mystical and fantastic, the other internal and psychological—the film points in the two directions of succeeding great German films.

Of the mystical descendants of *Caligari*, there was F.W. Murnau's *Nosferatu* (1922), the first of the film treatments of Bram Stoker's Dracula story, which uses many camera tricks and subjective devices to create a pure figure of horror (rather than the sexually attractive Dracula of later films). There were also Paul Leni's *Waxworks* (1924), a story of three figures in a wax museum who come to life to tell their stories, and the films of Paul Wegener and Henrik Galeen (*The Golem*, 1920, and *The Student of Prague*, 1926), who had begun making films of horror (an earlier version of *The Golem*) seven years before *Caligari*. The films of Fritz Lang, *Destiny* (*Der müde Tod*, 1921), a fantasy of the romantic past, and *Metropolis* (1926), a fantasy of the sterile future, are also in the *Caligari* tradition. *Destiny* is a romantic conflict between love and death (a conflict that also influences both *Caligari* and *Nosferatu*). When Death (personified in the film by a grim, black-clad man) takes away a woman's lover, he gives her three chances to redeem him, each of those chances symbolized by a flickering candle (the candle is used as a metaphor for human existence). Her three chances take her to a Muslim city, to Venice, and to a magical China, all of them distinguished by brilliantly textured sets, costumes, and architectural shapes, the last of them especially delightful in its use of camera tricks descended from those used by Georges Méliès. The woman fails at all three chances, but in the film's epilogue she performs a heroic action, saving an infant from death in a burning hospital. And, in dying herself, she joins her lover forever, conquering Death by dying. Although *Metropolis* claims to be about a world of science, the film is just as romantic. In his condemnation of the class system and the dehumanizing effects of industrial society, Lang builds a film that is visually magnificent but that, in its own way, is equally dehumanizing in its converting of men into dazzling architectural patterns. Lang's melodramatic plot, romantic social abstractions, and naïvely optimistic conclusion further reduce its power as social commentary (theoretically, its purpose) but cannot dull its effects as a kaleidoscopic visual experience.

Murnau's psychological realism

The greatest of the realist descendants of *Caligari* was F.W. Murnau's *Last Laugh* (1924), one of those half dozen or so representatives of the ultimate potential of the silent film as an art. The film's plot is extremely simple. A porter (played by Emil Jannings) at a posh hotel defines his self-respect and the value of his life in terms of this prestigious job, symbolized by his attachment to his ornate porter's uniform. Because he is old, the manager of the hotel strips him of that uniform and gives him another—the white linen jacket of the lavatory attendant. In losing his uniform, the porter loses his self-respect, as well as the respect of his neighbours, his friends, and even his own daughter. The story seems to end with the porter stooped in abject misery, a small, broken figure in the sterile whiteness of the hotel washroom (a man has been symbolically thrown into the toilet). But the film ends with a capricious reversal when Murnau inserts a subtitle, the only one in the entire film, to tell us that he has taken pity on the porter. In the film's epilogue, the porter miraculously inherits the millions of an eccentric American who died in the porter's arms in the hotel washroom. The now wealthy porter enjoys fancy food, cigars, and clothes; he showers the needy with gifts and money. Perhaps the contrived happy ending is needed to snap the absolute futility and despair of the porter's decline. Or perhaps the ending is a deliberate burlesque of the happy ending, also burlesqueing the Hollywood—and therefore American—notion that eventually poetic justice and good fortune will erase all social injustice and personal misery.

The remarkable feature of the film is its detailed care in the emotional rendering of every step in the porter's decay. The expressiveness of Jannings' performance—snappy, alert, vital in his porter's coat; broken, feeble, bent without it—partly creates the film's clarity and power; the expressive camerawork does the rest. Karl Freund, the photographer, frequently leaps into the porter's subjective vision of reality—his dreams of strength, his feeling that the world is collapsing on top of him, his drunken blurriness, his horror when the neighbours discover his dishonour. The camera both records how the porter appears from the outside and how he feels from the inside, constantly translating these feelings with the effects of lenses and the moving camera. Also remarkable about the film is its refusal to sermonize or draw facile moral conclusions. The porter's definition of himself in terms of a uniform is false but understandable. The same is true of his neighbours' false and surface definitions of the man's value. The hotel manager's decision to strip the porter is vicious but pragmatic. Perhaps the most vicious figure in the film is the hotel itself, the social institution that serves the wealthy few by sacrificing the human dignity of the many poor. The symbol of the hotel, a revolving door, underlies what Murnau saw as the inhuman, mechanical social process—as the old porter revolves out, another revolves in, a truly vicious circle.

Among the other psychological, social-realist films of the era was E.A. Dupont's *Variety* (1925), which also starred Jannings. In this story of desire, infidelity, and death among circus artists, the camera again plays a crucial subjective role in mirroring internal emotions; it swings aloft on a trapeze bar, spins around a room, tracks down a hallway.

The "street" films

The "street" films (a group of films that examined a whole social milieu) were more interested in the external problems of poverty, justice, and survival than in internal psychological states. Examples were Karl Grune's *Street* (1923) and Bruno Rahn's *Tragedy of the Street* (1927). These films frequently used the moving camera not for subjective views of experience but as an objective means of revealing a whole social pattern. G.W. Pabst's *Joyless Street* (1925) and *The Love of Jeanne Ney* (1927) are in this same realist tradition, combining melodramatic love stories with a carefully developed political, social, and moral background.

In a completely different direction, several of the German films of the 1920s tried to break the traditions of both fantasy and realism, as well as to break out of the restrictions of studio filming. Early in the decade Hans Richter and Viking Eggeling began making Surrealist, abstract films of form, shape, rhythm, and motion. In 1927 Walter Ruttmann combined these visual–musical principles with the documentary form of film in *Berlin: The Symphony of a Great City*, a candid study of Berlin

and Berliners at work and play, united by rhythms of editing and parallels and contrasts in visual forms. The ultimate intention of this "city symphony" (which started an international trend toward making other such city films) was to show the diversity of the Berlin experience, its joys and sorrows, warmth and coldness, without moralizing. The film also tried to use such kinetic effects of the cinema as rhythmic cutting and a dialogue of shapes and objects to make its points. These German experimental films of the 1920s were clearly a reaction and rebellion against the conventions of German studio films and their definition of nature as something built by a set designer, shaded by an electric light, and shaped by a camera lens.

Soviet Union. The Soviet cinema developed, defined, and ultimately codified the effects of editing—to which they gave the name montage. Like the Berlin symphony (which deliberately emulated the Russians), the Soviet film masters controlled the way that two shots could be joined together to generate meaning and feeling. Unlike the German cinema, the Soviets depended on the outdoors and nature, producing their interpretation of nature not by rebuilding it but by juxtaposing different views of it from shot to shot. In reproducing nature the Soviets also interpreted it, and that interpretation was a Marxist–Leninist view of history, social justice, and human necessities.

The Soviet cinema was created for the explicit purpose of propaganda and education rather than entertainment. Shortly after the Revolution of 1917, the state seized control of the scraps of the tsarist film industry that had been left behind. The Soviet discovery of montage was partly aesthetic—editing proved a very effective means of increasing intellectual awareness, creating social consciousness, and shaping emotional commitment—and partly accidental. Because of the many years of war and revolution, of blockades and counterblockades, there was very little raw film stock in the Soviet Union. So the potential film makers could work only at developing that potential in schools and workshops, doing film exercises rather than making films. Another significant accident is that one of the few Western films to pierce the blockades was Griffith's *Intolerance*, the finest example of the power of motion-picture editing that had yet been made.

Experimental editing in Soviet studios

In the workshop of Lev Kuleshov the future film makers performed countless editing experiments. They discovered that editing alone created the impression of unity of time and place in a film and could even assimilate a whole human being out of the individual parts of many different actors. The act of putting pieces of film together necessarily led the audience to connect the two pieces in time, space, meaning, or feeling. Even a face without expression could take on powerful significance depending on the emotion of the shots that preceded or followed it. For example, a preceding shot of a corpse made such a face appear sorrowful and a shot of a child at play made the same face appear quietly contented.

The experiments revealed that editing served three primary functions in a film. It could narrate, or communicate, the sequence of events, by shifting back and forth between the various figures in the sequence, cutting from a shot of one man to a shot of some force that would soon influence the man (Griffith's crosscut). But editing could also produce intellectual inferences. If the director cuts from a man standing in a pompous position to a statuette of Napoleon, the audience necessarily connects the two intellectually and applies the psychological and political implications of the statue to the man. Still further, editing generates a silent, undetectable energy and rhythm that work on the senses without an audience's being aware of them. A series of quick cuts or languid ones, a series of cuts of steadily darkening pictures or brightening ones, an exact metronomic rhythm of cutting, a series of cuts of larger and larger or smaller and smaller objects—these are all ways that an audience can be manipulated to feel something about the material without even realizing that it is being manipulated. The subliminal (or kinetic) use of cutting is very useful for underscoring a film's emotional effects as well as shaping an audience's moral responses to the action. Visual rhythms, like musical ones, are difficult to articulate because they work on the unconscious.

Eisenstein. Sergey Eisenstein was the greatest magician and musician of montage. He manipulated the kinetic rhythms of editing so successfully that an audience was never aware of manipulation, only of the film's tension, energy, and excitement. Eisenstein's theory of montage was one of collision—a crash of images from cut to cut, the images contrasting in light, shading, length, texture, size, speed, or emphasis. The result of this collision was an emotional experience of intense nervousness, a high level of anxious tension that never relaxed until the film ended. Eisenstein was remarkable for the intellectual thrust of his cutting. He also possessed a keen satiric sense that could produce comic intellectual inferences (juxtaposing a political leader with a shot of horses' buttocks) as well as serious ones. But Eisenstein was equally remarkable in his ability to handle visual images like musical notes and chords, for he was an orchestrator of visual images. His "music" was parallel to the dissonance in the compositions of the contemporary composer Sergey Prokofiev, and the later union of the two "musicians" for the film *Alexander Nevsky* (1938) was no coincidence.

Theme of brotherhood

Eisenstein's essential theme was human brotherhood, destroyed and betrayed by the vicious capitalistic system, eventually revealing itself as the brothers and comrades unite to take political action against the oppressor. Eisenstein consistently makes his protagonist the mass, not a single human being but a whole society. As opposed to the virtuous mass, Eisenstein personifies and individualizes his social villains. He either burlesques his villains (with animal faces as in *Strike* or with satiric, inanimate objects as in *October*) or turns them into the vicious stereotypes of melodrama (the smug officers and satanic priest of *Potemkin*). Eisenstein's social targets, the forces that oppress the people, are the "means of production" (the factory system of *Strike*, 1924), the military (*Potemkin*, 1925), the bureaucracy (*October*, 1928), and Christianity (which Eisenstein treats with either horrifying or comic ugliness in *Potemkin* and *October*).

Potemkin is probably Eisenstein's most unified, consistent, and controlled film. It is a simple story of a group of Russian sailors who seize control of a tsarist ship, build their comradeship and unity with the people of the Russian Black Sea port of Odessa, and eventually succeed in uniting all members of the fleet in brotherhood. Despite the story's simplicity, the flow of its rhythms turns each of its five sections into a musical movement, varying in tone but united in spirit. The sequences can be tensely expectant, chaotically violent, lyrically calm, horrifyingly brutal, or joyously triumphant. The most extraordinary movement in this visual symphony is the famed "Odessa steps" sequence, in which the vicious soldiers of the Tsar slaughter the defenseless men, women, and children, young and old, healthy and crippled, as these people respond joyfully to the harbour excitement of the liberated battleship. The emotional rhythms of the sequence's cutting perfectly mirror the brutality of the slaughter, providing an objective moral comment on the event as well as a subjective impression of how it felt to be there. But Eisenstein's aim is not despair but hope—the new spirit of the people that emerges with the slaughter. Among his subliminal symbols in the film are the battleship's guns, the sexual implications of which are intentional. At the beginning of the film, the guns are sheathed, impotent, lifeless; they are infertile servants of the Tsar. When the people seize the battleship as their own, the guns come to life, an expression of potency and fertility.

Eisenstein's most important rival

Pudovkin. Eisenstein's most important rival, Vsevolod Pudovkin, made films that were quieter, softer, more lyrical. They were also more obvious in their emotional effects and symbolism, using montage as a means of underscoring conventional dramatic effects rather than as a musical or intellectual element in itself. Like Eisenstein, Pudovkin's subject was Soviet political education and conversion, but, unlike Eisenstein, Pudovkin told a story of a single character's conversion, using that singularity to symbolize a mass movement: in *Mother* (1926), a

single woman discovers that revolution is the only means of securing justice in an unjust state; in *The End of St. Petersburg* (1927), a young man learns that to betray the revolution to the authorities in exchange for money is truly the act of a Judas; and in *Storm over Asia* (1928), a Mongol learns to rebel against his imperialist captors and to sweep them from his land. Underlining these gentler, more personal stories of people were Pudovkin's abilities with actors (he was a fine actor himself) and his tender natural imagery of fields, skies, earth, trees, and rivers, which also contrasted with Eisenstein's intellectual, inanimate, sarcastic, or brutal imagery.

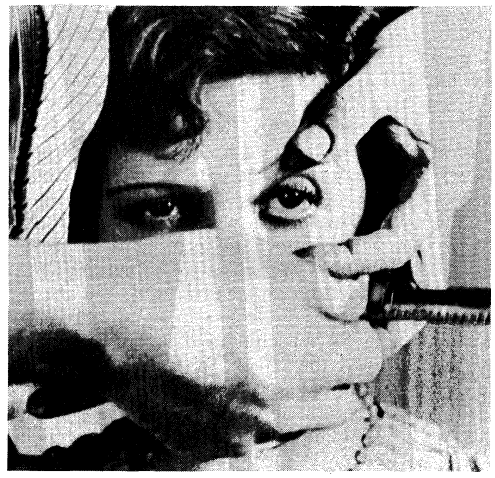
Pudovkin's two greatest contributions to cinema technique were his theory of the "plastic material" and his awareness of camera angle. Pudovkin noted that concrete objects often are more effective at revealing a character's emotion or producing an audience's social and moral awareness than any piece of acting or human gesture. His theory of building a film around specific objects was really the summary of a cinematic axiom that had been unconsciously realized by such directors as Chaplin, Griffith, von Stroheim, and Lubitsch. In an external, surface medium such as the motion picture, which merely records what can be seen, the unique power of the objects the camera sees is undeniable. Pudovkin's camera saw such physical objects as the brutal dismemberment of a fish (*Mother*), the parallel vanity of religious statues and people (*Storm over Asia*), the heavy emotional weight of a single silver coin (*The End of St. Petersburg*). Pudovkin also realized that how the scene or object is seen is equally important. The particular camera position—high, low, eye level—shapes the reaction of both the characters and the audience to the scene. An extremely low angle makes a bureaucrat, soldier, or policeman seem overpowering and awesome. An extremely high angle makes an ordinary citizen seem very small and insignificant.

Other directors. Aleksandr Dovzhenko was the third of the great Soviet directors who realized that natural imagery, montage, human feelings, and political commentary can be mixed together in an exciting, personal way. Dovzhenko was the most poetic of the Soviet masters—the most fanciful, unrealistic, elliptical, and imagistic. In his greatest silent films, *Arsenal* (1929) and *Earth* (1930), there is no story at all, merely a series of isolated vignettes unified thematically by consistent natural imagery and by the flowing rhythms of montage. Among the other significant Soviet film makers in the silent era who similarly combined emotional experience and didactic lessons were Dziga-Vertov, a master of montage documentaries, Abram Room, a more conventional realist (*Bed and Sofa*, 1926), and the documentarists Victor Turin (*Turksib*, 1929) and Esther Shub (who combined montage with old newsreel footage).

Dovzhenko's poetic imagery

France. France in the 1920s was more significant as a place of apprenticeship, experiment, and innovation than as a centre of a mature film industry with a unified aesthetic and theme. The experimentation of the 1920s, however, would produce in France during the 1930s some of the best early sound films and directors (René Clair, Jean Renoir, Jean Cocteau, Luis Buñuel, Jean Vigo) who had begun to master the medium in the silent era. The French experimented in the ways that the motion picture can mirror a distorted, dreamlike, or surreal version of nature. The contemporary artistic movements of Dadaism, Surrealism, Impressionism, and other experimental "isms" produced the films of Man Ray, Jean Epstein, Marcel Duchamp, Fernand Léger, Alberto Cavalcanti, Louis Delluc, and Marcel L'Herbier. Perhaps the most impressive of the Surrealistic short films (the short form was most useful for formalistic innovation since a film of pure forms lacked the unifying element of plot and could easily become tiresome) were René Clair's dizzy romp *Entr'acte* (1924) and the Salvador Dali–Luis Buñuel intermixture of dream, sexuality, and *non sequitur*, *Un Chien Andalou* (*An Andalusian Dog*, 1928).

In the feature film, France produced the spectacular visual experiments of Abel Gance (*J'accuse*, 1919, and *Napoléon*, 1927), a Surrealist interpretation of the U.S. poet Edgar Allen Poe's "Fall of the House of Usher" (1928) by Jean Epstein, and Clair's devastating comic attack on bourgeois assumptions and social conventions, *The Italian Straw Hat* (1927). The most important French feature of the decade was *The Passion of Joan of Arc* (1928), directed by a Dane, Carl Theodor Dreyer, and designed by a German, Hermann Warm. This hypnotic, sensitive sketch of Joan's trial and execution was not so much a story as a visual tone poem of suffering and inspiration. Like Eisenstein's work, the film is a musical flow of images, but, unlike Eisenstein's dissonant symphonies, *Joan of Arc* is a hymnlike choral mass of vital faces and architectural compositions.



Un Chien Andalou (1928), Surrealistic short film by Salvador Dali and Luis Buñuel.
By courtesy of the Museum of Modern Art Film Stills Archive, New York

The feature film in France

Scandinavia. The maturity and control of Dreyer points to the final major film tradition of the silent era. The Scandinavian silent film combined the mystical, romantic view of nature found in the German film with the concrete, objective depiction of vast nature manifested in the U.S. film. The great Scandinavian directors such as Mauritz Stiller and Victor Sjöström developed the interrelation of men and nature, of emotional states and natural forces, and of human urge and natural setting with mystical intuition. Both Stiller and Sjöström made powerful silent films in the United States, where images of nature could also play a significant role in shaping the human story. The introduction of synchronized sound not only erected a linguistic barrier that forced the Scandinavians back to Europe but also temporarily banished the world of nature (where dialogue was difficult to record) from the U.S. film.

IV. The sound era

1928 TO 1945

Experimentation and the first "talkies." The idea of synchronizing sound and picture was as old as the motion picture itself. In the earliest Edison–Dickson experiments, Dickson achieved a rough synchronization of word and image. Many of the inventors of the first cameras and projectors, in fact, devoted their efforts to sound synchronization after this early equipment had been created. There were two technical obstacles to dependable, stable sound filming: synchronization (keeping the image and sound permanently and perfectly together) and amplification (filling a large auditorium with sound that would be audible to many people at once). The amplification problem was solved first, with the invention in 1906 by the American Lee De Forest of the audion tube, which drove the sound into loudspeakers rather than requiring it to trickle out of horns or earphones. The synchronization problem was solved in the years just following World War I by the discovery of the oscilloscope, which was based on the principle of converting sound into light beams that could then be recorded on celluloid alongside the visual image.

Early problems of synchronization and amplification

Although the eventual stability of sound filming would

owe its life to this sound-on-film principle, most early U.S. experiments used the more clumsy sound-on-disc method, attempting to keep film and recorded disc synchronized ("in synch") electronically. As early as 1912, the Edison studio produced a musical film based on nursery tales and characters that roughly synchronized phonograph disc and recorded image. In this very early musical picture, there is no editing at all. The camera statically records 15 minutes of singing and dancing without stopping or moving. This stasis would be a consistent trait of sound films until 1930. In the 1920s De Forest made a series of Phonofilms, filmed recordings of singers, entertainers, and lecturers, exactly parallel to the first novelty performances in Edison's movie studio. De Forest's Phonofilms used a sound-on-film principle. So did the Movietone, another sound process that recorded static performances of talking and singing. William Fox bought the Movietone in 1927 and inaugurated a series of conversations with famous people (George Bernard Shaw, for example). In 1927 Fox also produced the first newsreel that combined narration and picture—the Fox–Movietone News.

The first and most influential triumph of synchronized sound used the sound-on-disc process: the Western Electric Company's Vitaphone. After failing to interest any of the major Hollywood producers in the Vitaphone, Western Electric sold the device to Warner Brothers, who were struggling for survival in 1925. Warner Brothers first used the novelty exactly as their competitors did, by making sound-film recordings of opera stars, famous personalities, and vaudeville entertainers. In 1926 Warners added a recorded ("canned") musical score and sound effects to feature films, replacing the live orchestra or piano in the pit. In 1927, the Warners added four singing sequences and a little dialogue to an otherwise silent feature film with recorded musical accompaniment—*The Jazz Singer*. Contrary to myth, this Al Jolson vehicle was not the first talking picture and not the first film to use synchronized sound. But it was the first commercial feature film to combine the musical performances and "canned" orchestrations that had been used earlier with some dialogue that replaced title cards and helped to move the film's plot. Sound had been integrated, if only fleetingly, with plot and not used as an independent novelty in itself. *The Jazz Singer* was such an enormous commercial success that every studio had converted to sound within two years. And the Vitaphone quickly converted itself to the sound-on-film principle.

Commercial success of *The Jazz Singer*

The initial effect of Hollywood's conversion to sound filming was commercial, technical, and aesthetic chaos. Sound filming was enormously expensive. The new sound cameras and microphones were not the sole expense, for sound filming required completely soundproof studios, where only a single shot could be filmed at a time. In the silent era, six production companies could share a single, huge room, shooting noisily but easily side-by-side. The theatres also required expensive sound projectors and speaker systems, as well as wiring to link them. In addition to these commercial difficulties, many of the studios' old stars were suddenly worthless, their new voices no longer matching their old images, and the studios imported new stars, writers, directors, and plays from Broadway.

The technical difficulties of the ponderous sound machines imprisoned the camera in a soundproof glass booth, to keep its noisy apparatus away from the microphone. The active, mobile fluidity of the camera in the silent era came to a very abrupt end. The rhythmic freedom of the editing process also suffered the restrictions of sound, for it was very difficult to edit sound and picture without knocking the dialogue out of synchronization. Space and time reasserted their tyrannical control over the motion picture. Even the actors lost their ability to move, for dialogue scenes required them to sit, stand, or huddle together in a stationary position so that the single hidden microphone could catch their words.

The earliest sound films used static, dead pictures as accompaniment for the dialogue. There was no longer motion in the motion picture. In *The Jazz Singer*, for example, the sound sequences use long, frozen views of Al Jolson's singing with no camera movement and very little cutting. Whatever happens in these sequences is because of Jolson, not the camera. The silent sequences of the film, however, though fluid and active, are less entertaining, and that was the point for the audience. Despite the unimaginative stasis of the cinematic technique, the sound sequences provided a kind of entertainment that they had not seen before—the speech and song of a compelling personality combined with a complete motion-picture story. The problem for imaginative film artists was to integrate that kind of entertainment with the earlier principles of "motion" pictures. Film critics and theorists were certain that such integration was impossible, that Hollywood's commercial greed had swallowed a magnificent art, and that the motion picture was dead as an art form.

Role of pictures in early sound films

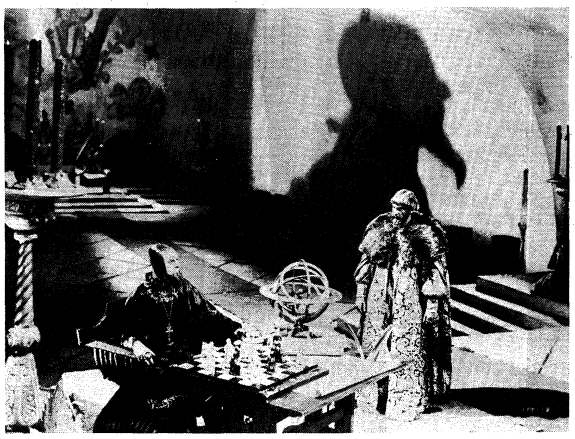
Ironically, that particular form of art was indeed dead. The introduction of sound did not simply add something to silent pictures; it also took something away—the way the juxtaposition of images and the texture of visual compositions themselves created social reality, human feelings, and thought. That reality, those feelings, and those thoughts would now have to communicate themselves, at least partially, in the same way they do in reality—with speech. The difference between the silent and sound film was the difference between a possible ballet version of *Hamlet* and Shakespeare's version. The aesthetics of the sound film and the silent film were as different from one another as the aesthetics of the theatre and the silent film or the theatre and the sound film. Once technicians had invented more serviceable sound equipment, the artists could discover the aesthetics of this new form—the way of integrating picture, word, cutting, noise, and music.

Some of the first interesting sound films were made in the United States, where sound produced its first complete revolution. Directors quickly learned that, although talking scenes had to be static, others without talk could be shot silently, using the old freedom of camera movement and cutting, then dubbing (adding) in the music and sound effects later. Ernst Lubitsch's *Love Parade* (1929) and *Monte Carlo* (1930), Rouben Mamoulian's *Applause* (1929), King Vidor's *Hallelujah!* (1929), and Lewis Milestone's *All Quiet on the Western Front* (1930) all reverted to silent principles for their active, nontalking, and musical sequences. Lubitsch was especially inventive at realizing that what is heard but not seen or seen but not heard could play very clever, ironic contrapuntal roles in a film. He also realized that, if one was going to make talking films, the dialogue ought to be good, and in his films (most of them written by Samson Raphaelson) the talk was very good indeed.

While directors of feature films tried to integrate the clumsiness of sound filming with their mirroring of the natural world, the animated cartoons of Walt Disney could indulge in all kinds of playful games, owing no allegiance to reality or nature. Disney used sound so cleverly in his second Mickey Mouse film, *Steamboat Willie* (1928), that he vaulted himself, Mickey, and the clever potentialities of sound to international attention. In the film Disney delighted in showing the surprising, ingenious sounds that could issue from cats, cows, pigs, and parrots. The climactic "Turkey in the Straw" number, in particular, is a visual–aural symphony of movement, rhythmic cutting, sound effects, and music. Symphony is an appropriate word for the effect, and one year later Disney inaugurated his series of "Silly Symphonies," which continued his experiments with abstraction, musical tone, rhythm, animated motion, and, by 1933, colour.

Early animation with sound

The early German sound films were among the best made anywhere. The German reliance on studio production and compositional style in the silents was especially suited to sound filming, which similarly required studio filming and relied on compositional style and texture for its primary visual appeal. *The Blue Angel* (1930), despite its U.S. director Josef von Sternberg, is a very Germanic study of a man's collapse (the man played by Emil Jannings). The careful control of visual contrasts, symbolic objects, musical motifs, and contrapuntal sound in the



Ivan the Terrible (filmed 1944–46), directed by Sergey Eisenstein.

By courtesy of Janus Films, Inc.; photograph, from the Museum of Modern Art Film Stills Archive, New York

film created a motion picture that seemed as cinematically free and emotionally satisfying as any of the German silent era. Fritz Lang's *M* (1931) was successful primarily because of the director's awareness of the power of subjective sound (and silence) as a means of tightening suspense. G.W. Pabst's early sound films, *Westfront 1918* (1930), a pacifist view of World War I, and *The Threepenny Opera* (1931), a heavy-handed but visually stunning version of Bertolt Brecht's play, also used the devices of sound—music, silence, asynchronous effects (sound and picture working in counterpoint, not together), subjectivity (the way the sounds feel to the character, not the way they really sound)—very effectively. When the Nazis gained control of the government and seized the film industry, the great German directors left the country, as did the spirit of creative freedom in German film making. After 1933 German films became remarkable instruments of propaganda, in which the "documentaries" of Leni Riefenstahl (*Triumph of the Will*, 1936; *Olympia*, 1936–38) and others were used as a means of allying the German mind with Hitler's objectives.

The early Soviet sound films were uninspired and clumsy. The great minds of the silent era had terrible troubles with the new equipment. Those troubles stemmed from several causes. Reliable sound machines were very slow arriving in the Soviet Union. The greatest silent film makers had all become politically suspect. And perhaps the very essence of the Soviet method, montage, was contrary to the demands of the sound film. Although both Eisenstein and Pudovkin wrote very perceptively on the theory of the sound film (they both argued that sound should work asynchronously with the visual), both deplored the idea of talking films—and most film critics feel that neither made sound films that measured up to his greatest silent work. Eisenstein did not complete a sound film until 1938, and his sound pictures (*Alexander Nevsky* and the two completed parts of *Ivan the Terrible*, made 1944–46) are more successful in their parts than in their entirety, and in their visual and intellectual conception rather than their execution.

France. The spirit of formal experiment and innovation among the French film makers in the 1920s led to their eager acceptance of the formal challenges of the sound film. Sensitive French directors quickly perceived the principle of asynchronous effects. In *Sous les toits de Paris* (1930; "Under the Roofs of Paris"), René Clair used a unifying street singer's song and cleverly integrated such sound effects as a repeating groove of a phonograph record into his story. In *Le Million* (1931), Clair juxtaposed a battle for a piece of clothing with the noises of a football game on the sound track. In *Boudu sauvé des eaux* (1932; "Boudu Saved from Drowning"), Jean Renoir used a shot of an etching—a military bugler blowing his trumpet—juxtaposed with the notes of a trumpet on the sound track to imply a seduction. In the same film, significant and symbolic musical motifs (the "Blue Danube" waltz, a brass band, a flute, an organ grinder's tune, a grunting song) underscore the film's action and ideas. Jean Vigo's *Zéro de conduite* (1933; "Zero for Conduct") and Jean Cocteau's *Sang d'un poet* (1931; "The Blood of a Poet") were predominantly silent films with musical accompaniment, functional sound effects, and a very sparing use of dialogue.

French use of sound

Having mastered the principles and procedures of sound filming so quickly and cleverly, the greatest French directors made some of the best films of the decade—indeed of any decade.

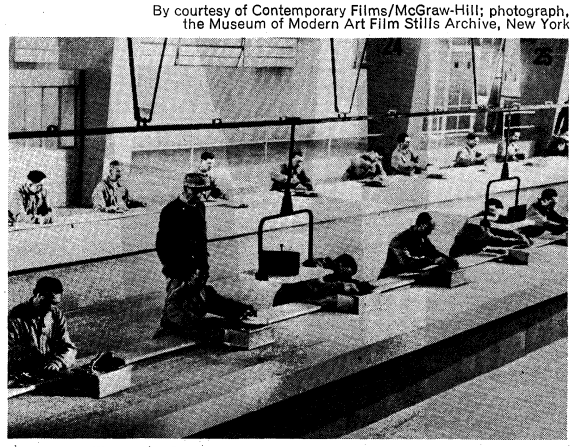
René Clair. Between 1930 and 1933 René Clair's career reached its peak with four sound films—two lyrical melodramas (*Sous les toits de Paris* and *Quatorze Juillet*) and two musical farces (*Le Million* and *À nous la liberté*). Clair's two greatest gifts were his feeling for music and his sense of decor. Songs, dance, musical interludes, and balletic chases dominate Clair's films, in which there was more singing than talking. In addition to evading reality by depicting a world where people sing, Clair handled the decor in an equally fanciful way, using visual shapes, patterns, and textures to echo his plot and themes rather than to mirror nature. In *À nous la liberté*, for example, a film that fancifully equates work (a factory) with slavery (a prison), every composition echoes the monotony and sterility of prisons by arranging people and sets in perfectly regular horizontal and vertical lines. Clair avoided the stifling demands of realistic talking pictures by avoiding reality and creating a world of fancy.

Injection of serious themes

Into this operetta world Clair injected serious themes. Central to most of his films is a male couple, who symbolize friendship or brotherhood in the abstract (*Sous les toits, Le Million, Liberté*). The plot of the film threatens to tear the friends apart, but they emerge victorious and friendly at the end. The foes of these friends become Clair's philosophic enemies: money, greed, possessions, social class, social propriety, societal assumptions about a meaningful and valuable life. Clair, heavily influenced by Chaplin, is the advocate of nature in its war with society, of people in their war with machines, and of fancy in its war with reason (hence, the unity of style and content in these fanciful films). Clair's ideal of happiness concludes *À nous la liberté* when the oppressive, prison-like factory continues to produce machines by itself, leaving human beings free to dance, sit by the river, and relax in the sun. Clair's sunny, optimistic view of man's innate potential was perfectly suited to the delicate, fanciful whimsy of his earliest sound films. But, when Clair's cleverness got heavy-handed (*The Last Millionaire*, 1934; *Beauty and the Devil*, 1949; *Beauties of the Night*, 1952) or when he tried to inject fantasy into a natural dialogue reality (*The Ghost Goes West*, 1935; *It Happened Tomorrow*, 1943; *Le Silence est d'or*, 1947), his delicate, uncomplicated ideas drowned in a sea of formalistic whipped cream.

Jean Renoir. Like Clair, Jean Renoir was also a proponent of nature, spontaneity, and freedom and the opponent of social form, convention, and assumption. Renoir's view of the conflict, however, was darker and more complex than Clair's, for he did not distill society into a

By courtesy of Contemporary Films/McGraw-Hill; photograph, the Museum of Modern Art Film Stills Archive, New York



A nous la liberté (1931), directed by René Clair.

Boudu sauvé des eaux (1932), directed by Jean Renoir.
By courtesy of Pathe Contemporary Films, Inc., photograph, the Museum of Modern Art Film Stills Archive, New York

vague, impersonal "it" but showed quite clearly that malignant social forces were the products of men. His comedy is more bitter and caustic than Clair's and his serious films more gloomy and menacing. Whereas Clair denuded reality of its natural texture, Renoir worked oppositely, to provide as much realistic texture and depth as possible, even developing a unique shot that relied on depth of field to produce visual and emotional contrast between foreground and background.

Renoir was also a master of complex cinema structure, developing parallel lines of action, parallel events, and parallel symbols in each film that seem to speak to one another within a rich intellectual system. The Renoir films also rely more heavily on individual performances, using some of the most impressive and able film actors of the period: Jean Gabin, Erich von Stroheim, Michel Simon, Marcel Dalio. And, finally, Renoir was a master of a cold, distant irony that consistently showed human beings struggling blindly toward worthless goals, pitifully unaware of their own blindness and the worthlessness of their pursuits. In *Boudu sauvé des eaux*, for example, a bourgeois bookseller rescues a despairing tramp from the river and seeks to continue the rescue process by introducing the man of nature to the world of culture (a bourgeois home and edifying books). By the end of the film, the former tramp is so engulfed by respectable, bourgeois culture that he has a wife, fancy clothes, money (won in the lottery), a mistress (except that Boudu's mistress is, ironically, the bookseller's wife), and a job (in the bookstore). Boudu decides to turn his back on this life and "unrescue" himself. He pretends to drown in a boating accident, but he actually returns to the open road as a bum once more. The careful structural pattern comes full circle as Boudu ironically saves himself as a human being by pretending to drown.

Renoir's mastery of contrast

Renoir's mastery of contrast is evident in such films as *Toni* (1934), in which he contrasts the sensual, anarchic passions of Spanish peasants with civilized society's laws and regulations; in *The Crime of Monsieur Lange* (1936) the utopian happiness of a communal publishing house is contrasted with its previous stifling sterility and inhumanity under a tyrannical owner; *The Lower Depths* (1937) contrasts the suffocating life within a social microcosm with the freeness of life in the outdoors; and *A Day in the Country* (made in 1936) contrasts the artificiality of bourgeois city people with the natural feeling of the people who live outdoors in the country. *Grand Illusion* (1937) and *The Rules of the Game* (1939) climax the decade, indeed the entire prewar French cinema, with their studies of polite societal murder—the chivalric murder of World War I, in which everyone (prisoners and captors alike) is caught in the rules of wars and concentration camps, and the genteel murder of spontaneous emotions and natural vitality by the rules, codes, and assumptions of the ruling classes. The two films depict the death of the old aristocracy, blindly, ironically killing itself and everyone around it with its artificial, inhuman definitions of human value and social conduct.

Marcel Carné. Death is also pervasive in the films of Marcel Carné (with scripts by Jacques Prévert), not because of deadly social assumptions but simply because men die and they are not happy. The nihilism of these films takes several forms: a tragedy of fate in which death overtakes the man fleeing from it and in which love cannot triumph over death (*Port of Shadows,* 1938); a farce in which men make mistakes of interpretation because they are selfish and blind and in which even the comedy revolves about murder and death (*Bizarre, Bizarre,* 1937); a spectacular historical extravaganza in which the theatre is more satisfying than life, in which people achieve great success at what they do not really want and which does not satisfy them, and in which love and happiness always dangle just beyond man's grasp (*The Children of Paradise,* 1943–45).

To make their journeys into blackness entertaining and moving, Carné and Prévert built extremely complex film structures (like Renoir), rich in ironies, levels of interpretation, and ambiguities. They captured stunning visual images, saturated in the style and period of the particular story, and manipulated a heightened poetic dialogue with great effectiveness. Carné also used gifted actors such as Michel Simon, Jean Gabin, Jean-Louis Barrault, and Pierre Brasseur. The complex richness of a film such as *The Children of Paradise* or *The Rules of the Game* reveals how far the sound film had come from its clumsy introduction only a decade earlier.

United States. The sound film came a long way in the U.S. as well, but, unlike the poetic, personal films from France, the U.S. film was still made in the context of an industrialized studio system. The basis of the studio system was quantity, the need to produce over 500 feature films each year for the over 75,000,000 weekly customers who patronized the chains of theatres that the studio corporation itself owned. The Hollywood studio was, in effect, an entertainment factory. Like any factory, its goal was to produce a large number of products of consistent, dependable quality. Also, like any factory, its organizing principle was the division of labour—separate departments for writers, performers, technicians, film cutters, publicity men, etc., each of which contributed its piece to the finished whole. A film was not an individual artistic and commercial enterprise but one part of a whole year's output. Uniting all the studio departments were the men in the management who planned the year's production, managed all the budgets, and kept the assembly line flowing. The introduction of synchronized sound—with its mechanical difficulties, technical intricacies, and expense—gave even more power to the industrial system that had already existed for a decade. From 1930 to 1945, the studio system reigned supreme in the United States.

The commercialism of the studio system

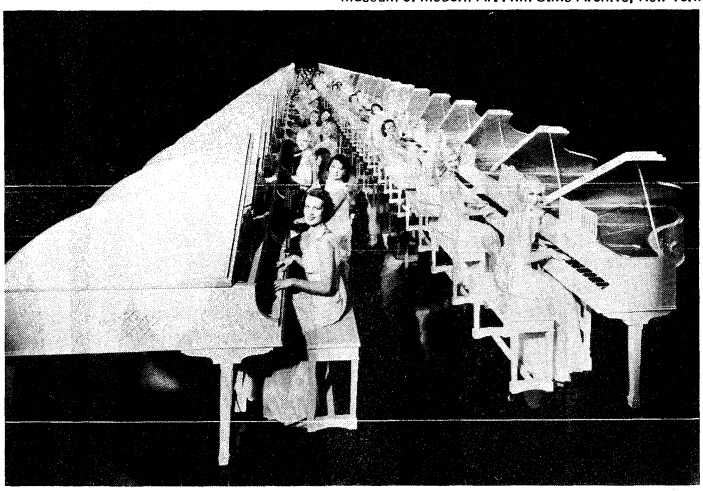
That system created a very clear tension between art and commerce. Art defies mass production and assembly lines. Indeed, most films of the studio era were competent, conventional pieces of unimaginative entertainment, far more interesting sociologically (in defining the attitudes and expectations of the audience) than artistically. Most Hollywood films substituted craft for inspiration and familiar genres for individual conceptions. The director who could execute a script with professional polish, on schedule and within the budget, and the script that fit into some familiar film cycle—the gangster film (now that audiences could listen to tough talking), the musical (a genre that sound made possible), the western, biography, marital comedy, historical romance—were the studios' most valuable commercial assets. The dominant cinematic style in the studios of the 1930s was clean, bright lighting, along with smooth, functional editing and unobtrusive camera work that did not detract from the dialogue, for most studio films were "talkies."

The studios. Within the general pattern, the two giant studios, M-G-M (Metro-Goldwyn-Mayer) and Paramount, were stylistic and intellectual opposites. M-G-M was a studio with almost dictatorial central control by Louis B. Mayer and Irving Thalberg. Paramount under

Adolph Zukor was a much looser organization, allowing individual producers and directors greater freedom. M-G-M was a studio of stars (*e.g.*, Greta Garbo, Jean Harlow, Joan Crawford, Norma Shearer, Judy Garland, Clark Gable, Lionel Barrymore, Mickey Rooney, Spencer Tracy), while Paramount was a studio of directors and writers (*e.g.*, Ernst Lubitsch, Josef von Sternberg, Cecil B. deMille, Leo McCarey, Preston Sturges, Billy Wilder, Samson Raphaelson). In the early 1930s Paramount was the iconoclast's haven—the home of Mae West, who ripped up the sentimental clichés of femininity and romance; of W.C. Fields, who ripped up the sentimental clichés of propriety and conventional ethics; and of the Marx Brothers, who ripped up high finance, higher education, democratic government, and anything else they could get their hands on. Paramount permitted the sexual excesses of von Sternberg and the sexual snickering of Lubitsch. M-G-M, on the other hand, flattened all excesses into a shiny, respectable wholesomeness. When M-G-M inherited the Marx Brothers, it smoothed their chaotic clowning with elaborate production numbers and trite plots. Lubitsch films for M-G-M were heavier, shinier, and sweeter than his breezy films for Paramount. M-G-M gladly sacrificed vitality and energy for its glossy, pseudo-elegant style.

None of the other studios was as distinctive as these two at the extremes. Warner Brothers was remarkable for its gangster pictures, biographies, and musicals. It had directors such as William Wellman, Mervyn LeRoy, William Dieterle, Busby Berkeley, and Michael Curtiz and stars

© 1935 First National Pictures, Inc., Renewed by United Artists Television, Inc.; photograph, from the Museum of Modern Art Film Stills Archive, New York



Elaborate production number from the musical film *The Gold Diggers of 1935*, directed and choreographed by Busby Berkeley.

Warner Brothers, 20th-Century Fox, and RKO

such as James Cagney, Edward G. Robinson, Paul Muni, Humphrey Bogart, Dick Powell, Ruby Keeler, and Bette Davis. The 20th-Century Fox Film Corporation excelled in historical and adventure films directed by John Ford, Henry King, and Henry Hathaway with Tyrone Power and Henry Fonda as stars; in show-business musicals with Don Ameche, Alice Faye, Betty Grable, and Dan Dailey; and in the comedies of Shirley Temple. RKO is most memorable for its musicals with Fred Astaire, smooth comedies with Cary Grant, and the adventure films and comedies directed by Howard Hawks. Universal excelled in the horror films—Dracula, Frankenstein, and the Werewolf—directed by James Whale and Tod Browning; in the comedies of W.C. Fields after he left Paramount; in the cheap comic films of Abbott and Costello, Ma and Pa Kettle, and Francis, the Talking Mule; and in the saccharine singing pictures of Deanna Durbin and Gloria Jean.

Most directors in the studio era were merely staff directors. These competent, proficient, unimaginative technicians took any script they received, shot it, and then passed the footage along to the editing department for shaping into its final form. The few directors who proved themselves particularly capable, who earned the right to shape the script before shooting and control the cutting after it, either transcended the Hollywood clichés or injected so much life into the cliché it was no longer recognizable as such.

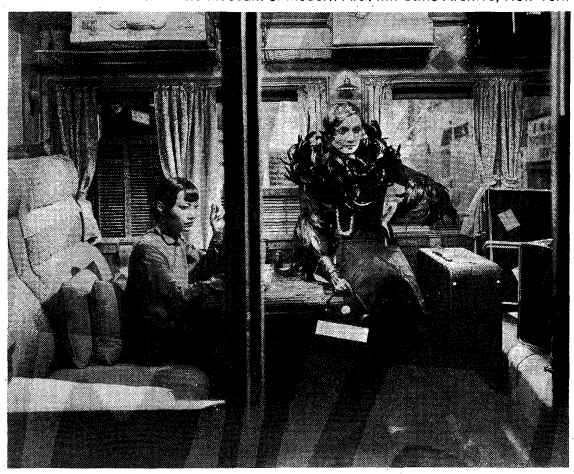
Ernst Lubitsch. Ernst Lubitsch was Hollywood's most remarkable director at avoiding formulas and even at playing games with the formulas. His central subject was sex, quite a central subject in an era that produced a formal moral code in 1933 to eliminate sex from the motion picture. Lubitsch showed that even faithful husbands have their moments of rakishness and that women are not statues but sensual beings with powerful sexual drives of their own. In matters of sex, Lubitsch was a gentle European cynic, rather than an American idealist like most of his contemporaries. The primary battle in a Lubitsch film was between sexual attraction and the permanent definitions of a person's life. In *Forbidden Paradise* (1924), *The Student Prince* (1928), *The Love Parade* (1929), *Monte Carlo* (1930), *The Merry Widow* (1934), and *Cluny Brown* (1946), the characters felt their social positions pull them one way and their desires pull them another. In *The Marriage Circle* (1924), *Lady Windermere's Fan* (1925), *So This Is Paris* (1926), *One Hour with You* (1932), and *Angel* (1937), the characters felt their marriages (another kind of social position) pull them one way and their desires another. In *Ninotchka* (1939) and *To Be or Not to Be* (1942), political beliefs pull one way, passion another; and in *Trouble in Paradise* (1932) vocation (thievery) pulls one way and desire another.

Lubitsch's handling of potentially censorable material

Lubitsch handled the potentially dangerous or censorable material in such ironic and clever ways as to take all the passion out of it and turn sex into a matter of intellect. In his silent films Lubitsch revealed sexual attraction by concentrating on inanimate objects. In his sound pictures, however, he added another dimension to his sexual comedy by making the picture and soundtrack work in opposition. In *The Love Parade* Maurice Chevalier can tell a risqué anecdote—if the camera stays outside the window and reveals him talking about a lewd incident that one cannot hear but must imagine. In *The Merry Widow* Maurice Chevalier plays erotic games with Jeannette MacDonald's legs underneath a café table. Lubitsch, however, keeps his camera on the faces above the table. Only Jeannette MacDonald's "Stop that" and "Don't do that" inform one of what cannot be seen but must be imagined. According to Hollywood clichés, sex was either an all-consuming flame or a nonexistent feeling for the pure of heart. For Lubitsch, sex was merely one kind of natural human activity—no more or less important than intelligence, cleverness, manners, wit, taste, humour, and style.

Josef von Sternberg. Josef von Sternberg was the director who most successfully and powerfully depicted the reigning clichés. Von Sternberg's sound films—*The Blue Angel* (1930), *Morocco* (1930), *Shanghai Express*

Courtesy of Universal Pictures, © Universal Pictures; photograph, from the Museum of Modern Art Film Stills Archive, New York



Marlene Dietrich in a scene from Josef von Sternberg's *Shanghai Express* (1932).

(1932), *Blonde Venus* (1932), *The Scarlet Empress* (1934), *The Devil Is a Woman* (1935)—were stories of men and women bewitched by overwhelming desire. At the centre of these films was Marlene Dietrich, the supreme enchantress, who turned men into milk. Whereas Lubitsch made sexual relationships playful, natural, and ironic, von Sternberg made them terribly exotic, relying on strange foreign settings, esoteric symbolism, and a lush visual style with grotesque contrasts of light and dark, minutely detailed ornate settings, and a breathlessly moving camera (usually managed by Lee Garmes). Perhaps von Sternberg himself saw the excesses of the grand passions in his films; his later Dietrich films are so excessive in their visual styles and symbolism that the director may have been chuckling slyly to himself about the naïveté of an audience that could accept such bizarre posing as emotional reality.

Howard Hawks. Howard Hawks not only eliminated the appearance of sex from his films; he eliminated most of the Hollywood emotions. Hawks's characters are remarkable for their doing, not feeling. His films refuse to moralize and sentimentalize, a striking characteristic in a moralistic, sentimental era. The Hawks character is wrapped up in himself—his job, his interests, and his personal sense of integrity. The conflict in a Hawks film stems from the integrity of one character running up against the integrity of another: in *Red River* (1948) the older ego of John Wayne confronts the younger ego of Montgomery Clift. The male ego of Humphrey Bogart meets the female ego of Lauren Bacall in *The Big Sleep* (1946), and in *Bringing Up Baby* (1938) the scientist's ego of Cary Grant runs up against the feline ego of Katharine Hepburn. *Twentieth Century* (1934) is the most complete and successful comedy of ego ever made, in which two people of the theatre, John Barrymore and Carole Lombard—the Svengali and his ungrateful Trilby—fight to an egotistic draw.

Hawks's characters fight more than they love, but in that fight there is a kind of love; for, if they did not love, they would not fight. Love in a Hawks film is a matter of mind—one ego's perception of the strength and integrity of another. When Hawks added this ironic and insightful view of human interactions to his fast-paced rhythmic action and his sharp cutting and functional camera style, he proved himself one of the best storytellers of the studio years.

John Ford. Like Hawks, John Ford was interested in the civilizing of the old West. But, whereas Hawks's stories were of strong-willed men imposing their wills on weaker ones, Ford's were moral allegories of virtuous men bringing moral light to spiritual darkness. His villains were either uncivilized savages who had not yet received the light of morality or, worse, men of civilization who deliberately warped and perverted law and order for their own selfish purposes (the banker in *Stagecoach,* 1939; the murderous brothers in *My Darling Clementine,* 1946; or Liberty in *The Man Who Shot Liberty Valance,* 1962). Ford's ultimate belief, however, is in human kindness. In *Stagecoach* the kind people inside that social microcosm are also the virtuous people, regardless of their specific social and moral errors. Ford turns John Steinbeck's political novel *The Grapes of Wrath* (1940) into a personal epic of one family fighting vague social and natural enemies rather than Steinbeck's specific economic system. The Joad family survives because of its human strength in combatting the outside forces and because, despite the struggle, they never lose their human compassion and feeling for one another and for others in a similar plight. Ford's films are not nearly as tough and fast-paced as Hawks's, but they compensate with their genuine feeling for human compassion, suffering, and sacrifice.

Ford's moral allegories of virtuous men

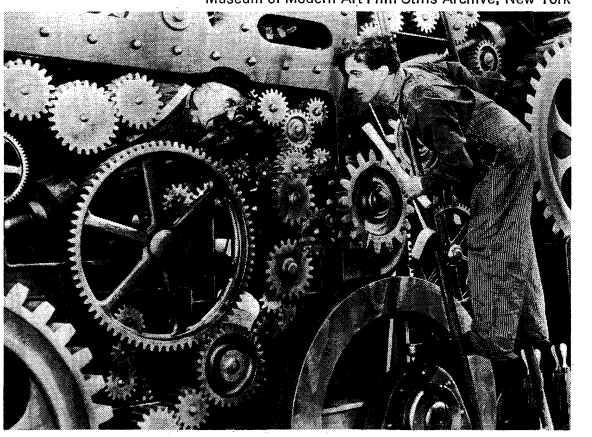
Frank Capra. Frank Capra was John Ford's counterpart in comedy. For Capra, as for Ford, the centre of moral and political virtue was the heart, whereas for both Hawks and Lubitsch it was the brain. If people all loved each other, acted warmly toward each other, there would be no wars, no poverty, no tyranny, no dishonesty, and no unhappiness. The Capra hero is personified by Longfellow Deeds (*Mr. Deeds Goes to Town,* 1936), Jefferson Smith (*Mr. Smith Goes to Washington,* 1939), and John Doe (*Meet John Doe,* 1941), three unaffected, simple, folksy Americans who solve all moral problems with their personal warmth and who convert all the hardened cynics with their sincerity. The Capra villain (often acted by Edward Arnold) is personified by someone who uses his wealth to buy power and reduce other human beings to his pawns and puppets (*Mr. Smith*; *You Can't Take It With You,* 1938; and *John Doe*).

To reduce conflicting social and political ideas to matters of sentiment was absurdly naïve. But Frank Capra's view of politics harmonized with the hopes of his audience, who felt the Depression at their backs and World War II staring them in the face. Despite their political naïveté, the Capra films are very entertaining, particularly those with the least serious scripts and the most of James Stewart's or Gary Cooper's bumbling naturalness and Jean Arthur's or Barbara Stanwyck's energetic wit.

Other studio directors. Among the other distinctive studio directors were Leo McCarey (*Ruggles of Red Gap,* 1935; *The Awful Truth,* 1937; *Going My Way,* 1944), who made Capra-style comedies of sentiment that were less energetic and more plodding than Capra's; Preston Sturges (*The Great McGinty,* 1940; *The Lady Eve,* 1941; *Hail the Conquering Hero,* 1944), who combined sharp dialogue, Lubitsch wit, a cynical view of Hollywood clichés, and Capra's love for the "little people"; George Cukor (*Dinner at Eight,* 1933; *The Philadelphia Story,* 1940; *Adam's Rib,* 1949), who directed stylish dialogue comedies in the Lubitsch manner without the Lubitsch insight and attitude; and William Wellman (*Nothing Sacred,* 1937; *The Ox-Bow Incident,* 1943), who was a softer, sweeter version of Hawks.

Chaplin. Still other directors, particularly Walt Disney and Chaplin, were free of studio restrictions altogether, maintaining their independence and innovation throughout the decade. Chaplin made no conversion to sound at all in his first two "talkies" (*City Lights,* 1931; and *Modern Times,* 1936). He preserved his tramp in a world of mime and silence, underscored by music that Chaplin himself composed. The two films are deliberate burlesques of synchronized sound, using clever sound effects and a very rare dialogue sequence to show how freely one could handle sound in a film. For his final three U.S. films Chaplin adopted dialogue but still maintained much of his spontaneity, wit, and ironic iconoclasm: *The Great Dictator* (1940), which contrasted the egomaniacal urges of a Hitler with the natural, human responses of a little tramp-barber (Chaplin played both roles); *Monsieur Verdoux* (1947), which acidly and ironically showed murder to be as legitimate a means of making money as any; and *Limelight* (1952), a tender, personal summary of Chaplin's whole career and commitment to life and art.

© Roy Export Company Establishment; photograph, the Museum of Modern Art Film Stills Archive, New York



Charlie Chaplin in *Modern Times* (1936).

Late Chaplin films

Orson Welles. Orson Welles was a film maker imported to Hollywood from outside the studio system. Unlike Hitchcock, Welles was never successful in Hollywood's

Orson Welles in *Citizen Kane* (1941).
By courtesy of RKO Radio Pictures, a division of RKO General; photograph, from the Museum of Modern Art Film Stills Archive, New York

terms. Like the earlier perfectionist and iconoclast Erich von Stroheim, Orson Welles lost more and more control over his films and eventually lost his job altogether. But *Citizen Kane* (1941), Welles's first and greatest film, ironically signified the death of the old Hollywood system even as Hollywood was trying to suppress it. The film aroused public controversy and contempt for its obvious and unflattering relationship to the life of the publisher William Randolph Hearst. Further, the film's structure was sprawlingly complex (yet completely controlled), its visual style idiosyncratic (shots using extreme depth of field, extreme shadowy backlighting, extreme upward angles), and its critical view of U.S. life and values was out of keeping with a nation engaged in a world crisis that threatened those values. For *Citizen Kane* showed the horrible destruction, emptiness, and deadness of a man who was supremely successful in the American terms of success.

Welles's influence

Charles Foster Kane was a man of fabulous wealth and possessions, but the one thing that he did not possess was love; for love, not being a material object, is not capable of possession. *Citizen Kane*, therefore, is a tragic story of the conflict between the material and the spiritual, possession and love. In the comic or romantic treatment of the American Dream (for example, Harold Lloyd's or Douglas Fairbanks'), the hero acquires both material success and love, the reward and the girl. *Citizen Kane*, however, shows that the two are mutually exclusive. If one devotes oneself to objects, one will come to treat people as objects and then will end up with nothing but objects. Ironically, Charles Foster Kane was dead before he died; his death began with the childhood decision to separate him from his family and natural environment all because of the rich metal (another object) on the family's property. The journey of Charles Foster Kane from vibrant, rebellious youth to dead tyrant and ogre was inevitable, the result of never knowing what was really valuable. In the individuality of its bitter theme, as well as the daring, innovative individuality of its cinematic technique, Welles pointed the way for a later generation of U.S. film makers. Although Welles's later work shows an occasional flash of *Kane*'s brilliance (*The Magnificent Ambersons*, 1942; *The Lady from Shanghai*, 1948; *Othello*, 1955; *Touch of Evil*, 1958; *Falstaff*, 1967), it never reaches that perfect integration of complex human statement, complex intellectual structure, and complexly rich cinematic style.

Documentaries. Also completely outside the studio boundaries and conventions was the U.S.'s struggling yet growing documentary movement. The most sensitive and effective U.S. director of documentaries in the 1930s was Pare Lorentz, whose *Plow That Broke the Plains* (1936), a study of the Depression's effects on the poor farmer in the "Dust Bowl," and *The River* (1937), a study of the beauty of the Mississippi River and its impact on the land and people surrounding it, combined factual photography, social criticism and commitment, compassionate human observation, poetic narration, effective musical scores, and a masterful control of editing. Lorentz was equally important as a producer and as an influence on the future direction of the U.S. documentary movement. He produced *Power and the Land* (1940), directed by the accomplished Dutch documentarist Joris Ivens, and brought Flaherty back to the U.S. to film *The Land* (1942). Lorentz also brought the power of documentary production—both in informing a people and shaping its values—to the attention of the United States government. The United States Film Service, which Pres. Franklin D. Roosevelt created specifically after seeing Lorentz' two films, was the first of a series of government film projects, eventually leading to the important government documentaries during World War II and the creation of the United States Information Agency after the war. On a more popular level, Louis de Rochemont's "March of Time" series, inaugurated in 1935, was a means of giving commercial film patrons a combination of interesting information on current issues and events as well as editorialized propaganda about those issues in the news-heavy years preceding the war.

England. The documentary movement in England played a much more central and vital role in the health of Britain's film culture. Whereas the commercial British film industry in the 1930s was primarily a colony of Hollywood, dominated in the era of sound by the common language that both nations shared, the most distinctive British products were documentaries. Funded and produced by the British government (the General Post Office or Crown Film Units), which accepted the principle of state production much more quickly than did the United States government, the most effective British documentaries were directed by John Grierson, Alberto Cavalcanti, Basil Wright, Humphrey Jennings, Harry Watt, Len Lye, Paul Rotha, and the Boulting brothers. The British historical epics produced or directed or both by Alexander Korda (*The Private Life of Henry VIII*, 1933; *Catherine the Great*, 1934; *Rembrandt*, 1936) were distinguished in their visual detail and texture and their smooth, realistic acting (notably by Charles Laughton).

Importance of documentaries in England

The commercial domination of Hollywood was so great in the 1930s in Great Britain that the government passed quota laws against U.S. films, intended to protect the native British industry. The laws only guaranteed that extremely cheap, incompetent "quickies" would be cranked out by British studios to satisfy the quotas, allowing the British cinemas to show all the Hollywood films they desired. England's greatest film stars (Laughton, Leslie Howard, Sir Laurence Olivier) went to Hollywood as soon as they were invited, as did England's one unique, imaginative director, Alfred Hitchcock.

Hitchcock's films

The Hitchcock film, both in England and the United States, was an exciting mixture of suspense and whimsy, murder and wry comedy, a driving narrative line and delightful digressions, social observation, and psychological aberration. Hitchcock's two dominant plots were the frantic chase for some vital "political" information (*The Thirty-nine Steps*, 1935; *The Man Who Knew Too Much*, 1935 and 1956; *The Lady Vanishes*, 1938; *Saboteur*, 1942; *North by Northwest*, 1959) or the psychological study of a single personality built with the same rhythms of the suspenseful chase (*Suspicion*, 1941; *Shadow of a Doubt*, 1943; *Spellbound*, 1945; *Strangers on a Train*, 1951; *Psycho*, 1960). As the Hitchcock films rush breathlessly toward their climactic solutions, the director weaves in bizarre interruptions that minutely examine an unfamiliar social milieu or wildly freakish occupation and that, ironically, prove not to be interruptions but vital strands of the plot. Hitchcock films succeed in making the completely incredible seem credible by piling one incredibility on top of another so that the incredible becomes logical, by developing the texture of each incredibility with great detail, and by moving from bizarre twist to twist with great speed and tension. If there is an essential Hitchcock theme, it would be the deception of surface appearances—that proper-seeming people can be very improper or those who seem mad are sane and vice versa. Because nothing is as it seems, the film's protagonist has

even more trouble solving the riddle, and the film's audience has even more fun watching the twists and turns.

As a technician, Hitchcock is a master of controlling the "plastic material." Hitchcock, like Lubitsch, manipulates concrete objects that mean so much more than they seem. Hitchcock is also a master of rhythmic, suspenseful cutting and particularly of intercutting sound and picture, underscoring the rhythm of his cuts with sound effects or using a sound effect to motivate a cut. He uses sound brilliantly by making many of his films revolve around musical motifs. Hitchcock, however, subordinates this technical control to creating a whole film world in which murderers, psychotics, and spies are no stranger than the little old lady next door, in which insanely exotic tales hang on the most trivial details, and in which the most unusual, abnormal settings became backgrounds for incredible chains of events that one accepts as credible because of their social detail, human comedy, and hypnotic rhythm.

1945 TO 1965

United States. During World War II, Hollywood contributed to the war effort both directly and indirectly. Many of the best U.S. directors—John Ford, Frank Capra, William Wyler, Garson Kanin, John Huston—made documentary films for the government. These films effectively mixed information and propaganda, combining newsreel and combat footage of battles and invasions with an editorial affirmation of the need to rescue the United States' freedom-loving allies and a warning about dangers to the United States' survival if those allies were to fall. The "Why We Fight" series, produced by Frank Capra and directed by Capra and Anatole Litvak, was the most notable group of war films to explain U.S. policy prior to the war and the necessity of fighting it after its brutal eruption (*Prelude to War*, 1942; *Divide and Conquer*, 1943; and *Battle of Russia*, 1944, were probably the three most effective films in the group). Other films combined combat footage of U.S. men in action with a stirring editorial tribute to human courage, personal sacrifice, and moral commitment (John Huston's *Report from the Aleutians*, 1943, and *The Battle of San Pietro*, 1944; John Ford's *Battle of Midway*, 1942; William Wyler's *Memphis Belle*, 1944). Whereas the propagandist "documentary" film played a minor and very tentative role in World War I, nations on both sides of the Atlantic had come to realize the power of film in shaping values and commitment by World War II.

Hollywood also contributed to the war effort by adding special taxes to theatre tickets. But Hollywood's greatest contribution during the war was probably its continuing to go about its usual business, producing the purely entertaining pictures that boosted the morale and spirits of both civilians and soldiers, taking their minds off the grim realities of the battlefield. The war, in turn, contributed to Hollywood's health, keeping the industry's profits up and the government's interference at a minimum. The U.S. industry's greatest commercial year was 1946, when box-office receipts climbed to just under $1,700,000,000. Box-office receipts fell every year thereafter (in an era of rising labour costs and inflation), sinking to about half the 1946 total by 1962.

Peak of Hollywood's success

The primary assault against Hollywood's audience came from television. More and more people stayed home every year to watch programs on their television sets. Gradually, the motion picture lost its role as the United States' primary means of mass entertainment. The United States courts were secondary assailants, finally ruling in 1949 (in the landmark Paramount case, which had begun in the courts 11 years earlier) that it was illegal for studios to operate their own chains of theatres for their films. The anti-monopoly ruling, however, was probably academic, like the legal ruling against the Patents Company trust in 1915. Commercial realities and audience demands preceded legal decision. There would be no audience to support either the huge theatre chains or the huge studios within less than a decade after the war. As the audiences shrank, the number of feature films, stars, contract players, technicians, executives, and publicity men also shrank. The studio system steadily lost its reason to exist as the yearly demand for large quantities of films steadily dwindled. Also, as the demand for feature films dwindled, the demand for the studios' supporting products—the comedy shorts, animated cartoons, travelogues, and newsreels—dwindled and died.

Hollywood's initial reaction was panic, which took several forms. In times of financial crisis, Hollywood had consistently imposed moral restrictions in the hope of luring the audience back with the support of the pulpit and the politician. The creation of the Hays Office in the 1920s and the Breen Code in the 1930s was succeeded by the industry's own witch-hunts as well as its subservient grovelling before congressional investigating committees in the late 1940s and early 1950s. Ironically, some of the very writers and directors who contributed to the war effort a few years earlier lost their jobs simply upon suspicion of being Communists, former Communists, Communist sympathizers, or "pinkish" leftists. Despite Hollywood's official piety, audiences continued to shrink.

Technical experimentation was another sign of panic. Hollywood introduced new cinematic gimmicks in the mid-1950s: 3-D, Cinerama, Smell-O-Vision, and others. Fortunately, several of the most gimmicky experiments were short-lived. Hollywood's intention was to provide a kind of entertainment that television could not. And television could throw neither props (3-D) nor smells (Smell-O-Vision) into the living room, nor could it successfully give the sensation of a roller-coaster or airplane ride (Cinerama) on an immense wide screen. The new devotion to mammoth spectacular films—*The Greatest Show on Earth* (1952), *The Robe* (1953), *The Ten Commandments* (1956), *Ben Hur* (1959)—was the result of the same panic. In these films, producers sought something grandiose that would stretch across the new wide screens and would overwhelm the puny inches of the image on a television screen.

Colour filming

The almost total conversion to colour filming in the mid-1950s can also be traced to the same source. Although a dependable Technicolor process had been perfected at about the same time as synchronized sound, only a very special studio offering (a lavish, historical film; a musical; the animated cartoon) could afford the additional time, care, and expense of colour filming. Technicolor originally required a very bulky three-lens camera, special technicians, and special artistic consultants. After the war, Hollywood needed colour to compete with a still black-and-white television. Several new and competing colour processes (Technicolor had previously been a monopoly supported by all the studios) lowered the costs and smoothed the difficulties of colour filming. Ironically, after television perfected colour in the 1960s, a motion picture had to use colour not to compete but to sell the suddenly valuable television rights to show the film on television.

Yet another reaction to the commercial chaos of the 1950s, in a completely different direction from Hollywood's preoccupation with size and dollars, was the birth and growth of America's "underground" film movement. The underground film maker rejected the huge, sophisticated equipment for his own portable, hand-held 16-millimetre camera; he rejected conventional plots for his own abstract, poetic, metaphoric, or satiric structures or non-structures; he rejected Hollywood's impersonal gloss for his own personal (if also technically rough) statement. The films of Stan Brakhage, Bruce Baillie, Ron Rice, Stan van der Beek, Kenneth Anger, Jack Smith, and Jonas and Adolfas Mekas rebelled against Hollywood's cinematic, social, sexual, intellectual, and psychological values. Although Hollywood and its patrons ignored the underground in the 1950s, the underground had ironically captured much of the aesthetics of commercial filmmaking by the late 1960s.

The U.S. commercial war between television and the motion picture was over in 1956, and the motion picture had lost it. Both motion pictures and motion-picture stars appeared on television, from which they had previously been banned. The studios began producing series and commercials for television, needing something to occupy

the vast sound stages and the large studio staffs. As the number of feature films dwindled each year, the sense of the U.S. film industry as a vital, fertile national movement also withered. Many of the greatest directors of the studio era had died in body or in spirit. Ford, Hawks, Cukor, and Hitchcock—the accomplished directors who survived—made films that deepened and broadened their own concerns but in no way blazed new trails in the possibilities of the motion-picture art. The new directors produced competent, entertaining, intelligent films in familiar, uninventive styles: George Stevens (*A Place in the Sun,* 1951; *Shane,* 1953), Fred Zinnemann (*High Noon,* 1952; *From Here to Eternity,* 1953), John Huston (*The Treasure of the Sierra Madre,* 1948; *The African Queen,* 1951), Elia Kazan (*Viva Zapata!,* 1952; *On the Waterfront,* 1954; *East of Eden,* 1955), William Wyler, Joseph L. Mankiewicz, Otto Preminger, and Stanley Kramer. These well-acted, clearly photographed dialogue films often merely used the camera and sound track to transcribe some hit play or best selling novel. Although many of these films yearned to say something socially, morally, and personally "significant," their social consciousness was extremely self-conscious and muted, strangled by fears of Hollywood political suppression, on the one hand, and by a devotion to Hollywood stylistic clichés, on the other.

Decline of comedies and musicals

The great comic films of the previous three decades withered into cheap, careless, formula entertainment with Abbott and Costello, Dean Martin and Jerry Lewis, Bob Hope, Red Skelton, Ma and Pa Kettle, and Francis, the Talking Mule. The dialogue comedies of Billy Wilder (*Stalag 17,* 1953; *Sabrina,* 1954; *Some Like It Hot,* 1959; *The Apartment,* 1960), however, effectively carried on the Sturges tradition of mixing satiric, unconventional comedy with a serious moral point. Another genre, the original musical comedies of M-G-M's waning years (directed by Vincente Minnelli, Charles Walters, Gene Kelly, and Stanley Donen), hit a peak of musical and visual imagination in the decade following the war, but by the mid-1950s Hollywood was producing mammoth, ponderous adaptations of Broadway musicals.

While both the U.S. film industry and film imagination decayed, the greatest contributions to motion-picture art came from smaller nations, the film industries of which were less dependent on quantity, where film making had always been more personal (if less slick) than in the United States, and where television took away the audience much more slowly. Indeed, the non-U.S. film saw its market expanding, not shrinking, for it found a new home on U.S. screens, which had been formerly very unreceptive to European "art" films, since sound erected its linguistic barriers. In contrast to the Hollywood film, the best films from Europe, India, and Japan were more sensitive to realistic social problems, to human psychological needs, and to flexible systems of moral values in a very complicated and rapidly changing world. The films translated into cinematic terms many of the intellectual and moral questions raised in the century's greatest novels, plays, and philosophic essays. These motion pictures were also more adventurous in their handling of the camera, editing, and the sound track—freeing themselves from the tyrannies of dialogue and of theatrical space and time that had begun to smother the U.S. film. The postwar European film, in particular, recaptured much of the visual freeness of the last decade of silent filming. And, just as the post-1918 European cinema helped the U.S. industry discover the fuller possibilities of its art, the post-1945 European cinema eventually gave the U.S. film its new direction and identity as well.

Neorealist movement

Italy. Italy's Neorealist movement showed human beings struggling with four basic problems of survival: against the war and the murderous political chaos that invaded their land, against hunger, against the poverty and idleness of unemployment, against the dissolution and corruption of the family. The Neorealist films straddled the boundary between fiction and documentary, emphasizing poverty rather than Hollywood glamour; ugly, bare hovels rather than tidy homes and chic apartments; and common, defiant, earthy people, such as Anna Magnani, as actors and actresses, rather than Hollywood's elegant gods and goddesses. The Neorealist films adopted the Marxist principle that unjust social structures pervert the essential human relationships and that both love and friendship become dangerous and elusive luxuries in a society in which there are not enough work, not enough money, and not enough food. Roberto Rossellini's *Open City* (1945) and *Paisan* (1946) were among the first Neorealistic statements, but the fullest and most moving of them were Vittorio De Sica's films, with scripts by Cesare Zavattini.

Scene from *Satyricon* (1969), directed by Federico Fellini.
By courtesy of United Artists, Inc.; photograph, Franco Pinna—Rapho Guillumette

Vittorio De Sica. De Sica's *Bicycle Thief* (1948) is remarkable in its ability to develop so much frustration and pathos from the simplest of stories. A worker, who needs his bicycle for the job that will support his wife and son, sees a thief steal that bicycle on his first day at work. The film becomes a desperate search for the bicycle, taking the man, his friend, and his son on a hopeless search through the slums and junk markets of Rome. When Antonio catches the thief, he discovers that the culprit is merely a starving young man like himself, a thief out of the same necessity that they all face. Unable to retrieve his bicycle, Antonio finally steals another (he is the film's second bicycle thief), only to be caught and abused by an unsympathetic and uncomprehending mob. The man's degradation is complete, and only the gentle affection of his young son provides him with a reason to continue to exist.

De Sica's later films (also with scripts by Zavattini) consistently examine human needs and desires with sensitive detail and compassion, juxtaposing the fervent human attempts to survive with the immense social, economic, and political chaos surrounding them: *Miracle in Milan* (1951), *Umberto D* (1952), *The Roof* (1956), *Two Women* (1961), *The Garden of the Finzi-Continis* (1971).

Luchino Visconti. The contrast of poor, humble man and malignant, impersonal society became the basis of hundreds of postwar Italian films. But, as the country (and the country's film industry) became richer, the Italian films drifted further from poverty, hunger, and social inequities. Whereas Luchino Visconti began with realistic, black-and-white films about the poor and squalid (*Ossessione,* 1942; *La terra trema,* 1948; *Rocco and His Brothers,* 1960), he soon drifted to historical films in colour (*The Leopard,* 1963), grandiose allegory in colour (*The Damned,* 1969), and mammoth literary adaptations in colour (*Death in Venice,* 1971), all of them adorned with sumptuous elegance and elaborate decor.

Change in Italian films.

Federico Fellini. Fellini also began with modest, personal studies of individual people and drifted toward the multimillion-dollar spectacle of *Satyricon* (1969). Fellini's journey from modest poverty to mammoth visual circus is quite understandable. The circus and variety hall had always been his favourite domains (which he proved in his intimate study of *The Clowns,* 1970), often forming

the backgrounds for the lives of the poor people he tenderly studied. Fellini's subject in his earlier films was not how the social milieu oppressed the poor and starving but how the poor and starving went about the business of trying to find love, the only necessity that made life worth living. Fellini's Cabiria (*Nights of Cabiria*, 1956), a prostitute who has plenty of money, gladly sacrifices it for the man who will take her off the streets. Gelsomina (*La strada*, 1954) is a foolish waif who requires affection more than food. Her death from the cold and snow is the fault of her brutal, callous lover-master, not of any social system in the abstract. Fellini's other early films (*Variety Lights*, 1950; *The White Sheik*, 1952; *Il bidone*, 1955; *I vitelloni*, 1953) are similarly close, personal, sensitive studies of love and rejection, human responsiveness and callousness, many of them featuring his own wife, the actress Giulietta Masina, as the source of their laughter and tears.

Fellini's later work expands the scope of his human circus. These films consistently reveal the director's urge to generalize, to convert his grotesquely compelling images of human faces and social events into intellectual symbols. *La dolce vita* (1960) contrasts the collapse of life's spiritual meaning (the commerciality and deception of the church, a constant Fellini theme; the impotency of the intellectual life) with the new vitality of sensual experience (the escape into riches, drink, and sex). The film *8½* (1963) is a bewilderingly complex study of the director's own life and art, a story in which a film director's imaginative fancies blend into his life and his life into his artistic creations, causing his agonized suspicion that he is a failure both as man and as artist. *Juliet of the Spirits* (1965) is Fellini's first colour circus, a hypnotic series of dazzling vignettes that contrast the stifling propriety of the bourgeois, conventional life with the lures of sensuality and amoral spontaneity. Fellini's devotion to flamboyant colour, grotesque faces and makeup, and sumptuously rich settings and costumes culminates in his *Satyricon*, a film that plunges the audience into a sensual Roman netherworld, where the spectator roams about aimlessly, like the film's protagonists, in a waking dream of colour, bizarre images, and sexual gratification, the only purpose of which is to be savoured and felt, like any circus.

Michelangelo Antonioni. Michelangelo Antonioni is far more introspective, more internalized, and much quieter than Fellini. The essential Antonioni theme is the anguish of modern life, and his essential human study is the way people (especially sensitive people) come to terms with that anguish. Contemporary life forms the backdrop for Antonioni's films: the world of the rich in *L'avventura* (1960), the literary world of *La notte* (1961), the stock exchange in *L'eclisse* (1962), the factory in *Red Desert* (1964), the chic world of fashionable success in *Blow-Up* (1966). Antonioni sets a sensitive human being (often the actress Monica Vitti) against these backdrops, using the camera and sound track subjectively to mirror the way she feels the hard, cold, intimidating world around her.

Antonioni's images and sound devices

Antonioni's favourite images are slick glass windows, white walls, shiny aluminum structures, polished terrazzo and stone, metal pipes, and barren, jagged rocks. His favourite sound devices are scraping, clanging, chugging, and thumping noises, as well as deathly silence. His films use little music and little dialogue, because they are about quiet people—sensitive characters who do not trust talk as a means of expressing their feelings. Also he finds talk less communicative than image, gesture, and noise. When Antonioni adopted colour photography in *Red Desert*, he used colour subjectively, to mirror the way the central character felt the world, not to present a dazzling Fellini circus for the eye. Most of Antonioni's films are not stories but circular journeys that leave the action at the end exactly where it was at the beginning, except for the increased awareness of the central figure. Though life does not look any brighter at the end of an Antonioni film, its betrayals, frustrations, and brutalities are more understandable and, hence, more bearable for both the central figure and the audience who has lived through the emotional experience with that character.

Other directors. The Italian film was extremely rich in the 25 years following the war—rich in imagination, in influence, and in profits, which have flowed into Rome's film studios from all over the world, especially from the U.S. Pier Paolo Pasolini began with the Neorealistic *Accatone* (1961) and progressed to biblical–Marxist spectacle, *The Gospel According to Saint Matthew* (1964); sexual–Marxist allegory, *Teorema* (1968); and sexual–allegorical spectacle, *Medea* (1970) and *The Decameron* (1971). Pietro Germi directs stingingly acid satires of provincial Italian hypocrisies and infidelities: *Divorce—Italian Style* (1962), *Seduced and Abandoned* (1964), and *The Birds, The Bees, and The Italians* (1966). Ermanno Olmi directs sensitive, Neorealist studies of sympathetic human beings caught in the gears of industrial, bureaucratic society—*The Sound of Trumpets* (1961) and *The Fiancés* (1963). Ironically, Italy capitalized not only on its sensitive artistic masterpieces but on the cheapness of its labour and film-production methods. In the 1950s and 1960s Italy sent Hollywood countless third-rate historical spectacles (often starring American musclemen) and cowboy films (called "Spaghetti westerns") as well as its "art" films by the De Sicas, Fellinis, and Antonionis.

France. *The continuing tradition.* Whereas the Italian cinema leaped off in a revolutionary direction after World War II, the French cinema first tried to repair the bridge to the past. Both *Les Enfants du Paradis*, which cost so much money, and *Zéro de Conduite*, which had been banned for its iconoclasm, could be released just after the liberation. René Clair returned to France to make sentimental, fanciful, occasionally effective films that were less innovative and much heavier than his earlier works. Jean Renoir also returned from Hollywood to make three of his greatest films, all of which were in colour: *The River* (1951), located in India; *The Golden Coach* (1953), set in early-19th-century South America; and *French-Cancan* (1955), which took place in 19th-century Paris. These films were magnificent in their visual splendour, in their saturation with the music of the particular culture, and in the director's own new sense of satisfaction with his life and art. As opposed to the bitter irony of his motion pictures of the 1930s, the late Renoir films develop the mood of "consent," that crucial term introduced in *The River* and developed specifically in his next two films, which are both about artists who have learned to consent to their lives.

Return of Clair and Renoir to France

Among the other postwar French directors who worked in traditional, graceful, almost literary cinematic styles were Jean Cocteau, who mingled realism and symbolism in *Beauty and the Beast* (1945), *Les Parents terribles* (1948), and *Orphée* (1950); Robert Bresson, who combined his quiet, sensitive attention to people and settings with his affirmation of both human values and religious faith (*The Diary of a Country Priest*, 1951; *A Man Escaped*, 1956; *Pickpocket*, 1959); and René Clément, who wandered from a delicately horrifying child's view of war in *Forbidden Games* (1952) to glossy, leaden, picture-postcard thrillers that were part Hitchcock and part parody of Hitchcock (*Purple Noon*, 1959; *Rider on the Rain*, 1970). Max Ophüls migrated from the U.S. to France, where he made several lavish spectacles that combined an ornately decorated society of turn-of-the-century Europe, an ironic view of romantic and sexual experience, and a dazzling control of camera, cutting, structure, colour, and composition. The Ophüls films, which contrasted the shams and artifices of society with the undeniable power of sexual lust and emotional need, were as intellectually subtle as they were visually hypnotic (*La Ronde*, 1950; *Madame de. . .*, 1953; *Lola Montès*, 1955).

Jacques Tati made a rather literal leap into the past with a series of essentially silent comedies that used music and brilliantly comic sound effects rather than talk, techniques such as those used in Chaplin's *City Lights* and *Modern Times*. Tati's Mr. Hulot, like Chaplin's tramp, punctures society's self-satisfied rules, forms, and values with his own natural spontaneity and with the director's miraculously clever visual gags. In *Mr. Hulot's Holiday* (1953), Tati's target is the bourgeois conventionality of a

stifling seaside resort; in *Mon Oncle* (1958), the inhuman sterility of a modern, bourgeois house full of maddening gadgets and plastic utensils. Life in the house is as plastic as its inanimate objects, for its owner also runs a factory that manufactures plastics. *Playtime* (1968) comically condemns the spread of U.S. culture in the form of both tourists and bureaucratic business procedures. Tati also condemns the non-Americans who surrender their own lives and culture so as to make money in the U.S. manner from those Americans. *Traffic* (1971) is a comic study of man's perverse attachment to his automobile, his conversion of himself into a machine, and his consequent divorce from both the world of nature and of spontaneous, natural feelings.

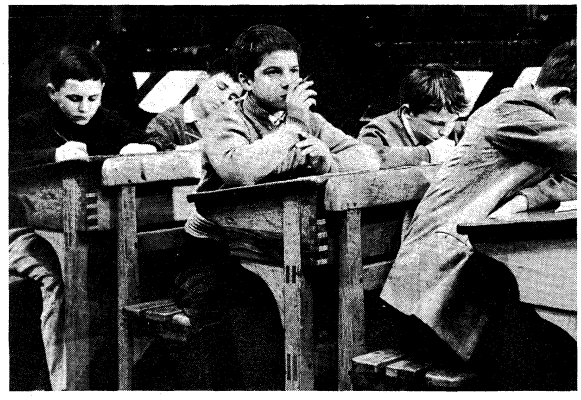
The New Wave. By 1960 an iconoclastic generation of French film critics and film makers rebelled against the classical, carefully written studio films that had dominated French film style since the introduction of synchronized sound. The young film makers applied to their own films the critical principles they had developed in writing for the magazine *Cahiers du Cinéma*. Jean-Luc Godard's *Breathless* (1960) officially marked the beginning of this New Wave; but several directors (Roger Vadim, Louis Malle, Claude Chabrol) had made New-Wave films before it, and two other extraordinary pieces of cinematic freedom, François Truffaut's *400 Blows* and Alain Resnais's *Hiroshima mon amour,* appeared in 1959. The parallel characteristics of all three films were the violation of linear time (huge leaps forward and backward and elliptical gaps in the story), an audacious mixture of varying cinema styles within the same film, and a spirit of spontaneity, improvisation, and freedom that was reflected in both the films' characters and in their cinematic style. For Truffaut and Godard especially, the spontaneous, irrational surprises and the breezy playfulness in the films of Vigo and the early comedies of Renoir and Clair served as models for the kinds of loose, buoyant films that they wanted to make.

Characteristics of the new films

Godard began as an Existentialist, passed through a period as Brechtian comic parodist, and then came to regard himself as a Marxist–Maoist lecturer. From the breezy Michel Poiccard in *Breathless*, a gangster in the Bogart mold who was charming, pleasant, warm, and completely committed to life as a series of spontaneous impulses rather than as a planned program, Godard switched to characterless, humourless, political debates in which his figures mouth banal platitudes (*La Chinoise*, 1967; *Le Gai Savoir*, 1968; *One Plus One*, 1968). Godard's spontaneity, humour, and human observation produced his best films, which either juxtaposed elusive and irrational feelings with the mechanical, rational repetitiveness of daily life (*Breathless*; *Vivre sa vie*, 1962; *Masculine/Feminine*, 1965) or became little parables of human and social existence that mixed parody of other films as well as literature with brutality (*Les Carabiniers*, 1963; *Weekend*, 1967).

Truffaut remained consistently committed to spontaneity, to human interaction, to a comic view of life in which love and sincerity are very difficult but still very possible.

By courtesy of Janus Films, Inc.; photograph, from the Museum of Modern Art Film Stills Archive, New York



Jean-Pierre Léaud as Antoine Doinel (centre) in *The 400 Blows* (1959), directed by François Truffaut.

His *400 Blows* shows the adolescent Antoine Doinel's breaking free from the stifling confinement of his childhood—his family, his school, a reformatory. Truffaut traces Antoine's later life in *Stolen Kisses* (1968), another story of Antoine's breaking away—from the army, from stifling (and comic) jobs—until he finally blunders into love. And in *Bed and Board* (1970), Antoine still loves his wife but also craves his freedom to experience. He wriggles free and then wriggles back to the marital relationship that really means something to him. Truffaut's art and the story of Antoine Doinel are intimately bound up with Jean-Pierre Léaud, the young actor Truffaut found for the first film and who continued to perform so vigorously and engagingly in the remaining Doinel films. Truffaut's non-Doinel films are equally committed to human sensitivity, spontaneity, and sincerity—in their cinematic styles as well as their plots: *Shoot the Piano Player* (1960) is a story of a honky-tonk pianist who cannot run away from his ability to love, and *Jules and Jim* (1961) is a story of three spontaneous people who love each other very much but cannot do so without also hurting each other and confining themselves and their own feelings. *The Wild Child* (1970), dedicated to Léaud, is a story of the film maker's own determined commitment (Truffaut plays the main role) to educating and to love. Truffaut's films are works of both education and love or, rather, education about love.

The films of Alain Resnais are much colder, quieter, and hollower than those of Truffaut and Godard, the works of a much more skeptical, older man. *Hiroshima mon amour* reveals the deadness of two people, their inability to love and respond to each other because of the deadness of their atomic world and the deadness of their own horrifying pasts. The Japanese architect and French actress are already dead inside, and Resnais uses images of cold, hard sterility to reveal that these people are stones. *Last Year at Marienbad* (1961) is another dazzling exercise in cinematic deadness, a series of flat, cold, gleaming images so confusingly interwoven, so bewilderingly arranged in space and time that it is difficult to make any rational story out of the film. It is a nightmarish experience of people frozen into statues, nature crushed into a painted landscape, and experience paralyzed into a series of photographs. *La Guerre est finie* (1966) is a much clearer and warmer story of a man who is virtually dead and whose life's work does not exist any more. He, too, is a prisoner of the past. In order to give his life meaning in the present, he must continue living as if his past were not dead and the war were not over.

Resnais's film world of living death

In the years since 1959, the New Wave has swept up many other directors. Claude Chabrol directed thrillers and case studies of psychological perversity (*Les Cousins*, 1958; *Les Bonnes Femmes*, 1959; *Les Biches*, 1968; *Le Boucher*, 1971), while Roger Vadim's films seem to fulfill the United States' stereotyped view of French sexiness (*And God Created Woman*, 1956; *Les Liaisons dangereuses*, 1959; *Barbarella*, 1968). Louis Malle jumped eclectically from classical, "literary" films (*The Lovers*, 1958) to irrational, incongruous comedy (*Zazie dans le métro*, 1960) to a combination of the classical world of sex and money with the child's world of spontaneity (*Murmur of the Heart*, 1971). Costa-Gavras combined political commitment with satiric comedy, exciting suspense and chases, and spontaneous cinematic tricks (*Z*, 1969; *The Confession*, 1970). The introspective, reflective, intellectual films of Agnes Varda (*Cleo from 5 to 7*, 1962; *Le Bonheur*, 1965) and Eric Rohmer (*My Night at Maud's*, 1969; *Claire's Knee*, 1970) have further enriched the French postwar cinema.

Sweden. The mid-1950s also brought the Swedish film industry to international attention, particularly through the films of Ingmar Bergman. Like the great Swedish silent films, Bergman's developed the interrelation of man and nature and of the reality of human existence with the metaphysical reasons for that existence. His films are dominated by men who have fallen out of touch with nature because of organized religion (*The Seventh Seal*, 1956), bureaucratic, bourgeois laws and assumptions (*The Magician*, 1958; *The Ritual*, 1969), a deadening

Bergman's interrelation of man and nature

The Seventh Seal (1956), directed by Ingmar Bergman.
By courtesy of Janus Films, Inc.; photograph, from the Museum of Modern Art Film Stills Archive, New York

and insensitive routine (*Wild Strawberries*, 1957), falsely idealized and romanticized notions of love and human relationships (*Smiles of a Summer Night*, 1955), and a loathing for an existence that cannot be rationally explained (*Through a Glass Darkly*, 1961). Many Bergman characters, feeling the total chaos of an inexplicable universe, escape from reality into a neurotic, selfish, fractured vision of people and the world (*Persona*, 1966; *The Passion of Anna*, 1969). Nature, however, reasserts itself in these films, with images of sunlight, trees, fields, skies, and flowers, showing the faltering, anguished humans its potential warmth and beauty if they will only allow themselves to see it. In his intelligent dialogue, glowing natural images, grimly ironic comedy, careful handling of music and the sound track, and capriciously clever manipulation of occasional cinematic tricks, Bergman combines his Existential quest with cinematic power.

Other contributions to postwar cinema from Scandinavia include the films of Alf Sjöberg (Bergman's mentor), *Torment* (1944) and *Miss Julie* (1951), the romanticism of Bo Widerberg (*Elvira Madigan*, 1967), and the titillating, inexpensively produced "stories" of ladies named Inga, Helga, and the like that are among the most popular entrées in the theatres devoted exclusively to films that in an earlier day would have been banned as pornographic.

Britain. The British film acquired a new individuality in the years following the war. The authenticity and realistic detail of the documentary tradition carried over into the fiction films, adding realistic texture and understated, realistic acting to genres as diverse as polished adaptations of literary classics directed by David Lean (*Great Expectations*, 1946; *Oliver Twist*, 1947), taut thrillers directed by Carol Reed (*Odd Man Out*, 1947; *The Third Man*, 1949), and quiet, realist comedies produced by the Ealing Studios, many of them with Alec Guinness (*Tight Little Island*, 1948; *The Lavender Hill Mob*, 1951; *The Man in the White Suit*, 1952). The year 1959, as significant a cinematic year in Britain as in France, was the year of both Jack Clayton's *Room at the Top* and Tony Richardson's *Look Back in Anger*. For the next several years, the British specialized in these "angry young man" films, which combined the British feeling for realistic acting and for the detailed textures of documentary reality with the foreign influences of cinematic freedom from France and of Neorealistic observation of human and social behaviour from Italy. Behind the new British movement was also the shift in the British class system: the angry but intelligent appeal of the former proletarian who had acquired an education. As the 1960s progressed, the British film, like the Italian cinema, lost its feeling of grimy, working class, black-and-white poverty and began producing colour studies of historical eras and existential rebellions of self-expression that were almost indistinguishable in style from films from the United States (*If . . .*, 1968; *Oliver!*, 1968; *The Lion in Winter*, 1968; *Nicholas and Alexandra*, 1971).

Eastern Europe. Exciting films and film makers also emerged in the postwar era in the eastern European countries of Poland, Czechoslovakia, Hungary, and Yugoslavia. In the 1960s the Pole Roman Polanski and the Czechs Jan Kadar and Milôs Forman all demonstrated personal commitment and artistic control within a small but healthy national tradition, just as DeSica and Rossellini had done in postwar Italy. Because of political and artistic suppression in Communist eastern Europe, most of these film makers migrated to Hollywood, where they shifted from national, personal, and social themes and modest styles to international casts, Existential questions, and tricky cinematic styles.

Japan. The international film festivals led to the Western discovery of the films of Japan, particularly of Kurosawa Akira. The thriving Japanese film industry, which by 1953 produced more feature films yearly than Hollywood, also demonstrated the artistic maturity and sensitivity of its product in the 1950s. Kurosawa's films were of two types: historical epics of the golden age of Japanese warriors called samurai (*Rashomon*, 1950; *Seven Samurai*, 1954; *Yojimbo*, 1961) and stories of contemporary Japan (*Drunken Angel*, 1948; *Ikiru*, 1952). Kurosawa's stories of old Japan are modelled on John Ford's and Howard Hawks's studies of the old West; Kurosawa uses western motifs, plot twists, a reliance on scenery and the outdoors, and even the western's means of building suspense and tension. Behind the violent action of these films and the heroic self-assertion of their central characters, however, Kurosawa has posed serious and significant

Influence of U.S. westerns on Kurosawa

By courtesy of Daiei Motion Picture Co.; photograph, the Museum of Modern Art Film Stills Archive, New York

Rashomon (1950), directed by Kurosawa Akira.

human questions. *Rashomon* remains one of the most powerful film treatments of the 20th-century idea that truth is in the eye of the beholder and that no single objective, factual explanation for a human event can ever be found. Ironically, after Kurosawa copied the U.S. western, U.S. westerns copied Kurosawa, turning both *Seven Samurai* and *Rashomon* into cowboy films (*The Magnificent Seven*, 1960; *The Outrage*, 1964).

Among the many other Japanese films that have been acclaimed in the West are Mizoguchi Kenji's *Ugetsu* (1953); Kinugasa Teinosuke's *Gate of Hell* (1954), Japan's first film to use colour; and Teshigahara Hiroshi's *Woman of the Dunes* (1964), which used the subjective, impressionistic camera with great beauty and success. Because film production was so inexpensive in Japan, the U.S. also imported many cheap, commercial films in which various prehistoric reptiles (*e.g.*, Godzilla, Rodan) reawoke to devour plastic models of Japan's major cities.

India. Although the Indian motion-picture industry is as productive as Japan's, very few of its films have been seen in the West. Most Indian movies have been produced for the subcontinent's populous domestic audience and therefore use local themes, settings, music, and actors. Satyajit Ray's "Apu" trilogy (*Pather Panchali*, 1954; *Aparajito*, 1956; and *The World of Apu*, 1959), the story of a boy's growing up and becoming part of Indian society, has been India's most impressive export. Striking glimpses of the Indian settings, actors, and culture can also be seen in Jean Renoir's *River* and in the films of James Ivory (*e.g.*, *Shakespeare Wallah*, 1965), an American who directs in India. The growing popularity of Indian culture, art, music, and thought in the West since the 1960s may eventually lead to more international distribution of that country's films.

V. Recent trends

Internationalism a major trend

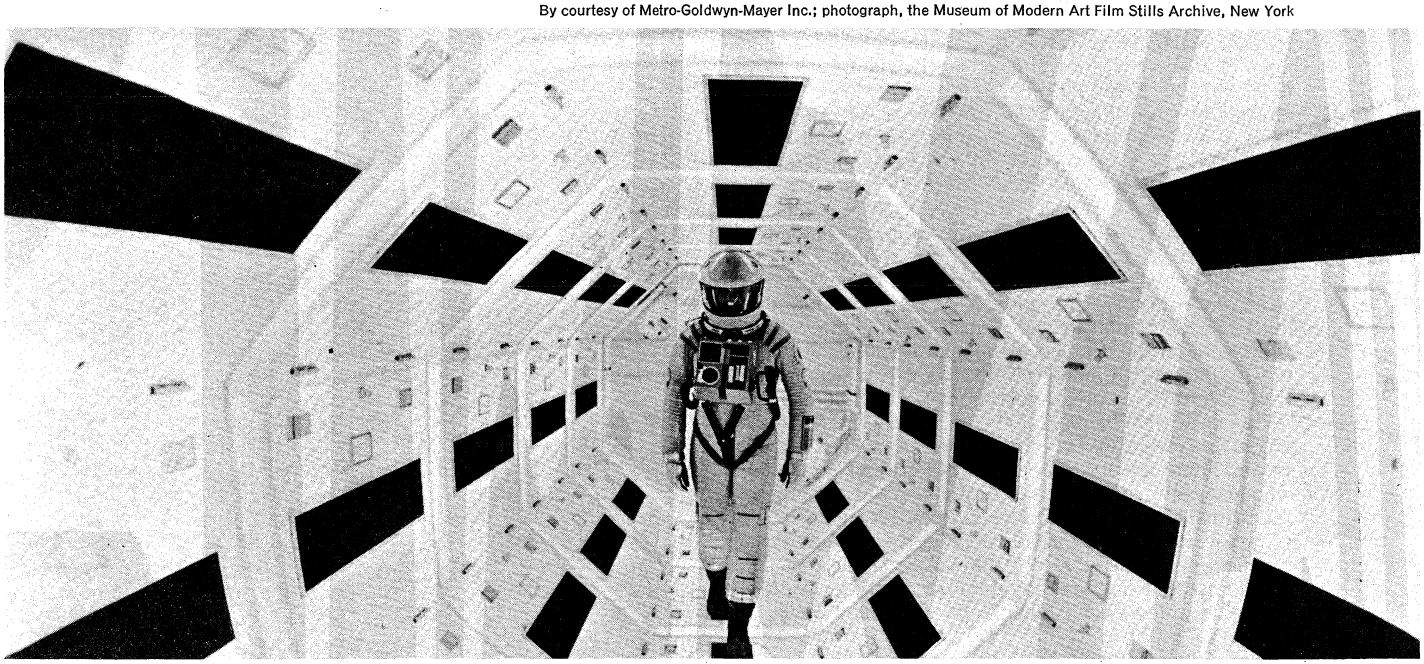
If nationalism was the major trend of the first two decades following World War II, internationalism would seem to be the trend after 1965. Not only does international distribution play a major role in the film industry of every country, but the international production (for example, an Italian director using British, Italian, French, and Swedish actors to film a German story) has levelled national and cultural boundaries. Perhaps this cinema internationalism was foreshadowed by the 45-year career of the wandering film director Luis Buñuel—a man without a cinematic country. He has made films in France (*Un Chien Andalou*, 1928; *L'Age D'or*, 1930; *Belle de jour*, 1966), in Spain (*Las Hurdes*, 1932; *Viridiana*, 1961), and in Mexico (*Los Olvidados*, 1950; *The Exterminating Angel*, 1962). The unity of this lengthy career stems neither from nation nor from a consistent cinematic style (Buñuel travels between total Surrealism and total social realism) but from the energy of the director's brutal, savage attack on social proprieties, on the Christian–Humanist interpretation of the needs of man, and on the pretentious falsity of any philosophical system that refuses to recognize the animal, bestial centre of human motivation.

The vitality, creativity, and commercial strength of film activity outside the United States eventually made their impact on the U.S. film industry. In the early 1960s, the attrition of both the art and the industry ended when Hollywood redefined its audience, its function, and, hence, its product. Since the motion picture was no longer the primary source of mass entertainment, it no longer needed to cater to the general tastes of the public—rich and poor, old and young, urban and rural. The industry addressed itself to the minority who were willing to see something they could not see on television. The new motion-picture audience was predominantly young (under 30), urban (or suburban), educated, and affluent enough to afford a high-priced ticket. Films could cater to even smaller minority audiences, such as urban blacks, devotees of rock music, homosexuals, "skin flick" addicts, and those interested in the documentary. Still other films were specifically designed for the whole family—a special attraction for a family outing. No longer did every family spend its weekly movie night at the "Bijou" (a U.S. colloquialism for old movie palaces often so named). There was no more Bijou, but, in its place, there arose smaller "art cinemas." A few anachronisms also remained, such as the Radio City Music Hall, New York City.

Influence of the elite audience

The catering to an essentially elite audience determined the content and styles of the films. The industry's uniform restrictions against sexuality, perversion, brutality, and unpleasantness disappeared. The Hollywood Code completely crumbled in 1966, to be replaced by a more flexible system of rating films according to their "level of maturity." In the new films, poetic justice did not always triumph; life did not always appear to be just and good; and virtue did not always defeat injustice. In fact, the new cinematic convention seemed to be the reversal of good and evil and the horrifying triumph of injustice. The members of the new elite audience were skeptical about social institutions and social justice and cynical about the idealistic beatitudes they heard in the mouths of politicians and saw in the movies of a previous generation.

By courtesy of Metro-Goldwyn-Mayer Inc.; photograph, the Museum of Modern Art Film Stills Archive, New York



Use of innovative cinematic devices of the 1960s in a scene from *2001: A Space Odyssey* (1968), directed by Stanley Kubrick.

Michael J. Pollard, Faye Dunaway, and Warren Beatty in a scene from Arthur Penn's *Bonnie and Clyde* (1967).
From the motion picture *Bonnie and Clyde*, Copyright © 1967 by Warner Bros.-Seven Arts, Inc.; photograph, from Pictorial Parade

The new hero was iconoclastic, rebellious, unsatisfied; he valued sincere emotional and sexual experience over moral platitudes and sanctimonious hypocrisies.

Along with their rebellious themes and characters, U.S. motion pictures adopted the rebellious, innovative styles of the European cinema. Films were no longer rational, linear stories but kinetic, emotional almost metaphoric experiences. Directors played as freely with space, time, and logical continuity as had the most fluid silent films. Tricks with the camera, the sound track, and editing techniques became the obligatory conventions of the new films: flashes forward and backward in time, kinetic uses of colour, split-screen effects, intrusive zooms, discordant and startling cutting, asynchronous sound effects, musical interludes that were metaphorically related to the feeling of the film but not its story. Many of the visual tricks of the U.S. underground films had been incorporated in the new commercial films. The most influential and impressive U.S. films of the 1960s inevitably combined an iconoclastic social statement with innovative cinematic devices. Such films included Stanley Kubrick's *Dr. Strangelove* (1964), *2001: A Space Odyssey* (1968), and *A Clockwork Orange* (1971); Robert Rossen's *Hustler* (1961); Sidney Lumet's *Pawnbroker* (1965); Arthur Penn's *Bonnie and Clyde* (1967); Sam Peckinpah's *Wild Bunch* (1969) and *Straw Dogs* (1971); Peter Yates's *Bullitt* (1968); George Roy Hill's *Butch Cassidy and the Sundance Kid* (1969); Peter Fonda–Dennis Hopper's *Easy Rider* (1969); Bob Rafelson's *Five Easy Pieces* (1970); John G. Avildsen's *Joe* (1970); Peter Bogdanovich's *Last Picture Show* (1971).

Commercial and artistic revolution

By 1970 the U.S. film industry and the industries of the rest of the world (they had become almost inextricable) had accomplished an immense commercial and artistic revolution. Hollywood was no longer the centre of the U.S. industry; more U.S. films were shot in New York, in Mexico, in Europe, and in various other locations than in the former film capital (indeed, the motion-picture industry in Hollywood became almost exclusively devoted to television production). Quantity film production had been transformed into a more individualized, personal, independent concern with quality. No longer considered the star's vehicle but the director's, audiences, critics, and even directors themselves had become conscious, indeed self-conscious, of the motion picture as an art form rather than as an entertainment medium. The art and history of the motion picture, therefore, had become subjects for serious study, and the film as a medium was accorded an equal cultural position with the older, traditional arts of the novel, drama, poetry, painting, and music. It is from such premises that the past history of the motion picture will evolve into its future. (Ge.M.)

BIBLIOGRAPHY

General histories: LESLIE HALLIWELL, *The Filmgoer's Companion*, 3rd ed. (1970), an encyclopaedic reference source for names, titles, terms, and genres; ARTHUR KNIGHT, *The Liveliest Art* (1957), and PAUL ROTHA, with RICHARD GRIFFITH, *Film Till Now*, 3rd ed. rev. (1960), two influential histories now dimmed by age and lack of critical perspective; PIERRE LEPROHON, *Histoire du cinéma*, 2 vol. (1961–63), useful reference work of names, dates, titles, and events; GERALD MAST, *A Short History of the Movies* (1971), a history of significant films and film makers presented in the context of the entertainment industry; RICHARD SCHICKEL, *Movies: The History of an Art and an Institution* (1964), a lively study of the entertainment industry.

Historical studies of specific periods: C.W. CERAM (pseudonym of K.W. MAREK), *Eine Archäologie des Kinos* (1965; Eng. trans., *Archaeology of the Cinema*, 1965), authoritative, well-illustrated study of cinema prehistory; PENELOPE HOUSTON, *The Contemporary Cinema* (1963), a survey of international trends after World War II; RICHARD DYER MacCANN, *Hollywood in Transition* (1962), personal observation of Hollywood's collapse in the 1950s; TERRY RAMSAYE, *A Million and One Nights* (1926, reprinted 1964), classic study of the cinema's infancy, with emphasis on American film between 1890 and 1915; GEORGES SADOUL, *Histoire générale du cinéma*, 6 vol. (1946–54), a mammoth, detailed study of the epoch of silent film.

Historical and critical studies of national film movements: JOSEPH L. ANDERSON and DONALD RICHIE, *The Japanese Film: Art and Industry* (1959), authoritative study of Japan's cinema history, directors, and major films; ROY ARMES, *French Cinema Since 1946*, 2nd ed., 2 vol. (1970), a survey of both the classical tradition and New Wave; ERIK BARNOUW and S. KRISHNASWAMY, *Indian Film* (1963), an authoritative study of India's industry, film art, and the relationship of film and society; KEVIN BROWNLOW, *The Parade's Gone By* (1968), nostalgic, well-illustrated study of American silent films and film stars; PETER COWIE, *Swedish Cinema* (1966), a survey from silents to Bergman; BOSLEY CROWTHER, *The Lion's Share* (1957), on the rise and fall of M-G-M; LOTTE H. EISNER, *L'Écran démoniaque*, rev. ed. (1965; Eng. trans., *The Haunted Screen: Expressionism in the German Cinema and the Influence of Max Reinhardt*, 1969), a study of the influence of the arts of painting, drama, and the novel on German cinema; DAVID STEWARD HULL, *Film in the Third Reich: A Study of the German Cinema, 1933–1945* (1969), Hitler's and Goebbel's use of the film as propaganda; LEWIS JACOBS, *The Rise of the American Film* (1939, reprinted 1968), a detailed study of American films through 1938, with emphasis on film trends and audience tastes rather than on individual works of art; SIEGFRIED KRACAUER, *From Caligari to Hitler: A Psychological History of the German Film* (1947, reprinted 1966), a psychological, sociological, and political study of German cinema, arguing that films unconsciously prepared the way for Hitler's accession; JAY LEYDA, *Kino: A History of the Russian and Soviet Film* (1960), an authoritative study from tsarist beginnings to 1958; RACHAEL LOW and ROGER MANVELL, *The History of the British Film*, 3 vol. (1948–51), a detailed study of the British silent film; GEORGES SADOUL, *French Film* (1953), a concise survey; ANDREW SARRIS, *The American Cinema: Directors and Directions, 1929–1968* (1968), classic definition of the "*auteur* theory" and its critical application to American films and film makers; PARKER TYLER, *Magic and Myth of the Movies* (1947, reprinted 1970), studio films seen as myth and dream, with a psychological analysis of the relationship of audiences to films of the 1930s; EDWARD WAGENKNECHT, *The Movies in the Age of Innocence* (1962), nostalgic reminiscences of silent films and stars.

Genre studies: RICHARD MERAN BARSAM, *Nonfiction Film: A Critical History* (1972), analysis of the art, themes, and values of primarily British and American documentary films; RAYMOND DURGNAT, *The Crazy Mirror* (1970), a sociological study of American comedy, reflecting audience values and tastes; JIM KITSES, *Horizon's West* (1970), Anthony Mann, Budd Boetticher, and Sam Peckinpah as representatives of the western; KALTON C. LAHUE, *Continued Next Week: A History of the Moving Picture Serial* (1964) and *The World of Laughter: The Motion Picture Comedy Short, 1910–1930* (1966), two surveys of significant stars, dates, producers, and titles; GERALD MAST, *The Comic Mind: Comedy and the Movies* (1973), thematic study of sound and silent comedies and the relationship between intellectual content and comic form; SHELDON RENAN, *An Introduction to the American Underground Film* (1967), a survey of significant films and film makers; PAUL ROTHA, with SINCLAIR ROAD and RICHARD GRIFFITH, *Documentary Film*, 3rd ed. rev. (1963), a classic study of the genre, written by one of its most influential practitioners, now somewhat dated and rigid, however.

(Ge.M.)

Motion Pictures, Technology of

The technology of motion pictures, based on the prior invention of still photography, developed rapidly after 1900 to meet the expanding needs of professional and amateur moviemakers. The technology, a complex one, consists, in fact, of several more or less independent technologies; that is, camera and projector design, film manufacture and processing, sound recording and reproduction, and lighting and light measurement.

This article is divided into the following sections:

I. History
- Early history
- Introduction of sound
- Introduction of colour
- Wide-screen and stereoscopic pictures
- Amateur motion pictures

II. Professional motion-picture photography
- Cameras
- Film
- Lighting
- Film processing and printing
- Sound-recording techniques
- Special effects
- Projection techniques and equipment
- Animation
- Motion-picture films in television

III. Motion pictures for scientific purposes
- Time-lapse cinematography
- High-speed cinematography

IV. Amateur motion-picture photography

Other articles in which specialized aspects of the subject are treated include SOUND RECORDING AND REPRODUCING; PHOTOGRAPHY, TECHNOLOGY OF; and OPTICAL ENGINEERING. Motion pictures as an art form are treated in MOTION PICTURES, ART OF; MOTION PICTURES, HISTORY OF.

I. History

EARLY HISTORY

The conditions necessary for generating the illusion of motion in a picture were set forth with remarkable clarity in a French patent of April 25, 1864, granted to Louis Ducos du Hauron.

> My invention consists in substituting rapidly and without confusion to the eye, not only of an individual but when so desired a whole assemblage, the enlarged images of a great number of pictures when taken instantaneously and successively at very short intervals. . . . The observer will believe that he sees only one image, which changes gradually by reason of the successive changes of form and position of the objects which occur from one picture to the other. Even supposing that there be a slight interval of time during which the same object was not shown, the persistence of the luminous impression upon the eye will fill this gap. There will be as it were a living representation of nature and the same scene will be reproduced upon the screen with the same degree of animation. . . . By means of my apparatus I am enabled especially to reproduce the passing of a procession, a review of military manoeuvers, the movements of a battle, a public fête, a theatrical scene, the evolution or the dances of one or of several persons, the changing expression of countenance, or, if one desires, the grimaces of a human face; a marine view, the motion of waves, the passage of clouds in a stormy sky, particularly in a mountainous country, the eruption of a volcano . . . . (From Terry Ramsaye, *A Million and One Nights*; © by Simon and Schuster, Inc., 1926.)

Persistence of the impression

Earlier devices, such as the Daedalum (later called Zoetrope—"wheel of life"—shown in Figure 1) had shown the possibilities of the principles set forth in the patent of Ducos du Hauron.

In carrying out these principles it is necessary to have the picture to view. At first, such pictures were drawn or painted by hand. In 1872 Gov. Leland Stanford of California, who was interested in the gait of racing horses, engaged Eadweard Muybridge, an English photographer, to take action photographs of a running horse. After many tries Muybridge was able to get successive pictures of a horse as it ran beside a wall. He used a battery of cameras facing the wall, and a picture was made as the horse went by each camera in turn. This was the first time that motion was recorded photographically.

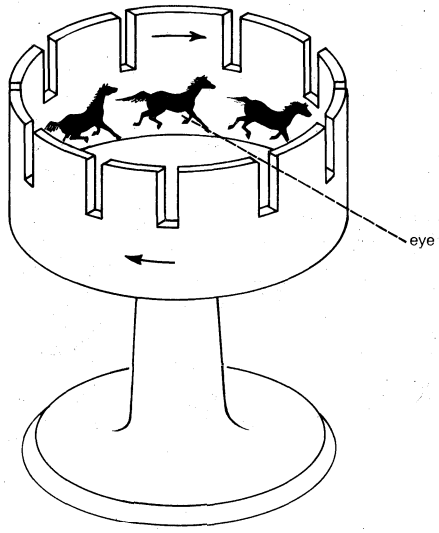


Figure 1: Zoetrope.

But, as long as photographs were made on glass plates, the pictures remained too bulky and fragile for even elementary cinematography. A few years later Émile Reynaud demonstrated a wide band of celluloid film, hand drawn and coloured, in his Théâtre Optique in Paris; and Thomas Edison, working with film made by George Eastman of much the same size as the present-day 35-mm standard width (see Figure 6), exhibited the Kinetoscope (Figure 2), an elaboration of the Zoetrope. The

By courtesy of J. Vivie, Paris

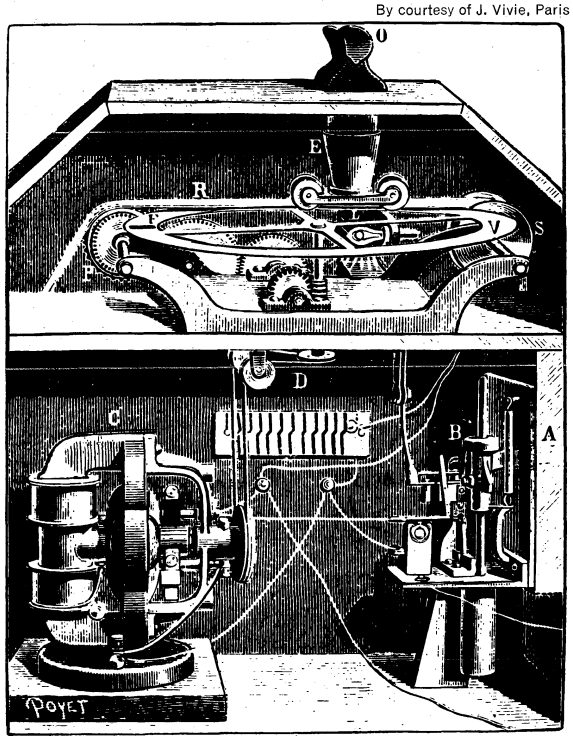


Figure 2: Edison's Kinetoscope, a device permitting one viewer to watch an illuminated strip of moving film one frame at a time through a narrow slit in a fast-moving shutter wheel. (A,B) Switching gear, (E) lens, (O) peephole, (R) film in continuous motion, (V) shutter wheel with slit (F). Engraving by Poyet.

pictures have been recorded on an extended film strip, R, which moves continuously in front of a light, L. They are viewed from a peephole, O, and lens, E, through a fast-moving shutter wheel, V, that has a narrow slit, F. This gives a short momentary view of each picture in turn. In the device, 46 separate pictures were viewed per second.

The pictures themselves were taken in a camera (the kinetograph) on a film that was drawn past an aperture.

Intermittent motion

The arrangement permitted the film to be exposed while it was stationary; it was then covered with a shutter while it was moved, and the sequence was repeated for another exposure. Several devices for obtaining the intermittent motion were used. At first the film at either end of the mechanism was stowed loosely inside the lighttight camera box. This became inconvenient as longer films began to be used, and the intermittent jerk was isolated to a relatively short loop between two continuously moving sprockets. Outside the sprockets the film moved uniformly and in time was wound neatly on reels.

At first most of the film exposed by Edison was photographed in a barnlike studio covered with black tar paper and called the Black Maria. It had large, movable pane windows at one end to let in appropriate natural light. The structure was mounted on a circular railroad track so that it could be turned with respect to the prevailing light. A sketch of the interior that appeared in 1894 showed a phonograph electrically connected with the camera, suggesting the possibility of a recording of sound pictures. Many years of arduous development and experience were needed, however, before acceptable sound pictures could be achieved.

The Edison camera and peephole Kinetoscope endured for a surprising time, remaining a feature of peepshow arcades as late as 1909. The device's great weakness was that it was wasteful of light, since the continuously moving film was cut off from the light for such a large fraction of the time. The small amount of light available made it impossible to project the picture onto a screen. One cause of the delay in the introduction of projectors was that the earlier so-called intermittents were very noisy, a situation that was intolerable in an auditorium.

In Paris in 1895, Auguste and Louis Lumière presented the *cinématographe*, in which the film was intermittently moved in both camera and projector. A special device incorporated a pin mechanism both to advance the film and to hold it in a firm position when stationary. Many designs for such devices were presented, ranging from complex machinery to the much simpler claw mechanism of the type shown in Figure 3.

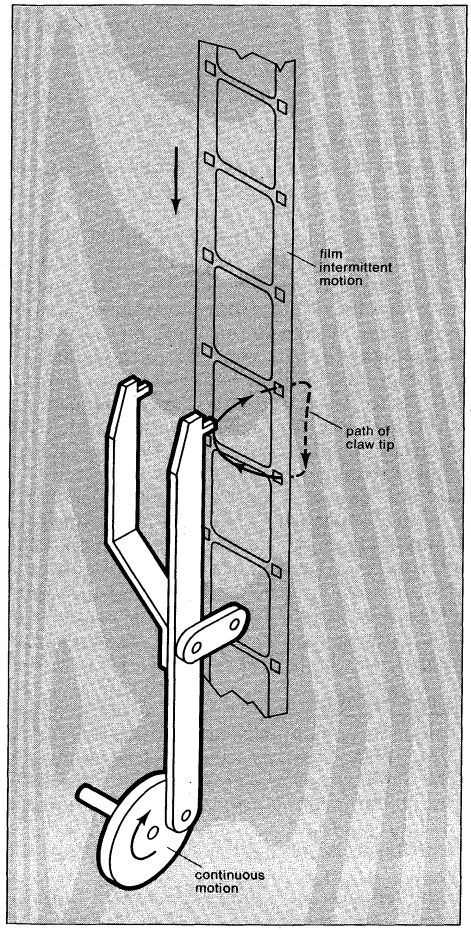


Figure 3: One type of the claw pull-down intermittent mechanism.

The intermittent was connected to a shutter. For projectors, the shutter was designed with two or three blades so that each picture frame was exposed two or three times and blanked out between these exposures for an interval equal to that required for the film motion from one frame to the next. This multiple exposure, with some light loss, reduced the flicker on the screen and made it possible to drop the frame rate from Edison's 46 to 15–18 frames per second. Since there was no flicker problem in the camera, the single-blade shutter was used there.

Growth of popular interest

Avid popular interest stimulated exploitation. The first naive titles, *The Baby's Meal, A Game of Cards, Street Dancers*, were followed by the more sophisticated *Cupid's Barometer, The Course of True Love*, and *Modern Brigandage*. News events appeared, such as *Moscow: The Crowning of Nicholas II; Stockholm: The King Arrives at the Exposition* (1896); *The Dreyfus Affair; London: The Crowning of Edward VII* (1902); and *Wall Street* (1896). Georges Méliès, in France, produced fantasy stories of the Jules Verne type, such as *The Voyage to the Moon* (1902); *Gulliver's Travels* (1902); *The Arabian Nights Palace* (1905); and *The Conquest of the Pole* (1912). Other narrative films included E.S. Porter's *Life of an American Fireman* (1903), and the very famous *Great Train Robbery* (1903).

When the art of projection was established, the importance of a bright screen picture was appreciated. Projectors used an arc lamp, at the time the brightest known light source. The machines ushered in the era of the nickelodeon (from Odeon, a theatre of ancient Athens and, much later, modern Paris, and the U.S. nickel, or five cents). Admission soon rose to $1.50 for *Quo Vadis* in 1913, as the auditorium grew from a vacant shop to a new theatre built for motion pictures.

INTRODUCTION OF SOUND

The popularity of the motion picture inspired many inventors to seek a method of reproducing accompanying sound. Two processes were involved: recording and reproducing. Further, the sound reproduction had to be presented in an auditorium and had to be quite good. This could not be achieved without a good amplifier of electrical signals. The invention of the Audion (triode electron tube) by the United States inventor Lee De Forest in 1907 made such an amplifier feasible and made possible an undistorted sound of sufficient loudness.

Next came the problem of synchronization of the sound with the picture. A major difficulty turned out to be the securing of constant speed in both the recorder and reproducer. Many ingenious ideas were tried. In 1918 in Germany, the use of a modulated glow lamp in photographically recording sound and a photocell for reproduction were studied. In Denmark in 1923, an oscillograph light modulator and selenium-cell reproducer were developed. De Forest tried a gas-filled glow discharge operated by a telephone transmitter to record a synchronized sound track on the film. For loudspeakers he experimented with a variety of devices but finally chose the speaker with horn. The operating signal was obtained from a light shining through the film soundtrack and detected by a light-sensitive device (photocell). These were used in a system called Phonofilm, which was tried experimentally in a number of theatres. In 1927 the Fox Film Corporation utilized some of these principles in the showing of Fox Movietone News.

Meanwhile, the Western Electric Company laboratories in the United States had been making extensive studies on the nature of speech and other sounds and on techniques for recording and reproducing such sounds. They experimented with recording on a phonograph disk and developed a 16-inch disk rotated at 33⅓ revolutions per minute; they improved loudspeakers, introduced the moving-coil type of speaker, and generally improved the entire electronic amplification system. The Warner Brothers movie studio became interested in all these de-

velopments and formed the Vitaphone Corporation to market the complete system. The first picture released was *Don Juan* (with music but no speech), in August 1926, followed by *The Jazz Singer*, featuring Al Jolson, with speech, in the fall of 1927. These two films introduced the era of sound pictures.

The first sound pictures

Further improvements were made in the disk system, but with time it became clear that recording on the film itself would be much more convenient. A serious problem was the distortion of the signal introduced by the glow lamp when recording the sound track on film. The Western Electric Manufacturing Company devised a mechanical light valve, in which the output from the amplifier causes a double-ribbon valve system to open and close by magnetic means. The light valve is arranged to modulate the light beam coming from a fixed lamp; the beam is projected optically onto the edge of the film to record the sound track. In 1930 the film recording supplanted the disk in the Warner Brothers Vitaphone system, and all other motion-picture makers adopted it.

There are two possibilities in the way the track can be recorded. The valve ribbons can be aligned either horizontally or vertically. In the first case, the optical track on the film will appear to be of varying density according to the modulation. In the second case, the densities will always be the same (black-and-white), but the track will vary in width. The two arrangements each have their advantages and disadvantages, and both have been used successfully.

INTRODUCTION OF COLOUR

The use of colour in photography and motion pictures is based on the fact that any arbitrarily selected colour can be produced by a mixture of red, green, and blue. In 1869 Ducos du Hauron outlined the principles of the two basic methods of colour photography, the "additive" and the "subtractive" methods. White light is composed of all the spectral colours, in equal intensities. It can also be composed of the red, green, and blue components, also in equal intensities. Light other than white can be composed of various proportions of these primary colours; for example, cyan can be composed of blue and green. In an additive system the colours are formed by the addition of specific amounts of light of the three primary colours. In practice, three pictures are taken behind a red, green, and blue filter, respectively; these are made into black-and-white slides, and each is projected by a separate projector through the appropriate filter, thus producing three images superimposed on a screen. Each spot on the picture is made up of added components of the red, green, and blue light in the proper proportion to give the colour of the original scene.

In a subtractive system the undesired components are subtracted from white to obtain the desired colour. By subtracting, say, the red from white light, the residual colour of cyan is obtained. By subtracting both red and green, the remainder of the spectrum, or blue, results. Subtracting all three colours yields black. In practice, the red-positive slide is converted into a "minus red" dye slide, so that the dark areas take away all red colour (and are cyan). Similarly, the green slide turns into a "minus green" slide, and the blue into a "minus blue" slide. The three slides are bound together into one so that each slide in turn subtracts its share from what has come through the preceding ones. When the composite slide is projected, it reproduces the colour view of the original scene. This is essentially what a colourist does in painting a slide by hand with transparent colours; transparent blue paint, for example, is simply a pigment that absorbs the red and green of the spectrum and lets the blue through.

Both systems have been used in colour motion pictures. Projection is simpler for the subtractive system because of the need for only one projector. Modifications in the additive system, however, can be used. The three projections can be made in quick succession, with a rotating set of colour filters. For motion pictures, this requires three times the picture speed for the one projector; or the picture space may be subdivided into many very small areas, so small that they will not be distinguishable on the projection screen, with very small filters of three classes in front of them when making an exposure. The emulsion is then processed, and, keeping the right filters in front of the small picture elements, the complete picture is projected. The essential difficulty here is that the slide will be dark because each colour will use only one-third of the slide area.

In 1906 Charles Urban and G. Albert Smith in the United States initiated an additive system of colour cinematography, which they called Kinemacolor. It used the system of successive colour projections, but to avoid extreme frame speeds it used only a two-colour process, namely orange red and blue green. The frames were exposed at 32 per second (giving 16 complete pictures per second) behind rotating colour filters. The film was processed in black and white and projected through rotating colour filters. Because of the two-colour process, the colour rendering was only fair. Also, because the two colours were exposed in succession, they did not cover exactly the same image. Thus, a white horse's legs in motion appeared successively orange red and then blue green. The system had some vogue; its outstanding film was *The Durbar of Delhi* in 1911 in London.

Around 1910 Louis Dufay made an additive-colour film that used fine-ruled areas on the film for the three colour elements; it was not put on the market until 1930. Eastman Kodak in 1928 produced Kodacolor 16-mm motion-picture film. This was also an additive process and used film with lenticular embossing that separated the rays of light according to where on the objective lens the rays came through. Thus, by putting three filters at the objective, the light on the emulsion was effectively separated into three small colour areas. After processing the emulsion in black and white, the film was used in a similar optical system for the projector. Both this and the Dufay process involved some loss of resolution and light-transmission capability. The Kodak film enjoyed a certain degree of commercial success, the name Kodacolor was later applied to an entirely different process.

After early starts with two-colour systems, a three-colour subtractive process (Technicolor) was introduced in 1932. This used a beam-splitting optical cube, in combination with the camera-objective lens to expose three black-and-white films. The three films individually registered the red, green, and blue tones of the scene being photographed. After processing, three film surfaces (red, green, and blue) were formed with gelatin layers that would retain more or less dye. After the surfaces were dyed, the films were contact printed on the final film base, the dyed images being superimposed to form the colour picture. Early famous pictures made by the Technicolor process were *Becky Sharp* in 1935 and *Gone with the Wind* in 1939. This system enjoyed a wide and durable commercial success.

WIDE-SCREEN AND STEREOSCOPIC PICTURES

Many people have felt that, while vision at the extreme sides of the vision field does not usually contribute much information to the eyes, it does add substantially to the illusion of reality when it is present. In 1929 Grandeur films were presented using 70-mm instead of the standard 35-mm film to give a wider field of view. Other film producers also experimented with wide films that showed a distinct improvement in realism, but they were introduced during a period of rapid development in sound pictures and did not last commercially.

In 1952 a radical attack was made on wide-field projection in the form of the Cinerama, using three projectors giving a field of view of some 70° or 80° (versus a more usual 40° or so) for most of the audience. An outline of the auditorium arrangement used is shown in Figure 4. This gave a remarkable increase in the illusion of reality, especially with such exciting and spectacular subjects as a ride down a toboggan slide. There were technical problems, including the necessity of carrying three cameras bolted together at the correct angles on the toboggan or other carrier, synchronization of the three separate films, and matching of the image structure and brightness at the joining edges on the screen. In spite

Heightening the illusion of reality

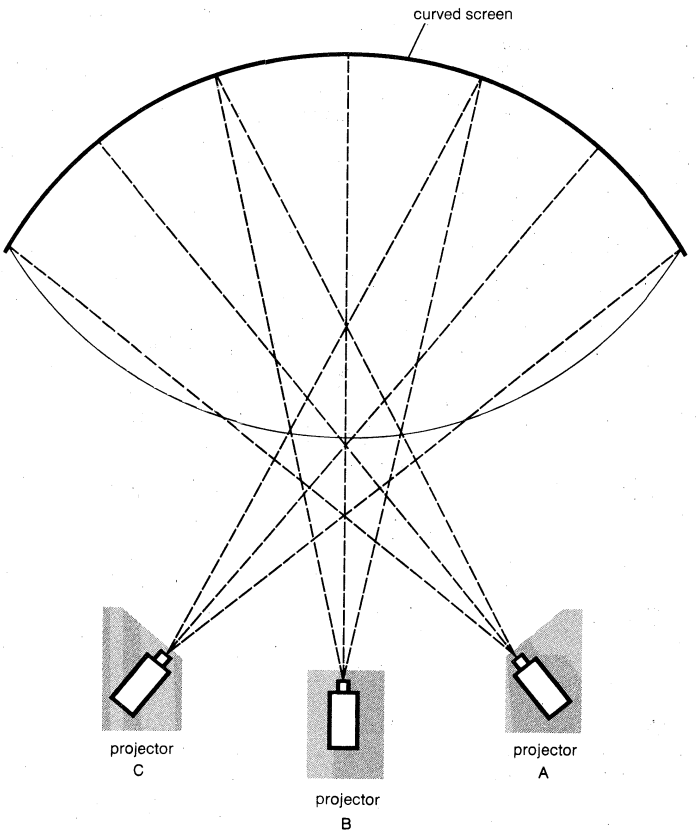


Figure 4: Cinerama auditorium.

of these problems and in spite of the difficulty of continually finding suitable subject matter, the presentations enjoyed a great vogue. A fourth film carried entirely separate sound tracks, which were connected to loudspeakers distributed throughout the auditorium.

A full circle of vision was presented by Walt Disney in a 360° projection at the Disneyland amusement park in 1955. At first, the projection involved 11 16-mm projectors and screens and, later, nine 35-mm projectors. The audience stood on a low platform in the middle. The result was extremely realistic. In one scene, showing the view from a cable car in San Francisco, the viewers were seen involuntarily to lean over on the curves, as if they were actually on the cable car. Such views, however, are not particularly adapted to general storytelling.

A less spectacular form of wide-screen system, introduced by CinemaScope in 1953, employed a special optical system to give a different magnification horizontally from that given vertically, thus changing the proportions of the picture on the 35-mm film to a relatively much wider one on the screen. This requires that the original picture be taken with a similar lens, used in reverse, to compress the picture on the 35-mm film. Such lenses are known as anamorphic optics.

In 1955 Todd-AO introduced a wider film (65-mm for the camera and projected on 70-mm film), with several sound tracks added (see Figure 6). It also used five sprocket holes per frame instead of the previous four. The wider film permitted projecting the picture on a wide screen with a single projector instead of the three used by Cinerama. In recent times wide-screen systems have tended to use a single wide film because of the complexities of the multiple films and screens of the Cinerama and similar systems.

Stereoscopic cinematography

Another project that has been introduced to improve the illusion of reality in motion pictures has been the development of stereoscopic, or three-dimensional, cinematography. Very simply, this consists in projecting (via the screen) to each eye of the viewer the different picture (that is, comparing one eye with the other) that it sees in normal vision.

One stereoscopic system uses two films, exposed side by side in a camera. The optical system is somewhat like that of a prism binocular and permits adjustment of the interocular (between-lenses) separation, to correct for differences in perspective as the scene is taken in close-up and somewhat longer shots, to make the final viewing seem natural. The system uses polarized light; that is, light that is vibrating in one specific plane. One projection lantern has a filter that polarizes its beam in one direction, and the other polarizes its beam at right angles to this. The projectors are adjusted so that the two images superimpose exactly on the screen. The observer in this system has to wear special viewing glasses that polarize light in the same relative directions as do the filters on the projectors. Thus, the left eye sees only the projection from the left lantern, and the right eye views only the right projection; together the two eyes see the projected image in the same way as they would have seen the original subject.

This process can give very realistic projections and has had some commercial success. A drawback is the need for the special glasses. Other systems have been devised to avoid this, but they are more complicated, have delicate adjustments, and are generally more expensive. Another problem is that only relatively few subjects presented in motion pictures gain substantially from the special depth perception given by the stereoscopic effect.

AMATEUR MOTION PICTURES

The public acceptance of motion pictures became so great in the 1920s that impetus arose to adapt the art to amateur use. One step was to devise means of reducing the film cost. A reduction in the size of the film was one approach. Various sizes, ranging in width from 28-mm to 9.5-mm, were explored in both the United States and Europe. Another approach was to study the overall process that required making and processing both the original negative and a positive print.

Reversal processing

This last gave rise to an exploration of reversal processing to obtain a positive rather than a negative from the original film. There are a number of processes that

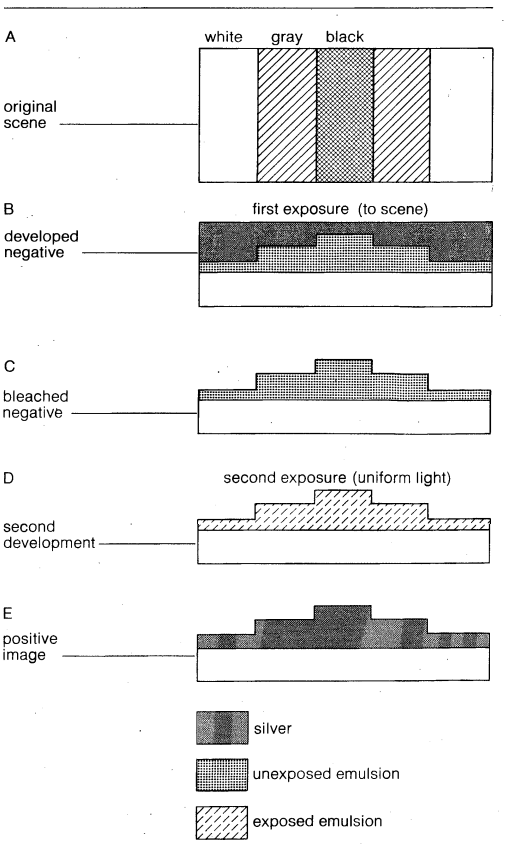


Figure 5: Principles of reversal processing of black-and-white film (see text).

can be used, but the one that has become almost universal is that developed in 1923 and outlined in Figure 5. An original scene (A) is represented by bars of various tones. It is developed as a negative, at B, in the usual fashion. Then the developed silver image is bleached away, leaving the unexposed silver halide, as at C. This is given a second and uniform exposure at D, which develops as a positive silver image at E. The first point that was noted in using this process was that it seriously reduced the latitude in exposure; *i.e.*, the exposure had to be correct for the particular lighting condition. This was especially unfortunate for inexperienced photographers. After some experimentation it was found that an appropriate adjustment of the second exposure improved the latitude. It was further found that this could be regulated automatically by exploring the image at D with infrared light, to which the emulsion was not sensitive. The total infrared light transmitted was detected by an infrared photocell. The more infrared that came through, the more light the second exposure was given. The reversal process not only saves the cost of an extra positive but also gives less graininess than when the positive is printed separately.

Amateur motion-picture photography soon became very popular, and other firms introduced their products. In 1923 Bell and Howell in the United States brought out equipment to use 16-mm film. In the same year Pathé introduced a 9.5-mm film camera that, along with the corresponding projector, became very popular in France. The 16-mm reversal film was introduced subsequently by a number of other manufacturers.

The Kodacolor film mentioned earlier was an improvement introduced in 16-mm amateur film in 1930. The later Kodachrome colour film was also introduced in 16-mm amateur film.

In 1929 a sound track was proposed for 16-mm film by omitting the perforations on one side, and a projector for it introduced on the market by Radio Corporation of America (RCA) in 1930. Narrower film was explored, and 8-mm film with cameras and projectors for it were put on the market by Eastman Kodak in 1932.

II. Professional motion-picture photography

CAMERAS

Principles of operation. The principles of operation of modern professional motion-picture cameras are much the same as those of earlier times, although the mechanisms have been refined. A film is exposed behind a lens and is moved intermittently, with a shutter to stop the light while the film is moving. In the process, the film is unrolled from a supply reel, through the intermittent to the gate where the exposure takes place, and then on to the take-up reel. Optionally, the film, after exposure, is passed through a sound-recording mechanism between the intermittent and the take-up reel. Currently, however, it has often been found more practical to record the sound on an entirely separate film or tape that is kept synchronized in some fashion with the picture film. The sound-recording stage is omitted in such cases.

Principal parts. *Body.* The principal difference between current cameras and the earlier ones is in the matter of sturdiness and rigidity. It was found that in order to maintain a high quality of picture it was necessary to eliminate vibration, so that the picture would remain sharp and also the film would not move about in the film frame. In some designs, the film chambers are separate from and attached to the camera body; in others, they are integral with the body.

Lens. Lenses have gone through a continuous evolution in the last half century, for both still and motion-picture photography (see OPTICAL ENGINEERING). The two major objectives have been to focus properly all the colours of the image at the film plane (*i.e.*, to make the lens achromatic) and to focus portions of a beam coming from different portions of the lens, the centre or the edges, at the same point on the film (*i.e.*, anastigmatic). Both objectives require solution for as large a lens opening as possible, in order to capture maximum light for the exposure, and for as wide a field of view as will be needed in the use of the lens. In order to solve these problems, lenses have been made with more and more components. Also, more types of glass have been discovered and developed, to give better achromatic performance. It was found, about 1939, that a special coating of the glass-to-air surface of a lens component could greatly diminish reflections from this surface without affecting other properties of the lens. The use of such coatings improved image contrast by reducing the stray rays that were produced by reflections in a multiple-component lens. Coatings also reduce loss of light by reflection in the desired rays. Coating developments have permitted making lenses with many more components than had previously been possible.

Achromatic and anastigmatic lenses

Long experience with both motion-picture and still cameras has shown the need for a variety of focal lengths (ranging from ultrawide angle to telephoto) to photograph scenes under the best conditions. To make changing focal lengths more convenient, the lenses have been mounted on a turret, so that one out of a set of three lenses may be quickly selected. For motion pictures this would mean an interruption in the action depicted. A continuous change would be more desirable.

When two lenses are used in a tandem combination, the focal length of the combination varies according to the separation between the two components. For example, when two thin converging lenses are mounted close together, the combined focal length is shorter than when they are separated a certain distance. Thus, the focal length of the combination can be continuously varied over a range merely by changing the separation.

This observation has led to the conception of camera lenses of variable focal length in which the variation is obtained by moving one or more elements. One simple design consists of two fixed convex (converging) lenses of unequal power with a movable concave (diverging) lens between them. When the central concave lens is located close to the front convex element, the combination focal length can be shorter (and the image therefore smaller) than when it is located close to the rear convex element. The design can be made such that, with the two convex elements remaining fixed, a distant view can remain almost in focus on the film as the middle element is moved. Exact focus for this arrangement, however, could not be attained. Thus, for lenses of this design, a cam device has in the past been provided to move the front convex element a short distance as the middle element is moved over its range, to keep the focus exact. This kind of lens has come to be called a zoom lens.

The zoom lens

By increasing the number of elements, the focus can be kept exact without the need of a correcting cam. Other improvements include increasing the range of focal lengths covered, increasing the effective lens aperture, increasing the angular field of view seen by the film, and improving the colour correction with radically new glass materials.

For a long time the change in focus was carried out manually. More recently, the use of an electric motor drive has allowed a smoother change, with less distraction to the cameraman.

The resolving power (in general, the ability to record fine detail) of the zoom lens has not in the past been equal to that obtainable from lenses of fixed focal length, and for this reason they were little used in professional feature motion pictures, being reserved rather for news and amateur pictures. The performance is improving, however, and some studio cameras now use zoom lenses.

Transport system. The general principles utilized in the film transport system have remained much the same over recent years, at least for the 35-mm film. The films are usually preloaded in lighttight reel cases, with an exposed loop between the supply and take-up reels. This loop is quickly fitted into the camera mechanism when loading.

The intermittent is usually a claw-type mechanism, somewhat more sophisticated, however, than that in Figure 3. Some cameras are equipped with pin-registering mechanisms, which hold the film firmly in place in the exposure gate, with the pins engaging sprocket holes.

In the early days of sound films, the noise made by the

intermittent and other moving parts in the camera was loud enough to interfere with the sound picked up by the microphone. Cameras were sheathed ("blimped") with outer, separate sound-absorbing materials. The sound insulation is now usually self-contained in the camera.

The film drive

Before the introduction of sound, the film and intermittent were driven by a large crank operated by the cameraman. With sound, considerably more uniformity in the speed of the film drive became necessary. For this and other reasons, the film drive in modern cameras is provided by an accurately controlled electric motor.

Shutter. This component keeps light from striking the film while it is moving from one frame to the next. A variable shutter opening can also be used to reduce exposure when it is necessary or desirable to do this without reducing the lens aperture. The shutter is rotary and is synchronized with the intermittent. In some cases (see below *Viewfinding and focussing*) the interval when the shutter is closed is used for viewfinding through the same lens that produces the image on the film. Here the shutter cuts the incoming light beam at an angle, and the front face of the blade has a mirror surface to deflect the light beam to one side toward the viewfinder.

Viewfinding and focussing. Viewfinding has always been a problem for both motion-picture and still cameras. If a separate viewfinder is attached to the camera, its line of sight to the scene is different from that of the taking lens, so that the picture seen is not exactly the same as the picture photographed. One solution to this problem consists of mounting the viewfinder in such a way that its angle can be adjusted so that the finder and the taking lens point at exactly the centre of the subject. Still another method is to use a partially reflecting mirror (beam splitter) in the path of the beam of light that is used to take the picture. This is objectionable because it takes away part of the light that would be used for the exposure, and, in addition, there is some possibility that the light beam containing the image to be photographed will be impaired.

Another possibility is to use a mirror to divert the taking light beam while the film is moving and the shutter is covering the film. This arrangement is not wholly free from objections. The arrangement opens a return path for light from the viewer's eyepiece to reach the film. The eyepiece must fit snugly around the eye while the viewfinder is in use, and the finder must be closed completely while it is not in use. In addition, since the camera shutter is closed only once per frame, the image will be subject to a distinct flicker, to which the cameraman must adjust himself.

Focussing has also been a perennial problem for the motion-picture camera. On the camera the position of the lens is precisely indicated on a calibrated scale. The actor's location on the set was formerly marked on the floor and the exact distance to the camera measured with a tape. The actor moved to previously marked places, and an assistant to the cameraman, called a focus puller, or follow-focus assistant, kept the lens in adjustment. Various electrical devices have now been introduced for remote adjustment by the assistant.

Distance measurement

Where a through-the-lens finder is used, focussing can be done directly, using the viewfinder image. Also, experienced cameramen can estimate distances quite closely.

Miscellaneous features. It is usual to generate some kind of signal in synchronism with the intermittent when an auxiliary, magnetic-tape sound recorder is used, so that the sound record can later be synchronized exactly with the picture. Where a separate film is used for the magnetic sound record, the sprocket holes are electronically locked with those in the picture film, maintaining the synchronization.

The camera is often supplied with electric motors to perform miscellaneous functions, such as to provide smooth rotation (panning) of the camera or to change the magnification in a zoom lens (or change lenses in a turret).

The camera is normally provided with footage indicators to indicate the amount of film left unexposed and frame counters where it is desired to superpose a second exposure. There can also be an "inching knob" to reposition the film to a given frame for multiple exposures. When the camera is used at a speed different from standard, a tachometer may be provided to indicate the actual speed.

The cameras that have so far been described are for the standard 35-mm film. Cameras for 65-mm film are generally quite similar, though heavier. The quality of 16-mm and 8-mm film has improved so much that a good deal of professional work is now done in these sizes. The 16-mm professional camera may differ from the 35-mm in the form of its case, in its use of a spring-operated film drive, and in its method of film loading, as a result of its development from a former amateur camera. On the other hand it may be a smaller version and have the same features as the 35-mm model by the same manufacturer.

Stereoscopic and Cinerama cameras. Three-dimensional cameras differ from standard cameras chiefly in being arranged to expose two films simultaneously side by side, using a twin optical system with prism arrangement.

CinemaScope cameras use a special 35-mm film with somewhat smaller sprocket holes and a lens or attachment that makes the reduction ratio from scene to film different in the horizontal dimension from what it is in the vertical dimension.

Camera supports. The camera must be mounted on a substantial support. In its simplest form this is a heavy tripod structure, with sturdy but smooth-moving adjustments and casters, so that the exact desired position can be quickly reached. Often a heavy dolly, holding both the

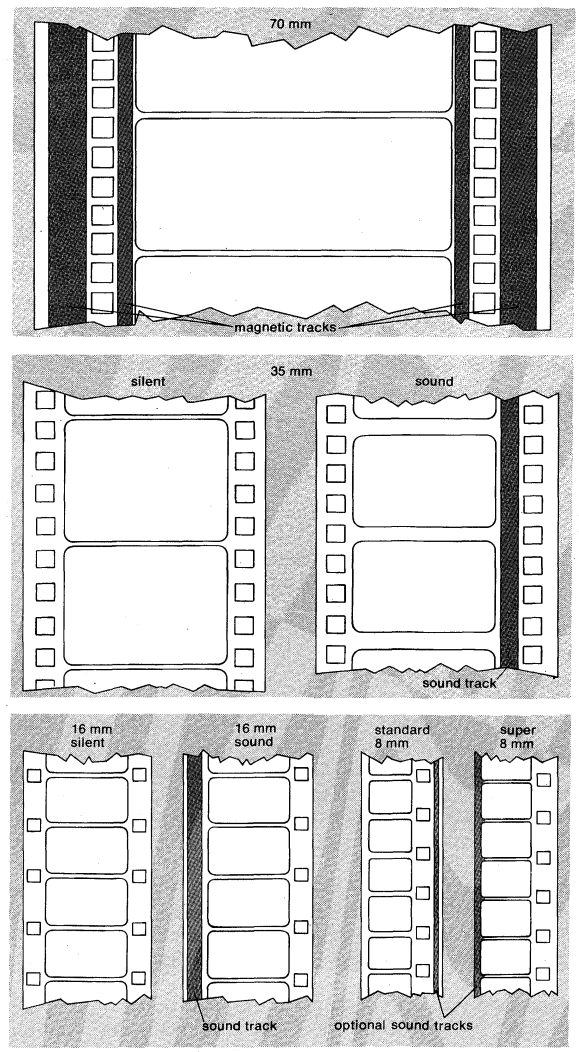


Figure 6: *Film formats.*
The sizes shown are approximate; the overall width of 70-mm film is 70 mm, etc.

camera and a seated cameraman, is used. This can be pushed or driven around the set. When shots from elevated positions are to be used, both camera and cameraman are carried on the end of a crane, also on a dolly. In some cases the assemblage is smoothly driven to follow the action being pictured, such as movement along a street.

FILM

Format and support. The more frequently encountered film formats are illustrated in Figure 6. The silent 35-mm film, however, is seldom used. There are some minor differences in the shape of the sprocket holes in 35-mm film between negative and positive film. The first 8-mm film was made by using 16-mm film, punched with twice as many sprocket holes of the same size and shape. One side, to the middle line, was exposed in one direction. The supply and take-up reels were then interchanged in the camera, and the other side was exposed in the other direction. After processing, the film was split into two strips, which were spliced into one. Experience with this format indicated that the sprocket-hole structure was unnecessarily sturdy, and a so-called Super-8 film was designed with the idea of reducing the sprocket-hole size and employing the space thus made available for a larger picture area. Both formats are in current use, but the Super-8 is becoming predominant.

Originally, the film base was some form of celluloid or cellulose nitrate. This material is highly inflammable, and extensive precautions were required in projection rooms to avoid film ignition because of the proximity of the projector arc lamp to the film.

In 1923, when 16-millimetre amateur film was introduced, cellulose acetate (or safety film), much less inflammable than the nitrate, was used. It was not considered desirable to adopt it for professional 35-millimetre film, largely because it was inferior in strength and dimensional stability. By the late 1930s an improved cellulose acetate safety film was introduced, and by the early 1950s it had generally replaced the nitrate film.

Black-and-white emulsions. The film base is coated with a light-sensitive layer of silver halide emulsion. Emulsion manufacture is quite complicated and delicate. The earlier emulsions were most sensitive to violet and blue light, as shown schematically in Figure 7, curve *a*.

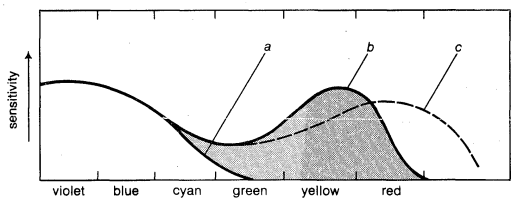


Figure 7: Sensitization of photographic emulsions (see text).

Toward the cyan and green, sensitivity drops rapidly. Such an emulsion is called natural, or ordinary. The result of such a characteristic is that in a natural scene reds and yellows appear black in the positive, and green appears too dark.

As early as 1873 it was found that dyes introduced into the emulsion could increase the sensitivity in the yellow and green, Figure 7, curve *b*. The change increased the natural appearance of the reproduced picture, and the emulsion was called orthochromatic. Later (1904) dyes were found to prolong the sensitivity into the red, and this emulsion is called panchromatic, Figure 7, curve *c*. The dates are fairly early for motion-picture application, but the development had importance in the general technology.

The overall sensitivity for picture taking has been increased greatly, from below about ten ASA before 1930 to several hundred and even several thousand. The ASA (American Standards Association) scale is an arbitrary rating of film speed; that is, the sensitivity of the film to light. If everything else is kept constant, the required exposure time is inversely proportional to the ASA rating.

Negative films designed for original picture exposure are usually faster (*i.e.*, have higher ASA ratings) than those for prints and are apt to be somewhat coarser grained.

Colour emulsions. The modern subtractive colour, which is so easy for the cinematographer to use and which gives such generally satisfying results, is sometimes thought to be one of the minor miracles of the 20th century. First introduced in 16-millimetre amateur films, it is now used in all sizes of motion-picture film and has largely replaced black and white.

The development of modern colour film

There are two major steps involved in making a dye image on motion-picture film. The first is to convert the negative silver image that is obtained from a normally exposed film into a positive dye image. The clue to how this can be done came from experience with a developer known as pyro (pyrogallol), once very popular with still photographers. A negative developed with pyro developer has not only a silver image but also a brown stain. Study of the process showed that the stain was caused by oxidation products given off locally by the developer in the development process. A substance in the developer reacts with these oxidation products to give an insoluble brown dye. The substance is called a dye coupler. Since the dye is not soluble, it does not wash off in the subsequent film treatment.

This suggested the possibility of bleaching to take away the silver image, leaving the dye image on the film. The first step was to find a developer and dye couplers that would produce the three dye colours that give a faithful three-colour picture rendering. The second step was to carry out the process in the film coating with three separate colours and keep them separate, all the way from exposure to the final three-colour image on the completed film.

The first portion of this second step is carried out by obtaining three emulsions that can be laid on top of one another and are sensitive, respectively, to the red, green, and blue of the exposing image without interfering with each other and that give corresponding silver layers that similarly do not interfere with each other.

It has been observed above that normal silver halide photographic emulsion is particularly sensitive to blue light and that one of the early problems was to obtain a more natural pictorial rendering by extending the sensitivity of the emulsion to green and finally to red light. The problem was solved by inserting appropriate dyes in the emulsion. The dye adds a peak of increased sensitivity, respectively, to green and red light, as in Figure 7. The triple-layer film then consists of, on the top, an ordinary blue-sensitive emulsion; below this, a yellow filter to cut off blue light; next below this, an emulsion with a sensitivity peak in the green, with the yellow filter cutting off blue sensitivity; and finally an emulsion with a peak sensitivity in the red, a valley in the green, and blue sensitivity cut off by the yellow filter. The sensitization can be chosen to locate and enhance the sensitivity peaks.

Thus, the blue layer responds to the blue light in the original, the green layer to the green light, and the red layer to the red light. These can be given a first development together, so that the individual responses will be indicated as silver deposits in the respective layers. The developer used is one which leaves no dye-coupler stains.

In what is called the nonsubstantive subsequent process, the dye couplers are introduced in a second development. Each colour layer is treated separately. Uniform red light is applied (from the bottom up) to expose the undeveloped silver halide in the red layer. It has no effect on the other layers because of their insensitivity to red. The film is processed with a developer containing a minus-red (or cyan) dye coupler. This leaves a silver and minus-red dye deposit wherever there was newly exposed silver halide in the red layer. Similarly, the blue layer, newly exposed with blue light from above and processed with a developer containing a minus-blue (or yellow) dye coupler, leaves a silver and minus-blue dye deposit wherever there was newly exposed silver halide in the blue layer. In the remaining green layer, a white-light exposure and development with a minus-green (or magenta) dye coupler converts the residual silver halide into a silver and minus-green dye deposit.

All the silver deposits and the yellow filter are finally bleached out. The remaining dye deposits serve to subtract from white light, in the manner that was described earlier, the correct part of the spectrum to leave the colour of the initial exposing light. For example, where this light was red, the final dyes absorb blue and green. Of the spectrum, this therefore leaves red light to go through the film.

The substantive process

In a modification called the substantive process, the appropriate dye couplers are suitably embedded in the emulsion in the appropriate colour layers to prevent their moving about during processing and contaminating the colours (an important problem). It is then possible to carry out the second exposure and development on all three layers in a single step with white light and with only one developer.

Examples of the nonsubstantive type of film are Kodachrome and Fujicolor. Examples of the substantive type are Ektachrome, Anscochrome, and Ferraniacolor.

The processes described can be modified to give a negative (or complementary-coloured) image by omitting the reversal step. This can then be printed to give a positive image on another film. This process can be useful when it is desired to make a large number of prints from a single matrix. The matrix can also be made as an internegative from an original positive film.

In professional commercial motion pictures, final exhibition prints can be made by a variety of combinations of these several processes, according to convenience, quality, and economic requirements. The combinations can include films in a variety of sizes, so that 16-mm projection prints can be made from 35-mm originals, or 35-mm prints can be made from 16-mm originals (for example, news films), and other sizes (such as 8-mm and 70-mm) may be involved in the operations.

LIGHTING

Equipment and techniques. One of the earliest problems encountered in motion-picture photography was the need for sufficient light for exposure. In contrast to still photography, the exposure duration was limited to something like half the time between one frame and the next. With the low-speed films and the relatively small-aperture objectives then available, a good deal of light was needed. At first, only sunlight and skylight were used. Artificial lighting had to be very strong to be effective and involved several problems. Incandescent-lamp lighting is rich in red rays but poor in blue, the reverse of what was desirable for the natural films shown in Figure 7, curve *a*. As with projection, in which a great deal of light was also needed, it did not take long to apply the brightest light source then available, the electric arc lamp. The term klieg light (arc made by Kliegl) became widely recognized in motion-picture production. Specialists developed a systematic technique, the general principles of which continue to this day and extend in their application from studio to location shots or other outdoor shots where preparation is possible.

First, a general level of illumination is established throughout the set to approximate daylight conditions. Actual outdoor daylight requires only a touching up procedure. Second, concentrated beams of light are shot onto people or areas that need to be clearly set forth in the picture. The shapes of the beams and their intensities depend upon the nature and shape of the area to be filled in. In the case of a travelling shot, in which the camera follows the action along a defined path, the fill-in light may have to be turned on and off as needed by the photographer. As a result of these different needs, a great variety of designs and intensities of lighting equipment have been produced.

Colour requires that much greater care be taken with the colour of the light directed onto the set. It is, for instance, much more of a problem to fill in general arc-light illumination with incandescent lights. These latter tend to be yellower and need filtering or a very cautiously selected location to avoid giving an unnatural appearance to the set on the colour film.

New tungsten-halogen lamps have been developed, using a quartz envelope. The halogen compound is included inside the envelope, and its purpose is to combine with the tungsten evaporated from the hot filament. This forms a compound that is electrically attracted back to the tungsten filament. It thus prevents the evaporated tungsten from condensing on the envelope and darkening it, an effect that reduces the light output of ordinary gas-filled tungsten lamps. The return of the tungsten to the filament means that the incandescent lamp can be run at a higher filament temperature and still have a long life. These lamps are now sometimes provided with a special multilayer filter to give a bluish light that approaches arc light.

Tungsten-halogen lamps

The lamps usually are now also provided with lenses to concentrate the light beam somewhat and prevent excessive light loss around the sides. They can also, when suitably focussed, give a relatively sharp beam. In a studio there are racks above and stands on the floor on which lamps can be mounted. The primary considerations are that they direct the light where it is wanted and that they not impede floor traffic around the set.

Light measurement. The measurement of the light is a development that took some time to come into the photographic art. It was not until the 1930s that a photoelectric device was developed that could be used easily to measure light intensity. Since that time there has emerged a body of knowledge on how the intensity is to be most effectively measured. This has been especially desirable since the advent of colour, which is more sensitive to exposure than is black-and-white film.

There are three approaches to this measurement: incident light, reflected light, and spot.

Incident light. Readings of this type measure the overall illumination on the scene. This is not always practical, as when the subject being photographed is distant, such as a remote building. These measurements require some adjustment for the general reflectance of the subject being photographed.

The meter for this type of measurement can be the usual portable meter used by amateurs, but it may be desirable to have the photocell covered by a translucent dome so that light coming from the side is included in the measurement. A more sophisticated meter is occasionally used with a photocell separate from a more sensitive indicating meter.

Reflected light. Readings by this kind of meter measure the average light coming toward the camera from the scene being photographed. This works well for average subjects but gives wrong exposures if the background contains either many bright areas, as in a beach scene, or very dark areas, as in front of a dark building. In such cases the photocell must be held not at the camera but very close to the subject of interest, to eliminate the effect of the background. This is also the case when the scene contains a good deal of backlight.

The meter here can also be the simple, portable, amateur exposure meter, but it is desirable that it have a relatively narrow acceptance angle for the light, to reduce the effect of the background. The procedure is the simplest and most popular with amateurs.

Spot measurement. These readings measure the light coming toward the camera from selected spots in the subject being photographed. The meter for this purpose has an optical system that covers measurement of a spot of about one degree. Such readings must be checked against the overall lighting on the set to assure that the distribution of light values fits the characteristics and range of the film being used.

Colour. With colour photography it has become necessary to measure not only the intensity of the light on the scene but also its colour. The meter used in this case is called a colour-temperature meter. A standardized and fairly accurate measurement can be obtained by taking the ratio of light intensities in the red and in the blue, which is what a colour-temperature meter does. More recently it has been found desirable on such meters to sample the intensities of three, rather than only two, colours to obtain the best results on colour film with artificial light.

FILM PROCESSING AND PRINTING

Processing. In the early days of motion pictures, films were processed by winding on flat racks and then dipping in tanks of solution. As films became longer, such methods proved to be too cumbersome. It was recognized that the processing system should have the following characteristics: it must run continuously; it must be lighttight and yet capable of being loaded in daylight; and it must be as compact as possible to provide a minimum air surface for the processing solutions. A general form evolved that is still in use (see Figure 8).

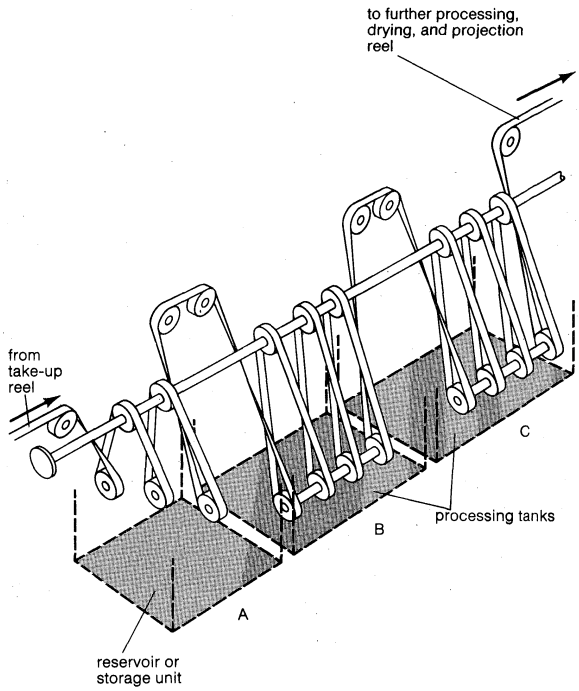


Figure 8: Basic steps in the continuous method of film processing.

For continuous operation the film must be passed continuously through the solutions and folded back over rollers that do not touch the emulsion surface. It must be handled very carefully, as the impregnation with solution weakens the support, and the sprocket holes should preferably not be engaged. Drive should, therefore, be accomplished by a light friction force at the edges.

Splicing on a fresh film without affecting the motion of the part of the film being processed is handled by using a storage unit or reservoir, as at A in Figure 8. This reservoir has a variable capacity so that the output end can be giving out film while the input end is stationary as the new film is spliced. Lighttight gates prevent all but a short length of film being light-struck at the very beginning or end of the film (and leaders may be used). The take-up-reel case is fastened in a lighttight way to the storage unit so that after splicing, the film is unreeled into the storage and processing units until the other end is reached, ready for splicing to the next film after changing cases.

Passing the film through solution

Many tank shapes have been tried. The long vertical tanks (as at B and C in Figure 8) provide for several passes of the film through each tank. The spools are designed so as to hold the film at the edges by friction. There are a number of types of drive, but all function gently to avoid strain. Sometimes the spools have multi-stepped edges to accommodate various film widths. The lower spools (or "diabolos") are more or less free but guided in a loose fashion so that they will not jam or tangle. The long vertical tanks give a minimum of air surface to the solution. The motion of the film through the liquid can be sufficient for proper contact of the film with the solutions, but sometimes submerged sprays with small jets of fine nitrogen bubbles are provided to increase the agitation.

Processing solutions age in continuous operation. They must be constantly checked by chemical or other analysis, and fresh replenishing solutions must be added as necessary. Lengths of processed film are examined to make sure that the processing is continuing normally.

The last receptacle in the processing sequence is a drying oven. There are several designs, some of which generally resemble the tank but without solution and are provided with heating elements. This receptacle does not need to be lighttight.

A variety of methods are used to make sure that processing is dust free.

The processing steps for the many different types of film are similar in principle, though there are variations in specific solutions and treatments. Reversal film requires a second exposure step, accomplished between trips through solution tanks. The introduction of sound track requires a much closer control of the development process to maintain the sound quality, for both variable density and variable area tracks. Colour film materially increases the number of processing operations and extra exposure steps and lengthens the size of the machine. It also further increases the need for close control of the processing operations. In cases in which the sound track is carried by the silver deposit instead of the colour dyes, special processing for the track is required.

Film prints to be used for projection are given a coat of wax over the sprocket-hole areas. This eases the film passage between the pressure plates at the projection aperture.

Processing solutions have been devised in viscous form, which are more rapid in operation than the liquid and permit the use of more compact equipment. These have been employed for special applications but have not been used generally.

Printing. The preparation of exhibition prints for sizable numbers of theatres is a large-scale operation. Printing takes a number of different forms.

In contact printing, the master film (or negative) is pressed against the raw stock; this combination is exposed to light on the master filmslide.

In optical printing, the master film is projected through a lens to expose the raw stock.

In continuous printing, the master film and the raw stock both run continuously. Continuous printing is usually contact printing but could be optical, through a projected slit.

In intermittent, or step-by-step, printing, each frame of the master film is exposed as a whole to a corresponding frame space on the raw film.

It is possible to print from one size master film to another size raw stock, such as 35-mm to 16-mm, or vice versa. In such cases the printing must, of course, be optical, and in the examples cited must be intermittent if there is a sound track. This is because, as indicated in Figure 6, 35-mm sound film has a spacing between frames which 16-mm does not. The sound track must be printed separately.

SOUND-RECORDING TECHNIQUES

The art of sound recording for motion pictures has developed dramatically. Apart from a continuous improvement in sound quality, two particular developments have markedly affected the technology and practices. The first was the discovery that there was no need to record the sound on the same film and in the same camera as the picture. On many occasions several pictures are taken from different points of view of the same scene, which has only one sound record. Or several sound records may be taken. It is far more flexible for subsequent handling (starting with the developing, which can be specialized) to take the sound recording or recordings in a separate box under the care of a specialist. In later editing processes, the editor can cut in such segments of the picture film as best fit the development of the story, but they are all anchored to the same time sequence of the sound tracks. For this it is necessary to have absolute synchronism between the various picture films and the sound recordings. When the last are on a sprocket-holed support, the

Separate sound recording

sprockets are all driven synchronously from a common power supply. A common starting point can be established at the beginning by a "clapper," or set of wooden jaws about a foot long, snapped together in the picture field. The instant of clacking then appears in both picture and sound tracks.

The second change in technology was the advent of magnetic sound recording. The big advantage of this technique is that it provides improved sound quality, thus permitting greater flexibility in editing procedures. When a sound track is copied, the copy is not as good as the original, and a copy from the copy is poorer yet, and so on as successive copies are made. The amount of degradation at each step, however, is less if the process gives a more faithful rendering in the first place. Thus, magnetic recording permits the use of successive copies in the editorial process while still giving a final copy of adequate quality for exhibition purposes. Also, magnetic tape or film does not need to be developed but can be used as soon as made and does not need to be protected from light. The result is that most of the present original sound recording for motion pictures is done magnetically. On the other hand, most of the 35-millimetre film projected in theatres carries a photographic sound track, as most theatre projectors in current use are equipped with optical sound reproducers, thus requiring the magnetic sound track to be converted to an optical sound track in the exhibition print.

Much of the equipment and procedures in motion-picture production is more or less the same whether the process used is optical or magnetic (see also SOUND RECORDING AND REPRODUCING). The microphone is an example. Microphones of many different types have been used for sound recording. These may differ in sound quality, in directional characteristics, in practical convenience of use, and in other characteristics. Conditions that may dictate the choice of a particular microphone include the presence of minor echoes from objects in the set or reproduction of speech in a small room, as distinct from that in a large hall. Painstaking adjustments are made by careful attention to the choice of microphones, by the arrangement of walls and furniture on the set, and by the exact positioning of the actors. For recording a conversation the preferred microphone is sensitive in a particular direction to reduce extraneous noises; it is moved toward each actor in turn. The boom used to hold the microphone in these cases resembles that used for small lights. Microphones can also be mounted on a variety of other stands. If necessary, a chest microphone that is hidden under clothing may be used.

Choice of microphones

When several microphones are used on the set, their outputs are often mixed and reproduced on a single film or tape. The mixing is done at a small console, and the operator must use a high degree of skill. Sometimes this is done in the later editing or dubbing process.

Magnetic recording. As far back as 1900, Valdemar Poulsen, a Danish electrical engineer, devised means of recording sound on a steel wire. Around 1928 the German firm Allgemeine Elektricitäts-Gesellschaft (AEG) experimented with a thin iron oxide layer deposited on paper or plastic tape. It was not, however, until well after the end of World War II that its advantages were recognized for sound motion pictures.

The system makes use of a magnetic iron ring with a wire winding around it. In the recorder this carries the electrical signal. The ring has two gaps in it, filled with shims of nonmagnetic material. The plastic tape with magnetic iron oxide coating is drawn past one of these gaps at a constant speed. The current in the winding causes a magnetization of the ring, which in turn leaves a permanent magnetization in the iron oxide coating. The wave shape of the permanent magnetization follows the wave shape of the signal current. The recording of fine details on the track depends upon the fineness of the nonmagnetic gap in the ring and on the speed of the tape.

In the replay condition the tape is passed over a similar ring with nonmagnetic gap, and the magnetization in the iron oxide coating induces a current in the winding. A measure of the quality of the reproduction is the fidelity with which the wave shape of the current follows the wave shape of the original current that was supplied to the recorder and the amount and nature of the noise that was added.

It has been found that, if a certain fixed amount of current (usually called "bias") is added to the recording signal, the quality of the reproduction is distinctly improved, both as to faithfulness in wave shape and as to noise. The amount of bias added is usually adjusted to a value giving the greatest improvement. With this improvement the quality of the reproduction obtainable in practice is higher than that usually provided by the use of the photographic sound track.

The tape base is available in two forms. One is with sprocket holes of the kind used for picture recording. Usually this is without emulsion, since no picture is carried. Occasionally, however, especially with the 16-mm format, it is desired to record the picture also, and for this the film carries both picture emulsion and magnetic track. Sometimes, if there is no picture emulsion, the base is made of different material from that used for picture film.

Tape bases

More compact recording apparatus is possible with a straight quarter-inch-wide tape having no sprocket holes. In this case, a second magnetic track is necessary to maintain synchronism with the camera.

Optical recording. Among new developments in optical sound recording is a technique using a fast-acting galvanometer rather than the double-ribbon light valve to control the amount of light reaching the sound track. The record on the film is a variable-width track.

The principal present use of this type of recorder is to make a master optical negative for final exhibition prints after all the editing and dubbing have been finished. From this optical negative the sound track can be printed photographically on the exhibition release prints. Sometimes the release prints are printed as positives directly from the final magnetic track, rather than from a negative made from this track. This is known as electronic, or electrical, printing, and it may have advantages from the standpoint of quality.

Dubbing. This is a term used to indicate a sound track that was added during an editing stage rather than recorded at the same time as the picture. It can include the addition of incidental background noise to add to the realism of a scene. It can also provide for the substitution of a voice recording made under more favourable conditions than those prevailing during picture taking and can be made before or after the picture is taken. In pictures that present singing of operatic quality, for example, the singing may be accompanied by facial expressions that are not attractive in close-up (which in an opera would not be noticeable to an audience). Such scenes can be prerecorded or postrecorded, due care being taken to assure exact synchronism between the sounds and the lip movements. In musical productions in general, a better recording of the music can be obtained in a separate session in a sound studio than on the picture set, with the result that filming is made with only a token orchestra. Sometimes an addition dubbed in can be only a commentary on the action in a picture. In all of these cases the sound and the picture recordings are merged in an extra editing session of the production film.

There is a further use of the dubbing procedure. It is often desired that a film made in one language be translated into another, for an entirely different audience. If the actor cannot speak the other language, it is possible to dub the speech in that language from another speaker. In such cases, if this is to be done credibly, it is necessary to make the speech in the other language fit the character and cadence of the original. If the actor's face is visible in the picture it is also necessary to fit the words of the translation so that the lip movements are not too disparate. Dubbing of this sort done skillfully can be remarkably credible.

Translation dubbing

SPECIAL EFFECTS

Perhaps the simplest special effect is the gradual dissolving of a scene by a slow reduction of contrast until it

disappears into either a uniform field or another scene. This is frequently carried out in a preparation of the final film negative. Another method involves "wipe" patterns, in which the scene grows from a small part of the frame until it encompasses the whole frame. This is accomplished by the use of a "travelling mask" on an auxiliary pair of films (one the negative of the other). On a first printing through the positive mask the first scene is exposed on the fresh film. On a second printing through the negative mask the new scene is printed on the unexposed part of the fresh film.

To print conveniently and simultaneously from several films requires a special-effects printer, which is usually a specially designed optical printer. The machine becomes relatively complicated when several films are involved in the special effects.

Another type of special effect is the projected background. If, for example, it is desired to photograph the actors before a remote background, such as a street in Paris or the Pyramids, an ordinary shot of the location, perhaps taken from stock, is projected in the studio onto a translucent screen. Necessary precautions involve synchronizing the taking film with the projected one and adjusting the colour. It is usual, also, to use a pin-registered intermittent for the projected film so that the background scenes do not show vibration with respect to the actors. Some very convincing pictures can be taken in this way.

The term special effects also includes more ambitious simulations. A background can be painted on a backdrop, as in the live theatre. Sometimes a large glass sheet is used; the actors and the immediate surroundings are photographed with built-up scenery, but the more distant objects are painted on the glass, with the junction merged into a feature of the locale.

It may even be desirable to compress the actual distance in a condensed background by falsifying the perspective with exaggerated building and street angles that merge with the foreground. In this technique the camera point cannot be changed, as it would reveal the false perspective at the merger.

Use of models for special effects

A practice that is not really a special effect but can be noted is the use of full-scale models of heavy stone, brick, or other materials made in cast plastic foam. Some remarkably convincing structures, inexpensive and light to handle, can be built in this way. Scale models often require changes in the time scale and a great deal of care about perspective. A disaster at sea, for example, or an automobile plunging off a cliff can be simulated with scale models and a speeding up of the frame-taking speed.

PROJECTION TECHNIQUES AND EQUIPMENT

Apart from the addition of sound, the most notable change in projectors has been the progressive introduction of automation. Automatic feed for the arc-lamp carbons, automatic film changeover from one reel to another, and other automatic features have been developed.

Picture projection. In projection practice, a cue mark is placed on the film, consisting of a small ring scraped in the film emulsion at the upper right-hand corner of the picture as seen on the screen. This is put near the end of the film at about the time the second machine should be ready. A second cue mark, somewhat later, indicates that the first machine should be faded out and the second machine faded in. The film is usually edited so that the changeover corresponds with a fade-out–fade-in sequence in the narrative. Three machines are often installed, the third being a spare in case of a breakdown (or during maintenance) in one machine.

Lamps. An important element of picture quality on the screen is brightness. The introduction of the arc lamp initiated a constant effort to increase the brightness. A typical lamp design is shown in Figure 9. A fairly large concave-mirror condenser concentrates the light from the most brilliant part of the arc (the crater on the positive carbon) onto the film aperture, and from there to the objective lens. The arc lamp is operated with direct current to avoid the periodic brightness variation every 60th second that would occur with alternating current. A double screw feed moves the carbons together as they burn. The brightness was first increased by increasing the arc current. In time, it was found that if the positive electrode was "cored" with a softer carbon mixture, as in Figure 9B a greater stability and higher crater brightness could be obtained. Then if the core were impregnated with fluoride salts the crater and the immediately surrounding gas became phosphorescent, the brightness increased, and the light became whiter. This was called the flaming arc. The current was then further increased but at the cost of crater stability. The stability was restored by rotating the positive carbon and inclining the negative, as in Figure 9C. This is now called the high-intensity arc. All these developments have improved picture quality, especially with colour film and have also permitted larger screen pictures.

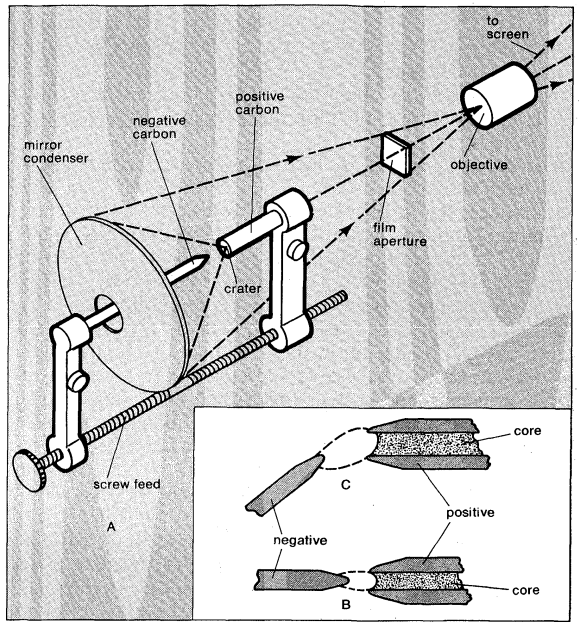


Figure 9: *Typical projector-lamp design.*
(A) Principal parts. (B) Coring of positive electrode with a softer carbon mixture. (C) Rotation of positive carbon and inclining of negative carbon to stabilize the crater, resulting in a high-intensity arc (see text).

Meanwhile, other developments have been studied. A high-intensity form of incandescent lamp, with halogens as noted for studio lighting, and high-intensity xenon glow lamp have been developed.

As the light on the film aperture has increased, so has the heat. To some degree this is abated by the use of a heat-absorbing filter. But the darker areas of the emulsion still tend to expand over the short exposure in the aperture, causing the film to bow in the interval and go out of focus. The film is cooled by a small jet of air to mitigate this.

The intermittent mechanism used in current projectors is generally of the Geneva-cross design (see Figure 10), sometimes called the Maltese cross, machined to close tolerances to give steady pictures and minimum noise.

Screens. The screen used in projection has itself been the subject of development. As screen and picture grew larger, it became necessary to keep the viewed image bright. Light from the screen is wasted if it comes back over the heads of the audience, is too low down, or is too far to the sides. With an ordinary matte-white screen, the return light comes back in a wide cone, less intense as the angle between the incident and reflected beams increases. A good deal of light is wasted at larger angles. Screen surfaces have been developed that are somewhat more specular, or mirrorlike, in their reflection. The design involves a choice between saving light and keeping a wide, usable audience area. One recently developed screen surface is embossed with shaped indentations to form the return light in narrow up-and-down angles, with wider angles from side to side.

Innovations in screen surfaces

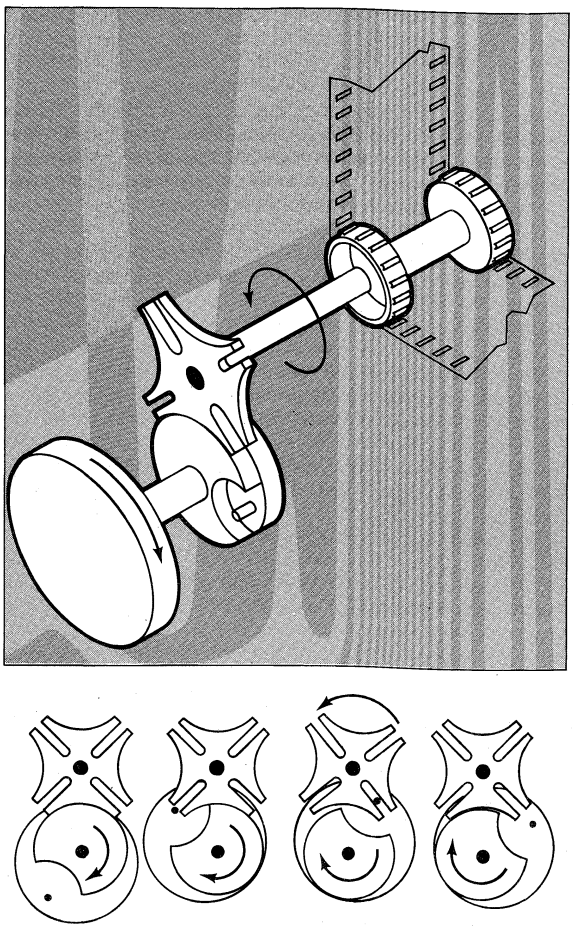


Figure 10: Full Geneva-cross intermittent. Bottom shows successive positions of the intermittent mechanism.
By courtesy of N.V. Philips Gloeilampen Fabrieken

Greater image brightness permits a slightly higher illumination level in the audience without sacrifice of contrast. A moderate level of illumination in the audience is less tiring during prolonged viewing. With sound (see below *Sound reproduction*) it is necessary to perforate the screen with fine holes that reduce reflecting surface and therefore image brightness.

Rear projection (that is, through a translucent screen to an observer on the other side) is occasionally used for exhibition purposes, as well as for special effects. The directional characteristics of a translucent screen are more pronounced than those of a reflecting screen, and a "hot spot," or extra-bright local region, is often observed directly in line of projection from the incident beam; some provision is usually necessary to mitigate this. Another problem is the lateral scattering of light within the thickness of the translucent screen itself, which shows as a haze around the sharp edges in the image, giving loss of crispness or definition. Because of these effects, special lenticulated plastic screens have sometimes been used.

Stereoscopic projection. Stereoscopic projection requires two projectors running in synchronism. In some, the two projectors are incorporated in one machine, but it is also possible to use two machines with interlocked drive. When light polarization is used for image separation, the screen reflection must not destroy the polarization, as would a plain-white matte screen. Aluminized screens are frequently used.

Sound reproduction. The sound track on the film is usually picked up by a "sound head" on the projector. On full-sized film (35-mm or wider), when the track is optical, the recording is placed earlier on the film than the corresponding picture material. Consequently, the sound head must be placed later in the film path so that the two will be projected at the same time. Since the film travels down through the projector, the optical head is located below the picture aperture. When the sound track is magnetic, on full-size film, it is recorded later on the film than the picture and consequently must be scanned earlier in the film travel path. Thus, the magnetic head is located above the picture aperture on the projector.

In 16-mm and 8-mm film (standards are not completely set on the Super-8), both optical and magnetic tracks are located ahead of the picture. Consequently, both optical and magnetic heads are located past the point where the film has passed the picture gate.

Sound amplification

The sound signal, whether from an optical or a magnetic track, is suitably amplified and fed to loudspeakers. In general, modern speakers reproduce with fidelity a fairly wide band of sound frequencies. For theatrical use the speakers embody two or even three parts. The low notes involve a mechanism different from the high notes, and the third, when used, is an intermediate device. These are usually all mounted on one structure. In the case of a relatively large screen, the most realistic sound from one speaker is obtained when it is mounted directly in back of the middle of the screen. To give good-quality sound, the screen must be perforated.

Room design (especially as the room becomes quite large) requires a certain amount of acoustic sophistication. The nature and geometrical shape of the wall and ceiling surfaces become important, as does the general shape and arrangement of the room. The reverberation time (the time for a sound to die down once the source stops) has to be appropriate for the room size.

All these factors become even more important when stereophonic sound, from two or more tracks, is used. The location of the loudspeakers must be planned with care, and their total acoustical power output must be in proportion to the size of the room. In some systems, five loudspeakers are used behind the screen, and special effects are obtained from speakers over the auditorium, all on separate magnetic tracks.

ANIMATION

The picture in motion-picture film does not have to be a photograph; it may be drawn or painted. A start in the technique was made in France about 1908 by Émile Cohl, when he depicted white matchstick figures on a black background. Later developments employed cartoon figures, a concept highly developed by Disney in the United States in the period from 1928 to 1940. Employing a technique of exaggeration in proportions of figures and motion, Disney created feature-length movies with great commercial success. In more recent times, the animation technique has been used on shorter films, such as television commercials or occasional short comic films. The principles are also found to be very useful in teaching films, since they can show exaggerated forms or motions that would not be obvious enough in their natural proportions and can show exploded views of intricate assemblies, with motions added to bring out interrelations that might otherwise be obscure.

Use of animation in teaching films

The practical technique leans heavily on the use of clear-plastic sheets on which the drawings are inked or painted. Sometimes each frame is completely drawn and coloured. Stationary parts may, however, be reproduced on a succession of sheets in which the moving parts are completed as they occur, or various parts may be moved with respect to each other to form the action. The clear plastic permits the lower sheets to be viewed as distinctly as the upper sheets. The final result is convincing largely because of the skillful depiction of the successive steps in the motion shown. It is not always necessary for each frame of the final film to represent the exact status of the motion at that instant; especially in the case of fast or erratic motions, it is often possible to use the same picture for each group of two or more successive frames.

The frames are photographed in a so-called animation stand, which supports the camera at the desired height over the plastic sheets. The sheets can be moved about on calibrated, screw-operated platforms. Very complex stands are needed to depict perspective by separating the plastic sheets by suitable distances and moving the nearer sheets faster than the farther ones.

Another animation technique consists of the use of puppets in a three-dimensional scene. The puppets are moved or changed between the successive frames.

MOTION-PICTURE FILMS IN TELEVISION

Films have been used in connection with television in a variety of ways. The most important is the use of previously made films as program material. The second is the taking of a film during a television program to serve as a recording for further use or possibly for exhibition directly with a projector.

Problems of using films in television

The principal complication arises from the different frame rates in the two arts. The 24-frame-per-second rate was adopted as standard for films when sound was introduced, and, for a variety of reasons, this has proved to be a wise decision. In the United States, the electrical-power supply operates at 60 cycles per second. In the early days of television there were a number of points in the complete television system, from studio to broadcasting station to home receivers, where the power supply could cause interference (cross talk) with the television equipment. It was possible to minimize the effects of this on the received picture by setting the frame frequency at 30 frames per second, and this was done. It had the advantage that the brightness of the received picture could be raised without making flicker apparent. Cross-talk-prevention techniques have improved, and the power supply is no longer a valid reason for the 30-frame-per-second standard, but because of the flicker advantage this has not been changed.

In Europe the power supply operates at 50 cycles per second. In the early days this was the justification for a 25-frame-per-second standard. It is no longer so, and the difference between 24 and 25 cycles per second gives only a very small flicker advantage, but the television standard has not been changed.

In using films made at 24 frames per second on a television transmitter operating at 30 frames per second, advantage is taken of the fact that because of interlace in the television picture the picture is scanned at 60 fields per second (two fields make up one complete picture). Then alternate frames on the film are scanned three times and the intervening ones twice. Thus, the total time taken to scan two frames is $^3/_{60} + {}^2/_{60} = {}^5/_{60} = {}^1/_{12} = {}^2/_{24}$ seconds. This is exactly two frames on the 24-frame-per-second standard. To work this out practically requires a special "limping intermittent," of which there have been several designs.

In Europe, 24-frame-per-second films are run at the 25-frame-per-second speed without perceptible visual effect. The sound pitch is raised by slightly over 4 percent, about a third of a whole tone interval. This is perceptible to a musician but not particularly objectionable to most hearers. More recently in Europe, film intended for television use is often taken at 25 frames per second.

The second important complication arises from the fact that the pull-down time for film (the time during which the film is moving from one frame to the next, with shutter closed, in camera or projector) is much longer than the field retrace time (the time from the end of one field to the start of the next, when the electron beam is blanked out) in television.

The solution makes use of the storage capabilities of the camera tube. The image is projected in a short flash on its sensitive surface during the field retrace time, after which it is scanned by the tube's electron beam. The instants have to be picked carefully because of the relation between fields and frames, though this is less important now that the new tubes permit overlap between the light incidence and the beam scanning without causing interference.

The problems occur also in film recording of a television picture from a cathode-ray screen, but recording of television pictures is now usually by means of magnetic tape.

Besides these more or less mechanical problems, there were also problems in contrast and colour reproduction, in the use of film both as source material for television and as a recording medium for received television pictures. These have required considerable and continuing study, and some reasonable compromises have been reached (see also TELEVISION).

III. Motion pictures for scientific purposes

As soon as motion pictures were invented, they were applied in the recording of scientific phenomena. The recording of an experiment in which a number of things happen at about the same time is especially appropriate for motion-picture recording.

TIME-LAPSE CINEMATOGRAPHY

There are many occasions on which the cinematography can be carried out at normal speeds, and these need no further discussion. There are other situations in which the changes occur very slowly, so slowly that the eye does not discern the change. One example is the opening of a flower blossom. In such a situation the technique used is to take successive pictures at intervals of, for example, an hour, taking great care not to move the camera or the plant and to project the resulting film at normal motion-picture speed. The projected picture will disclose many details in the development from the bud to the completed flower that are not apparent in ordinary visual observation. Other phenomena can be studied in this way. These techniques require merely a standard camera that can take single exposures, plus a timed triggering device that can take the exposures at the desired intervals. The rest of the technique is mostly a matter of preventing undesired motions of camera and subject.

HIGH-SPEED CINEMATOGRAPHY

Motion pictures have also been used to study phenomena that occur so fast that they cannot be recorded on normal cameras. An immense amount of ingenuity has been applied to the solution of many problems in this field, and a short discussion can only summarize the techniques used.

Optical systems outside the camera. *Shadowgraph.* The film can show sharp shadows of fast-moving bodies, indicating their speed and any changes of attitude they undergo. With a rapid enough exposure this procedure can also show the sharp shadow of a wave front in air.

Schlieren optics. The procedure shows changes in the condition of a test area of space in the optical path, even when it remains fully transparent. A simplified outline of

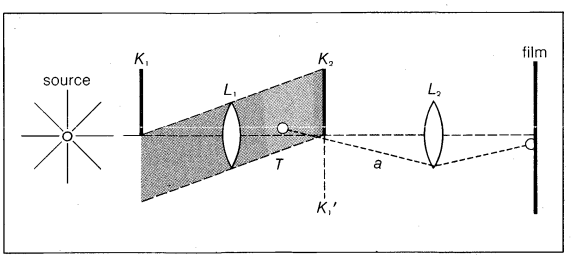


Figure 11: Schlieren optical system (see text).

the arrangement is shown in Figure 11. A source sends out a beam of light through the apparatus to a film. Half the field is cut off by a knife-edged screen, K_1, which is imaged by a lens, L_1, to a position, K_1', in the same plane as a matching screen, K_2. The knife-edges of image and screen exactly coincide. A test position, T, is sharply focussed on the film. Thus the light flow past K_1 to K_2 is all cut off from the film if the space at T is completely uniform. Any nonuniformity, such as caused by a wave front in the air at T, causes a scattered light beam to evade the screen, K_2 (path a), and reach the film.

Schlieren interferometer. This is a modification of Figure 11. Apparatus that is too complex to describe here but that polarizes the light beam—separating it into two slightly different paths and then reintegrating it into one path—is inserted before and after the area T, to bring out details of the phase of the light as it progresses from left to right. These details appear in the form of light and dark phase bands, or fringes. They outline the disturbance of the wave by any nonuniformity existing at T. Counting the number of fringes displaced gives a measure

of how great the nonuniformity is. The arrangement is especially adaptable to detecting and measuring irregular flow phenomena and wave fronts in the air (or another fluid) at *T*.

Time variation of spectrum. A spectrum of a self-luminous subject is often obtained. Sometimes the spectrum varies quickly, as in an explosion, and it is desired to study the variation. The light output from the spectrometer is led into a motion-picture camera that can handle the speed needed, and a film record is taken. The film can then be studied at leisure.

Framing. The normal-speed camera records on the film a succession of frames, each showing a complete picture but taken at successive instants of time. Not all high-speed cameras record in this way, but some do, and the various methods of separating the frames and recording them represent key elements in the technology. As a framing method, the intermittent mechanism has limitations, since it is not capable of being pushed very much beyond the normal speeds. Even if the mechanism itself could function, the sprocket holes in the film are not strong enough for the necessary accelerations.

Rotating prism. When a cube (or many-sided prism) is rotated in the optical path in a camera, it moves the image periodically past the picture aperture. If the film is moved in the same direction at the same speed, the image, while it is appearing, is fixed on the film. The problems are that, during its appearance, the image does not move with an exactly constant velocity, and the prism introduces some optical distortion of the image. These can be partially remedied by a variety of modifications, but there are still residual imperfections. The device is used, however, up to about 10,000 frames per second. A variation, in which an internal mirror drum with many facets is used instead of a glass prism, is usable up to some 40,000 frames per second. In evaluating the performance of the system, the size of the frames and sharpness of the images must be considered, as well as the number of frames per second.

Rotating-mirror system. Such systems take a number of forms, but Figure 12 shows a reasonably typical arrangement. The objective lens forms a primary image at

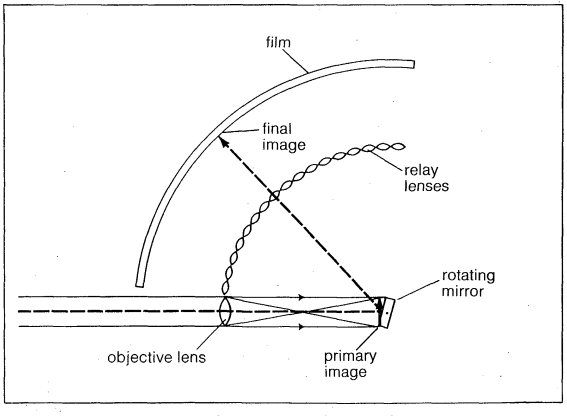


Figure 12: Arrangement of the principal parts of a rotating-mirror system (see text).

the rotating-mirror position. The mirror reflects this image through an arc of fixed, individual relay lenses in succession on the stationary film. The relay lenses having the primary and final images both in focus keep the final image fixed at each place in succession on the film as the beam traverses each lens. With some rotating-mirror systems it is possible to go to some 5,000,000–10,000,000 frames per second (again keeping in mind that frame size and image sharpness are important as well). The arrangement wastes light, so that bright original subjects are necessary for a satisfactory exposure. The system has been used to study explosions and plasmas and to study subjects by reflected light with very high intensity illuminants.

Image dissector. The central idea of this technique is to cut up a picture in the same way a television picture is cut into horizontal scanning lines. The bright lines are made very narrow, with black or empty space between, which permits superimposing other, similar pictures—also in fine lines separated by empty spaces. A number of pictures can be superimposed in this way without the lines interfering with each other. Lenticulations (embossed lenslike shapes) on the film form the scanning lines and concentrate the incoming light into fine lines on the film, according to where the light comes from on the objective lens.

The method will also work with two sets of lenticulations, one horizontal and the other vertical and superposed. This permits the storing of more pictures on the record film—in this case, as dots instead of thin lines. A number of devices based on these general ideas have been used. The frame speeds achievable vary with the specific arrangements but are generally lower than with the moving-mirror mechanisms.

Streak camera. The number of frames in Figure 12 may be increased to infinity by changing the optical system to put the primary image on the film and leaving out the relay lenses or, more simply, to replace the mirror with the film on a rotating drum. This requires compressing the individual frame from a two-dimensional into a one-dimensional image or slit. The equipment becomes a "streak camera." It is useful when there are one or more distinct demarcations in the one-dimensional view of the subject, and it is desired to record their motion along a more or less straight line. This is useful, for example, in recording the motion of the front of an explosion cloud or of a projectile. It leaves no unexplored areas of time, as does a framing record.

Illumination. The need for sufficient light is a most important requirement in high-speed cinematography. This need is so great that it is often undesirable to keep the light on longer than absolutely necessary, so that in many cases the light being turned on and off acts as a shutter. High-intensity, short-time sources have been studied extensively. Most sources consist of electrical discharges or, in some cases, arcs in air or a gas. In most cases they involve sophisticated electronic control methods.

Separate shutter. In cases in which the original subject is sufficiently self-luminous, such as in explosions, a fast shutter may be required. Since mechanical shutters do not go to very high speeds, various types of electro-optical shutters are used. Typical is the Kerr cell, in which a substance between electrodes can be made opaque or transparent according to the voltage on the electrodes. It can be made extremely fast acting in response to a sharp voltage pulse. Sometimes a relatively slow but positive "capping" shutter is required, to cut off light before and after an experiment. This is usually a mechanical shutter.

Electro-optical shutters

Image converter. On some occasions it is necessary to amplify the light after it has left the subject and before it reaches the film. In recent years "image converters" have been developed that can do this. Such devices usually project the light image onto a photosensitive surface that emits electrons. The electron beam is amplified and projected by electron optics onto a luminescent surface, which is then photographed optically on a film. These devices can also act as shutters by the use of a triggering pulse on one of the electrodes.

X-rays. X-rays are sometimes used as the source of illumination, but these rays do not permit the usual optical procedures. Slits in lead screens must be used instead of lenses and mirrors. As is the usual practice with X-ray films, both sides of the film are coated with emulsion. Also, an intensifier is used, consisting of a suitable fluorescent screen backing the film.

IV. Amateur motion-picture photography

Throughout the world the number of amateur cinematographers has increased steadily since the 1920s. As the quality of amateur equipment and the performance of the narrower film have improved, this equipment has been used on more and more professional tasks. News events can be covered more effectively with a compact amateur camera than with cumbersome professional equipment.

In addition, there is a growing number of semiprofessional makers of motion pictures. Their products include teaching films, used by schools of all types. Such specialized topics as geography, history, and archaeology can be effectively treated. Also involved are business and commercial films, such as promotional movies of various kinds. A new field that has become very popular is the making of amateur entertainment films by budding actors and directors.

The field of amateur-motion-picture technology is quite varied and forms a substantial part of the overall industry.

Camera. The need for simplicity of operation, lightness and compactness, ruggedness to withstand amateur abuse, and low cost have brought many camera developments that are gradually finding their way into the professional field.

The external shape of the camera has gone through a number of changes since its introduction in 1923. The present-day amateur camera takes a variety of forms. It is generally (but not always) hand held and often has a large pistol-grip handle to improve steadiness. Film reels are carried inside the body, and loading is possible in daylight. Loading has been so simplified that it sometimes consists merely of inserting a cassette, without the need for threading the film.

Modern amateur cameras

The lens most popularly used is of the zoom type, with a ratio of effective focal lengths running in some cases up to 10:1. In some instances this is manually moved; in others it is battery-motor operated. Exposure is nearly always automatically controlled. A choice of frame speeds is often available on the camera, usually from 16 to 20 frames per second.

Film. At one time 16-mm film on safety base was the most popular amateur film. In recent times, however, 16-mm has become a professional and semiprofessional film and has been replaced in the strictly amateur field by the 8-mm and Super-8-mm. The reasonably good quality of these latter and their lower costs and greater compactness have been the chief causes.

The 8-mm and Super-8-mm colour film is available in several emulsion speeds for daylight and also artificial light use and packaged in rolls, magazines, or cassettes.

The processing, carried out by a laboratory, involves the same general type of equipment as described above for professional film. Editing is done in a much more informal way, and some relatively inexpensive equipment for this is available for amateurs.

Very little printing is done for the home-movie amateur, although sometimes extra copies are made from original film. Printing for the semiprofessional, however, is another matter. Films for promotional or sales use or for teaching may be printed in fairly large quantities. Special printing machines are available that can print and process several copies at the same time, often on wide stock prepunched with sprocket holes. After processing, the wide stock is slit into the several copies. The choice of specific arrangement is almost entirely a matter of economics. One problem has been the lack of final standardization in packaging of the amateur film, which influences some automation steps. This sort of printing is mostly done as a reduction from 16-mm original, inasmuch as printing from 8-mm original to 8-mm stock does not give the highest quality results.

Machines for multiple printing

Lighting and light measurement. This is generally much more modest than for the professional. The semiprofessional, however, uses some of the smaller lighting equipment of the professional. The home-movie amateur is still more modest; for artificial light he has used 65-volt lamps on 110-volt supply to give intense light with a relatively short (but reasonably adequate) life. Special 650-watt tungsten-halogen lamps are now available for this purpose, in some cases held in reflectors that can fit on the camera.

Light measurement is currently nearly always accomplished automatically with a photocell in the camera. In many recent cameras it is combined with the viewfinder, and readings are taken through the lens by using a bypassing, partially reflecting prism in the taking beam. Some one-third of the light is used for viewing, together with a range-finding prism, and for a photocell; an amount that does not seriously reduce the amount of light available for film exposure.

Sound recording. All of the three small film sizes carry an optional sound track, optical or magnetic. The way in which the sound track is used, however, varies a good deal with the nature and interests of the photographer.

Semiprofessionals by themselves cover a wide range of interests. The makers of serious teaching films and of promotional films generally use much the same sort of techniques used by the full professionals, with some simplifications and a greater degree of informality. They almost always use 16-mm film and record the sound on a separate film (or tape), registering magnetically. There are some cameras, however, that ordinarily record picture only but in which a sound-recording magnetic module can be inserted to record both picture and sound. It can be used even with a recording amplifier having built-in mixing to be done on the set. The final exhibition print of the film usually has an optical sound track.

The film as finally used is likely to be 16-mm, but there is some tendency to reduce the format in printing to 8-mm, particularly to Super-8 and in general with the magnetic track. There are occasions when the 16-mm original is enlarged to 35-mm, but this is hardly in the amateur category.

Reduction in format

The home-movie amateur generally restricts himself to 8-mm film. Few 8-mm cameras, whether standard or Super-8, include provision for simultaneous sound recording. If sound is used, it is dubbed on the magnetic track afterward. There are some cameras, however, that include a special contact to send a pulse at every frame over a wire connection. This can be used with a tape recorder, on a separate track, to synchronize picture and sound. The film is projected on a special projector that has an arrangement to lock the speeds of the projector and tape recorder (and start the two in phase) so that the recorder sound matches the picture on the screen. These devices are not yet generally standardized.

Projection. The projection installation varies with the nature of the amateur and the use involved. For the serious semiprofessional, the 16-mm film is projected, with a sound track (usually optical) pickup from the film. This is usually the situation with teaching films, amateur entertainment films, and some forms of promotional-film uses. The equipment includes a moderate-sized screen and usually a single loudspeaker, and installation can range from permanent to portable. The projector itself in such cases is very much reduced in size from the professional projector and usually uses a high-power incandescent lamp (frequently of the tungsten-halogen type) or an enclosed tungsten arc lamp. For some promotional uses the projector is compact and throws the light on a small integrally mounted rear-projection screen.

For both teaching and professional use, Super-8-mm projectors have been designed to employ the semi-professionally produced sound track on the film and project onto moderate or small-size screens using a tungsten-halogen lamp. These have been designed for cassette-packaged film with many automatic features. There are several forms of cassettes and of sound pickups, and the field is only partially standardized.

Finally, for the home-movie amateur, projectors are offered with or without sound pickup—often permitting up to seven frame speeds (arranged with special frame repetition at low speeds to avoid flicker)—with zoom lenses to permit quick adjustment of projected picture size and accepting films in reels or various cassettes (but not on the same machine). Some machines accept only standard 8 and others Super-8; some accept either with a small adjustment. Many machines use tungsten-halogen lamps, and some use standard high-power incandescent lamps. Some machines, as noted earlier, are specially designed to be synchronized with a double-track tape recorder. The possibilities are quite varied.

The screens used run all the way from small portable screens to fixed moderate sizes installed in a small audito-

rium. Some amateurs still use a white cardboard or a white wall, but these do not generally give a high light effectiveness.

BIBLIOGRAPHY. R. SPOTTISWOODE (ed.), *Focal Encyclopedia of Film and Television* (1969); and L.A. MANNHEIM (ed.), *Focal Encyclopedia of Photography*, rev. ed. (1969), are two fairly complete one- and two-volume encyclopaedias on their subjects, for the nonspecialist. TERRY RAMSAYE, *A Million and One Nights*, 2 vol. (1926); and LEWIS JACOBS, *The Rise of the American Film* (1939), are two extensive, chatty, and unusually detailed stories of the development of the motion picture. RAYMOND FIELDING (ed.), *A Technological History of Motion Pictures and Television* (1967), reprints collected landmark papers from the *Journal of the Society of Motion Picture and Television Engineers* on the history of these two arts; the SOCIETY has also compiled *The Technique of Motion Picture Production* (1944), that covers many then-current details of this art. D.V. KLOEPFEL (ed.), *Motion Picture Projection and Theatre Presentation Manual* (1969); and JEAN VIVIE, *Projection des images animées et reproduction des sons* (1968), are projection manuals on the modern technology of motion-picture projection. GEORGES SADOUL, *Cinéma d'Aujourd'hui*, vol. 1, *Georges Méliès* (1961), vol. 29, *Louis Lumière* (1964), are two little volumes describing the early experiences of these pioneers, mostly in terms of their pictures, rather than the technology. There are also others in the series. JEAN VIVIE, "Motion Pictures and Education in France," *J. Soc. Motion Pict. Telev. Engrs.*, 78: 950–955 (1969), is a chatty story about the development of the narrow-gauge film for educational use in France. J.S. COURTNEY-PRATT, *A Review of the Methods of High-Speed Photography* (1957), and *Proceedings of the Fifth International Congress on High-Speed Photography* (1962); WILLIAM G. HYZER and WILLIAM G. CHACE (eds.), *Proceedings of the Ninth International Congress on High-Speed Photography* (1970), include detailed techniques in high-speed and scientific cinematography. R.J. ROSS, *Television Film Engineering* (1966); H.M. TREMAINE, *The Audio Encyclopedia* (1959); R.W.G. HUNT, *The Reproduction of Colour* (1967); Y.N. GORokhovskii, *Spectral Studies of the Photographic Process* (1965; orig. pub. in Russian, 1960); C.E.K. MEES and T.H. JAMES, *The Theory of the Photographic Process*, 3rd ed. (1966); and RUDOLF KINGSLAKE, *Applied Optics and Optical Engineering*, 5 vol. (1965–69), are more extended discussions of specific portions of motion-picture technology. A. BONI (ed.), *Photographic Literature* (1962), is merely a bibliographic listing, but useful to hunt up occasional articles. Periodical literature (all monthly): *Journal of the Society of Motion Picture and Television Engineers*; *British Kinematography, Sound and Television*; *Popular Photography*.

(P.Me.)

Motion Sickness

The term motion sickness was proposed by J.A. Irwin in 1881 to provide a general designation for such similar syndromes as sea sickness and train sickness. This convenient term, imprecise for scientific purposes, has nevertheless gained wide acceptance.

Definition. Manifest, or frank, motion sickness comprises a constellation of symptoms resulting from sudden exposure to periodic unnatural accelerations, and its course is self-limited through the mechanisms of adaptation. The nonacoustic part of the inner ear must function if the symptoms are to develop, but secondary factors, tending either to increase or decrease susceptibility, always are present. Normal persons are inherently susceptible, and susceptibility in the same person may differ in different gravito-inertial force environments. Overt manifestations include pallour, sweating, salivation, drowsiness, and (most important from a practical standpoint) the nausea syndrome. First-order responses give rise to second-order responses and complications of ever-increasing diversity, which may lead to prostration. The diagnosis of frank motion sickness is so easy that it is not a matter of professional concern, but mild symptoms may pose a diagnostic problem. In addition, confusion may arise because similar symptoms may be elicited in the absence of "motion"; *e.g.*, by syringing the ear with cold water, or by manipulating the visual input. Prevention is the key to good management in all cases. Motion sickness, theoretically, is preventable, but countermeasures may be impractical in many circumstances. Once symptoms are well established, quick relief may require intensive therapy.

Despite the prevalence of motion sickness and its age-long challenge to the investigator, most of the events and processes underlying the manifestations of motion sickness have not been elucidated. Some of the elements of a conceptual framework are discernible, however, and will be presented here in truncated form.

Vestibular systems. The nonacoustic portion of the inner ear comprises the otolith apparatus and semicircular canals, collectively termed the vestibular organs. In each ear the sensory receptor mechanisms in the three mutually perpendicular canals are stimulated normally by rotary motions of the head or head and body. The paired otolith organs, their receptor mechanisms in the shape of four curved plates constituting *in toto* a significant portion of a sphere, are stimulated continually by gravity and also by sudden linear accelerations. The vestibular organs are encased in a bony labyrinth and, while protected from external mechanical stimuli, must react slavishly to accelerative forces insofar as their response characteristics permit. The important point to be made here is that vestibular stimulation eliciting motion sickness may result from exposure to motions in conveyances either by virtue of (1) linear oscillations (as when a ship rises and falls); (2) angular oscillations (as when a ship rocks); (3) Coriolis acceleration (as when a ship pitches and rocks simultaneously); (4) Coriolis accelerations generated by two angular velocities, one contributed by man and one by machine, as when nodding the head on a carrousel; and (5) passive rotation at constant velocity in the laboratory when rotated in the "barbecue spit" or earth horizontal axis.

Types of motion causing motion sickness

The vestibular organs have a large representation in the brain stem, cerebellum, and spinal cord. By virtue of an extensive and complex neuronal circuitry, mainly at subcortical levels, the vestibular systems participate in such functions as the maintenance of postural equilibrium and in the coordination of eye-head-body motions. Although evolutionary adaptation has ensured that such participation is effected in elegant fashion in naturally occurring accelerative environments, this guarantee does not hold true for exposure to all types of artificially generated accelerations. This limitation accounts for the great individual and intraindividual differences in susceptibility to motion sickness but, of course, explains nothing.

Mechanisms and manifestations. When a susceptible person is exposed to an adequate accelerative stress, the resulting abnormal pattern of impulses to the vestibular system exceeds the capacity of the vestibular systems to maintain a state of equilibrium, or homeostasis, with two distinguishable consequences. The first consequence is confined to the vestibular systems proper and is manifested by disturbances in normal responses. The second consequence involves a "failure" in homeostasis, allowing the irradiation of vestibular activity to reach foreign servation systems; the manifestations evoked include the typical symptoms of motion sickness and are superimposed on any manifestation of the first. The fact that, in this case, manifestations appear that have their immediate genesis in nonvestibular systems implies a linkage of a temporary nature with the vestibular systems. Thus the main chain of events involves the vestibular systems, a facultative linkage, and the nonvestibular systems that will be considered in turn. Through selection of subjects and control of stimulus it is possible to bring about vestibular perturbations without causing motion sickness. Typical responses include nystagmus (rapid involuntary eye movements), disequilibrium, and "sensations" of turning that have the characteristics of reflex phenomena. Continued exposure to a constant level of accelerative stress leads to their disappearance through the mechanisms of adaptation; a specific mechanism involves the establishment of new central nervous system integrative patterns, and a general mechanism may involve activity arising from the vestibular systems. Adaptation both to these sensorimotor perturbations and to motion sickness takes place in the vestibular systems proper. Vestibular aberrations and hence motion sickness can be avoided when an otherwise intolerable level of stress is reached by incremental steps; indeed, this is one of the keys to the prevention of motion sickness.

Perturbed vestibular responses without motion sickness

Little is known concerning the facultative linkage. The

fact that irradiating vestibular activity is demonstrably open to modulating influences points to the use of common pathways in the brainstem reticular formation. The sometimes long delay in appearance of motion sickness suggests that a chemical linkage also may be involved.

Response of nonvestibular systems

The manifestations having their immediate origin in nonvestibular systems are far more widespread than the overt signs and symptoms of motion sickness mentioned earlier. These responses do not have the characteristics of typical reflex phenomena and are absurd in terms of the needs of the organism. Unfortunately, there has been no systematic attempt to categorize these manifestations as first-, second-, or third-order effects and their complications. The release of antidiuretic hormone (which reduces the output of urine), and the appearance of stress hormones (catechols and corticoids) in the urine testify to the general nature of the signs and symptoms when the stress is sufficient and sufficiently prolonged.

The recovery from motion sickness during continual exposure to the stressful accelerations is complicated. The nonvestibular systems must first be freed from outside influences (vestibular activity, chemical agents), after which restoration takes place spontaneously through homeostatic events and processes. Curves depicting the time course of these events overlap, but as yet these curves have not been clearly defined.

The responses evoked on sudden return to a stationary environment depend on the level of adaptation acquired in the vestibular systems and upon whether restoration was complete in nonvestibular systems. If adaptation is incomplete, one experiences both a loss of muscular coordination and motion sickness; if complete, motion sickness is absent.

The roles of vision, distraction, illness

Secondary influences. It may be concluded that any factor tending either to evoke or inhibit a response characteristic of motion sickness will affect susceptibility accordingly. Vision plays an important role. A pilot, for example, free of symptoms in the cockpit, may experience motion sickness in the navigator's closed compartment. Mild symptoms of motion sickness have disappeared under the influence of experimenter-directed tasks that may have preempted central nervous system pathways used by irradiating vestibular activity. Illness may raise susceptibility greatly; indeed, a provocative test may bring out the symptoms of the disease before they would otherwise have appeared.

Prevention. The prevention of motion sickness is a highly individual matter. Persons differ not only in their inherent susceptibility to unusual accelerations but also in the roles played by secondary factors, in the acquisition and decay of adaptation effects, and in the effectiveness of drugs. Assessment of these factors could be made in the laboratory, but for the great majority of persons it is a matter of self-appraisal under travel conditions by land, sea, or air. The more complete this assessment and the more precise the knowledge of the duration and conditions of travel, the greater the opportunity to exploit preventive measures. It is particularly important to guard against the false security that may be engendered by the delay between the onset of exposure and the onset of overt symptoms.

It will be apparent that certain options suggested in the following discussion are not available to everyone and that certain trade offs are "judgment" decisions.

Minimizing accelerations. The single most effective measure is to fix the body, especially the head, with reference to the vehicle. Minimizing accelerations in a particular vehicle involves such things as seat location and time of day in an aircraft and cabin location and time of year in a ship. Choice of vehicle for a given mode of travel may be important; *e.g.*, flying above turbulence in a jet aircraft and use of stabilizers in a ship.

Incremental exposure. This measure is at once highly effective and highly impractical except in using private conveyances. Any incremental exposure to stress on a particular trip in common carriers is a fortuitous matter. The traveller does have the option of gradually increasing the duration of trips and, to some extent, choosing favourable conditions. On long ocean voyages, during an initial period of calm at sea, moving about would increase the rate of adaptation; and this might even be extended to less favourable conditions under the protection of drugs. This is a bit risky, however, for the inexperienced.

Secondary factors. Unless vision is used to "lock" onto the horizon or to engage in activities resulting in competitive inhibition (distraction), there is less tendency to motion sickness if the eyes are closed. Certain illnesses, especially those with gastrointestinal symptoms, may increase susceptibility to motion sickness. Susceptibility is lessened if the most pleasant atmospheric conditions are sought or arranged and if the thought is kept in mind that the very motions evoking symptoms also result in adaptation, which may be near at hand.

Drugs. Persons who are in the passenger category do well to test a "new" drug for its side effects prior to travel. A combination of l-scopolamine hydrobromide and dextroamphetamine sulfate, taken 40 minutes before departure, affords some protection for several hours. Four hours later the dose is repeated unless the trip will soon be over. There is need of caution both in continuing the prescribed dosage and in using larger doses that afford increased protection. Promethazine hydrochloride, with minimal risk of side effects, is a good substitute for scopolamine. Ephedrine is only a fair substitute for amphetamine but has minimal side effects.

Several drugs in the antihistamine category decrease susceptibility to motion sickness, including diphenidol, dimenhydrinate, cyclizine, and meclizine. The last named is effective over periods up to 24 hours.

Today many persons are under therapy with drugs that may cause nausea. It is highly desirable, therefore, to seek professional advice with regard to continuing these drugs.

Therapy. The inexperienced person with motion sickness needs the best available professional help. Once symptoms of motion sickness are established, appropriate preventive measures are taken, as have just been described; but drugs taken by mouth may either remain in the stomach or be regurgitated. The most effective measure, recommended only for long trips, may involve injection or rectal administration of a soporific drug in an amount to ensure sleep.

BIBLIOGRAPHY. K.E. MONEY, "Motion Sickness," *Physiol. Rev.*, 50:1–39 (1970), provides the most recent comprehensive review in which motion sickness is treated mainly as a medical syndrome. The conceptual approach in the present article represents essentially a condensation of a report by A. GRAYBIEL, "Structural Elements in the Concept of Motion Sickness," that appeared in *Aerospace Med.*, 40:351–367 (1969). Anti-motion sickness drugs are considered at some length by S.C. WANG, "Emetic and Antiemetic Drugs," in *Physiological Pharmacology*, 2:255–328 (1965). Recent evaluations of representative anti-motion sickness drugs are described in two articles by C.D. WOOD and A. GRAYBIEL in "Evaluation of Sixteen Anti-motion Sickness Drugs," *Aerospace Med.*, 39:1341–1344 (1968); and "Effective New Motion Sickness Remedy," *ibid.*, 41:932–933 (1970).

(A.G.)

Motivation

The term motivation popularly refers to the causes of behaviour; *i.e.*, whatever it is within the individual that incites action. The term motive is commonly applied to fairly specific causes. Thus, it often is said that someone is highly motivated for success and, more specifically, that success serves his motive for achievement or for power, status, recognition, or some other more specific motive. Yet, motivation can hardly be the sole basis for success; one must have ability, training, knowledge, and many other characteristics to achieve specific goals. What this popular definition effectively implies would seem to be the mainspring of success: the characteristics that drive one person to achieve a particular end that another individual with similar endowments apparently lacks.

The word motivation comes from the Latin term *motivus* ("a moving cause"), itself derived from *motus* ("moved"). The implication that it refers to the springs of action seems evident in such phrases or words as motive power, motive force, automotive, and the like. Not until the 20th century, however, did "motivation" seem to enter

the vocabulary of philosophy and psychology. The concept was not in the *Dictionary of Philosophy and Psychology*, publication of which was completed in 1911; the word motive was listed, but in the sense of an awareness of desire or purpose rather than as a moving force.

HISTORICAL OVERVIEW

According to the hedonistic principle, human actions are governed by a calculus or combination of pleasure and pain. Although hedonism in this sense is a motivational factor, the full advent of motivational ideas among psychologists was not to come until the latter part of the 19th century, following the development of Charles Darwin's theory of evolution (*Origin of Species*, 1859).

Among the significant effects of Darwin's theory were these: first, that man and more primitive animals were seen as being parts of the same continuum, rather than as sharply demarcated from one another as rationalist ideology had held; second, that the structural and behavioural characteristics of all animals, including man, were regarded in terms of their survival value or of their usefulness for adaptation to a particular environment. These effects were far reaching. As to the first effect, it suggested that if less completely evolved animals have instincts, so also does man (and if man has reason so also may other animals). This has been summarized as leading to the "brutalization" of man and to the "humanization" of animals. The effect led to the development of comparative psychology and to an emphasis on instincts in the control of human behaviour.

Impact of Darwin's evolutionary system

The second effect was equally, if not more, important. It made relevant the study of behaviour as a means of adaptation to situations and environments. Prior to this time there had been little interest in the empirical study of behaviour, and psychological approaches to the study of man largely had been confined to determining the contents of consciousness, the elements of which these contents were constituted, and the rules for the combination of the elements into larger units. This view of the task of psychologists (held by Wilhelm Wundt [1832–1920] and Edward B. Titchener [1867–1927], for example) had no room for motivational determinants as they are defined by contemporary theorists. But evolutionary theory produced functional psychology, whose adherents were mainly interested not in what *is in* awareness, as Wundt and Titchener were, but rather in what consciousness *is for*; *i.e.*, how awareness functions in the organism's adaptive interactions with the environment. Functional psychologists also broadened the task of psychology to include not only conscious functions but the whole range of the behavioral repertories of animals. (Behaviourism, a later development, simply eliminated the study of conscious states from the functionalist program.)

For the study of motivation, the first consequence of evolutionary theory, then, was a primary stress on notions of instinct, man's behaviour being interpreted as the manifestation of many instincts. William James (1842–1910) wrote that human beings have more instincts than do any other animals, and many other writers attributed instinct after instinct to man. One writer facetiously went so far as to propose that there is an instinct to believe in instincts. Almost any behaviour came to be called instinctive. Unfortunately, instincts were proposed without empirical support or with only the most casual sort of evidence. There was little scientific value or gain to this uncritical, proliferative invention of instincts, and there was such a powerful reaction against this trend that the word instinct came into disrepute, almost disappearing from the psychological vocabulary for a number of years. The term began to return to popularity in the 1950s, however, the earliest reappearance with some frequency being found in the writings of certain European zoologists known as ethologists. (Ethology is the scientific study of animal behaviour.)

Motivational concepts survived, however, and it is to three individuals that this survival may be attributed: William McDougall (1871–1938), Sigmund Freud (1856–1939), and Robert S. Woodworth (1869–1962). All of them wrote of instincts, but the real emphasis of these men lay on energetic sources of conduct, by whatever names they might be called.

Instinct as the driving force

McDougall identified his point of view as hormic psychology, the term hormic being derived from the Greek word *hormē* ("impulse" or "striving"). He defined the instincts as being the prime movers of conduct, stating that without them the organism would be inert. Purpose, goals, and striving he saw as the hallmarks of human conduct. He wrote that the instincts in human beings determine not only perception of and attention to relevant objects but also generate emotional excitement when such objects are perceived and produce impulses to act toward the objects in some way. McDougall saw mankind essentially as driven by irrational instincts, and he posited a need to control these as the main reason for socialization. In his theory of personality he provided for the control of instincts through the concept of what he called the sentiment, much as Freud did in his theory when he invented the notions of the ego and the superego. McDougall's contributions are important, especially in the history of academic psychology, because his stress on impulse and striving had to be acknowledged. While his extreme emphasis on instincts was later abandoned by others, his concern for impulsive sources of action was not, being incorporated into the contemporary concept of drive.

Like McDougall, Freud emphasized the irrational, impulse-driven aspects of human conduct and proposed that it rested on an instinctive basis. Freud was most especially concerned with the unconscious character of these motivational factors (McDougall was not) and suggested the existence of fewer specific instincts than did McDougall. Since Freudian theory has survived as a pervasive influence, it will be treated more fully below.

Woodworth, who brought the term drive into the psychological lexicon, regarded instincts as useful concepts but distinguished two features in the preparatory and consummatory behaviour of which he said they were composed. One of these features he called the mechanism, the other the drive. The automobile is an illustration of what Woodworth meant: it has mechanisms for imparting motion to the wheels and mechanisms for steering them, and all of the structures are present, whether the machine is moving or at rest. The mechanisms require a supply of energy or a force to make them go. It was this force that Woodworth identified as drive. Thus Woodworth performed the service of distinguishing between mechanisms and drive as two important components of action. His neutral term drive served in lieu of the more controversial term instinct.

THEORIES OF MOTIVATION: MOTIVES AND GOALS

Although there are a number of theoretical views of motivation, it seems fair to say that no one of them can be considered as fully adequate. In a number of respects these theories are not directly comparable. To some extent, the adequacy and comparability of the theories are difficult to judge in that they tend to define the problem and function of motivation differently, especially since their origins lie in attempts to cope with differing sets of phenomena. Before entering into descriptions of particular theories of motivation, it will be helpful, first, to comment on the following three general issues that pervade the subject.

Homeostatic mechanisms. Homeostatic regulation is a familiar concept when exemplified by the thermostat in a residential heating system. Such a device is sensitive to temperature variations, and, when the temperature falls below a selected point, it will switch on the furnace; likewise, the thermostat is set to turn off the source of heat energy when the temperature has risen to some designated value. The thermostat serves thus so to regulate the production of such energy by means of the furnace that the temperature of a room or of a house remains relatively constant within a narrow range of variation. The tendency for such an arrangement to produce more or less constant temperature can be referred to as homeostatic, although the term homeostasis ordinarily is applied to biological systems. In physiology there are constant levels that regu-

latory processes function to maintain. For example, when a person's body temperature rises, perspiration increases, functioning to regulate the level through the loss of thermal energy through evaporation. When body temperature falls, other regulatory processes (*e.g.*, shivering, increased muscular activity) come into operation to generate more heat.

Many writers have depicted behaviour as a regulatory process, called into play by some imbalance in bodily or psychological forces, and functioning to restore a state of equilibrium. Imbalances are considered as being motivational states and as ranging in kind from such conditions as deficiencies of food or water in the body to deficits or excesses in the external stimulation that impinges on the body's receptor structures. Homeostasis can be used as a relatively precise physiological term, but often it is employed in such a loose way that it can be made to fit a much broader set of circumstances. The notion of homeostatic regulation runs through much of the theorizing about motivation; depending on the set of assumptions that characterize any given writer, however, it has been disputed and denied.

External and internal determination. In some theories (McDougall's for example) motivation seems to be treated as having an internal bodily locus and as being an impulse that propels the organism into action. In other theoretical systems, however, the emphasis is on sources in the surrounding environment that may be said almost to pull or push the organism toward them or from them.

Role of motivation in behaviour. Behaviour varies in many ways; it occurs or it does not; it is produced in varying amounts and at different speeds; it is initiated slowly or rapidly; it may be intense or weak, broad or narrow in its scope. These variations are essentially *quantitative*; *i.e.*, any given behaviour (such as eating) can be described in terms of its different degrees as being strong or weak, fast or slow, and so on. Behaviour also may be seen to vary in *kind*: there may be eating, drinking, or mating; approach or withdrawal; concentration on one goal rather than on another. Such qualitative differences are commonly employed to indicate what usually is termed the direction of behaviour; the quantitative variations mentioned first in this paragraph reflect an energetic component in any kind of behaviour.

To some theorists it is the direction of behaviour that requires the motivation concept. Direction implies qualitatively different goals, purposes, choices, preferences, and decisions; it is such processes that these theorists hold to require motivational factors for their explanation. By contrast, other writers suggest that motivation is not a necessary notion in accounting for direction, preferring to rely on other concepts (such as habit) to explain qualitative behavioral variation. In the view of such authors it is the quantitative dimensions, reflecting changing levels of energy expenditure, to which the motivation concept is pertinent. In their view, motivation serves not to direct but to energize behaviour.

In what follows, the three issues just summarized will be apparent in the exposition of representative theories; the discussion will focus on classes of theory, giving an account of a specimen typical of each. No effort will be made to detail or even name the richly diverse variants within these classes.

Psychoanalytic theory. Classical Freudian theory emphasized internal determinants and was homeostatic in character. It seemed to ascribe to motivation both the functions of energizing and of directing behaviour, perhaps emphasizing behavioral *qualities* most.

Freud modified his theory from time to time, so that the version summarized here represents his last formal efforts. In his accounts he postulated two major sources of instinctual energy: *Eros* (so-called life-giving instincts) and what he referred to as death or destructive instincts (sometimes collectively called *Thanatos*). These instincts were said to be located in a presumed, large unconscious part of the personality known as the id, from which other parts of the personality (what he named as the ego and the superego) were held to emerge. The personality was portrayed as providing a structure for the discharge of the instinctual energy, but the instincts themselves had directional components as suggested by their names.

Freud postulated death-destructive instincts in the light of observations of a general prevalence of aggressive and suicidal acts among his patients and in the rest of the world about him. He seems, however, never to have set forth the detailed nature of these instincts as clearly as he did for Eros. (In passing, it may be noted that other psychoanalysts have not always accepted the instinctive character either of Thanatos or of Eros. Some of them have indicated that the German word *Trieb* in Freud's writings could have been translated by the term drive as well as by the word instinct.)

Any infant was seen by Freud as containing a relatively undifferentiated amount of so-called psychic energy (named libido) that, in the course of each child's development, underwent successive changes in the parts of the body where it achieved the most pleasurable discharge. Psychosexual stages

Freud spoke of erogenous zones (the mouth, the anus, and the genital area); if all proceeded properly, the libido was first to be expressed through the mouth (oral stage), next via the anus (anal stage), and then, in the period of infantile sexuality, was to focus on the genitals (phallic stage). The phallic stage was held to be followed by a latency period during which sexual expression was at a minimum; the eventual genital stage, if reached, would occur in the psychosexually mature adult.

Freud characterized instinctual energy as requiring discharge and as accumulating if not vented. It seems obvious that there was a limit to the amount of energy that could be accumulated before discharge was necessary. (This formulation has been interpreted as following a hydraulic model.) Discharge was described as often being delayed, since the environment rarely can be so arranged as to meet the young child's needs immediately. Hence, frustrations in dealing with the environment were expected to be experienced by the child at a very early stage in life. These frustrations are seen by some psychoanalytic theorists as essential for initiating the development of the ego.

The unbridled discharge of libidinal energy at any stage (especially in the Victorian society of Freud's middle-class, turn-of-the-century Vienna) was described as meeting with disapproval and punishment. Yet, it was felt, discharge must be arranged, and another reason for the development of the ego was theorized to make this possible.

The ego was depicted as a kind of executive of the personality structure. It was said to serve the id, tending to reduce or to prevent frustration by permitting discharge of libido while also carefully acting to disguise the ways in which the discharge was permitted so that punishment and disapproval would not follow. The ego could be said to follow what Freud called the reality principle in its operations, whereas the id was described as following only the pleasure principle. The ego was supposed to react by producing anxiety when punishment or disapproval were foreseen to be in the offing. Anxiety in Freudian theory is a major activity for keeping expressions of libidinal energy in check through the ego. Another part of the personality, the superego, was offered as an internalized set of parental (and other social) prohibitions and standards, having the ability to punish the ego and to induce anxiety via a threat to diminish self-esteem.

According to Freud the ego then must mediate among the demands of the id, prohibitions and threats of environmental reality, and the moralistic functions of the superego. The ego was presumed to employ a number of devices, sometimes called dynamisms, or defense mechanisms, that permit discharge of energy from the id yet disguise the nature of the discharge to avoid censure by the superego and the anxiety that this would bring. Among these so-called ego-defense mechanisms were named such dynamisms as repression, sublimation, reaction formation, regression, isolation, undoing, projection, introjection, turning against the self, and reversal. Two of these many different mechanisms can be sketched here as illustrative. Defense mechanisms

When a young boy reached the phallic phase, his libidi-

nal energies typically were to be directed to his mother as the object suitable for gratification of his phallic wishes. (Freud named this the Oedipus complex.) The boy was said to be threatened, however, by the fear of what his father would do to him (castration anxiety) were he to follow through sexually on his impulses toward his mother. This anxiety commonly was expected to be so great that to overcome it he was supposed to repress (forget, become unaware of) his impulses and the memories associated with them and to enter what Freud called the latency phase. Presumably, the maintenance of this severe repression (by which Freud accounted for the amnesia people commonly report concerning their first years of life) requires the expenditure of a great deal of energy. At any rate, the repression was seen as protecting the ego from the anxiety it would have without such unconscious forgetting.

The ego also was empowered to channel energy into behaviour that counteracted behaviour the id instincts would generate if left unmodified. For example, if expression of anal tendencies were the basic impulse, the ego was held to channel this energy into the behaviour of excessive cleanliness; *i.e.*, to employ the mechanism of reaction formation, thus avoiding the anxiety that direct expression of the anal impulse would engender, while still permitting some discharge of energy. When successful, all of the dynamisms of defense were understood to have the function of permitting discharge to a greater or lesser degree while avoiding anxiety.

This brief summary has been designed to underscore the all-important roles that Freud attributed to instinctual energy and to anxiety in the production and control of behaviour. There are few theories of behaviour that have been so firmly committed to and dependent upon motivational determinants as Freud's. Later psychoanalysts, as well as some of Freud's contemporaries, such as C.G. Jung, O. Rank, and A. Adler, disagreed with Freud in many respects. But the emphasis on unconscious motivational determinants of conduct remained a central core in the thinking of most of them.

Drive theory. It will be recalled that Woodworth characterized drive as the force of energy that activates the mechanisms that subserve behaviour. As the term instinct began to fall from fashionable use in the 1920s, "drive" emerged as the major motivational concept of the functionalist tradition in psychology. There were two primary interpretations of drive. The first was that of drive as an internal, bodily stimulus condition epitomized in the theory of the U.S. physiologist Walter B. Cannon (1871–1945) that the experience of hunger arises from the contractions of the empty stomach and that the feeling of thirst stems from dryness in the mouth and throat; other drives (such as that said to energize sexual behaviour) were understood to have similar sources in local stimulation involving appropriate bodily structures. The organism was described as being goaded into regulatory acts by these uncomfortable stimulus conditions, the relief of which was to occur when consummatory behaviour took place. This formulation clearly emphasizes internal sources and has a homeostatic character; and it was actually Cannon who coined the term homeostasis.

While a substantial body of empirical data supported Cannon's local, peripheral conception of drive, there were many contrary results as well, so that the theory began to lose favour. Behaviour that commonly is interpreted to signify hunger persists in experimental animals whose stomachs have been removed or have been denervated; similar findings are obtainable for evidence of sexual behaviour and for signs of thirst. Human beings, for example, may report that their feelings of hunger persist even when the stomach remains full of food while passage to the small intestine temporarily is blocked. The hunger experience persists with no notable change according to ulcer patients who have been vagotomized; that is, those in whom a branch of the vagus nerve has been cut to alter the actions of the stomach. Men who suffer desert thirst persist in drinking even after wetting their mouths and throats. Sexual desire continues to be reported by humans who have undergone castration and other operations involving the genitals (including denervation). Although there are clear experimental grounds for attributing partial validity to the theory, evidence of these kinds patently is incompatible with it.

Drive as a central state

A second interpretation characterized drive as a central state. The basic idea here was that in some way depletion or (in other instances) repletion of body tissues sets up more or less diffuse internal conditions that instigate the organism into activity. Such activity was seen as regulatory, since it was thought that the more active an animal is the more likely he is to find the goal objects necessary to relieve the internal tension arising from depletion or repletion of tissues.

Several types of evidence were compatible with this interpretation. For example, as the length of time that an organism is deprived of food or water increases, eating or drinking is observed to increase in amount and rate when food or water becomes available. Similarly, animals tend to cross an electrified grid to reach an appropriate incentive more frequently after prolonged deprivation (within limits) than after short periods of deprivation. Hormonal states seem to be important to the readiness with which some activities are likely to occur in many species; seasonal sexual behaviour, heat in females during only a part of the sexual cycle, seasonal migrations, and maternal behaviour are examples. These suggest a diffuse, internal, hormonal basis for the instigation of the activities in question. When animals have food available at all times, periodicity in eating also suggests internal regulation, as does evidence for such specific hungers as in the preference and heightened intake shown for dietary substances (*e.g.*, sodium salt compounds), of which deficiencies have been established experimentally.

It is understood that in all of these instances, an incentive (food, water, a potential mate, for example) is present at the time the behaviour indicative of drive occurs. Another source of evidence for the central-state interpretation is general activity, studied extensively by the U.S. physiologist Curt P. Richter and his students who found, for example, an activity rhythm in laboratory rats over 24-hour periods. Activity was low following feeding, remained low for some hours, and increased to a peak as the next feeding time approached. This rhythm clearly seemed to support the notion that internal factors instigate nonspecific activity; the general principle seems to be that as drive mounts activity increases.

Internal conditions, then, leading to homeostatically regulatory behaviour, were the prime notion of drive theory (as far as biological needs or drives are concerned). It was recognized readily that unmodified biological drives alone were an insufficient basis to account for all of animal motivation, and the concept of secondary or learned drives was added to complete the picture. In the development of a learned drive, a stimulus situation (or, particularly in humans, a symbolic process) was thought to gain motivating power by becoming associated through experience with a biological drive state. As an example, consider the learning of fear among laboratory animals. A two-compartment box that has a hurdle between its two sides may be employed. If a rat is placed in one compartment and shock is applied to his feet by electrifying the floor, he is quite prone to escape to the other (shock-free) compartment. If the administration of the shock is signalled ahead of time by flashing a light or by a buzzer tone, the animal will begin to learn to avoid the shock by jumping into the electrically safe compartment before the shock is given. Learning to escape the shocks presumably is motivated by the rat's painful experiences, but learning to avoid them is inferred to be motivated by the animal's anticipation of the pain, this posited anticipation being referred to as fear. In this situation fear is a learned drive, and the anticipatory avoidance developed by the animal is quite persistent. It was believed that the addition of many such learned drives based on a full range of biological motives (such as hunger) would help provide a full account of the complexities of human motivation.

As drive theory developed and was incorporated into a general behaviour theory, it was suggested that a drive contains two components, only one of which was consid-

ered motivational. One component was identified as embracing the stimuli associated with the drive and as being correlated with such factors as the local sources of the stimuli and the length of time the organism is deprived of consummatory behaviour. The other component was the energizing factor; *i.e.*, it was posited that any internal state corresponding to drive somehow could prime into readiness any learned habit or unlearned behaviour of which an animal was capable. The drive stimuli were thought to impart selective direction to behaviour, but this function was said to be associative rather than motivational. The energizing function, however, was conceived as being motivational and nonspecific; it would prime for action any and all behavioral tendencies. Thus a particular habit originally may have been learned under the impetus of one drive but was held to be subject thereafter to activation by other drives as well. In this form of drive theory, drive was described as being internal and was homeostatically conceived but was not assigned a directional function for behaviour; direction was attributed either to habit or to unlearned stimulus-response connections.

In this formulation the promptings of drive were held to be stilled by consummatory behaviour, and activities followed by such consummatory behaviour were represented as what animals learned. In other words, learning was supposed to occur as a consequence of reward or reinforcement, conceived by theorists as drive reduction or tension relief.

Together with psychoanalytic motivation theory, to which it has some strong resemblance, drive theory dominated the study of motivation by North American psychologists from about 1920 to about 1950. After 1950 drive theory declined in influence. There are several reasons for the waning of the theory, and these reasons have led to the suggestion of alternative theoretical views of motivation. The summaries of alternative accounts in the following sections incorporate some of the reasons for the decline of drive theory.

Arousal theory. Arousal or activation theory came to prominence in the early 1950s. It arose primarily from studies of emotional behaviour in conjunction with selected neurological findings. A number of other discoveries or findings that appeared at about the same time, independently of arousal theory, not only offered challenges to drive theory but also seemed to fit well with the notion of arousal theory.

The electroencephalogram (EEG) is a record of the electrical activity of the brain. Such records for humans and other animals exhibit a number of patterns. Readily identifiable variations in these electrical rhythms were interpreted as reflecting inner states by a number of investigators. For example, it was suggested that a pattern of rapid, low-amplitude fluctuations in electrical potentials recorded from humans and other mammals corresponds with alert, excited, and emotional states; regular, slower, and larger amplitude fluctuations in potential tend to appear when a subject seems to be in a relaxed state; even larger and slower fluctuations are likely to be observed when a subject shows signs of sleep. The suggestion was that there is a continuum covering the range from low to high arousal, and that several different levels of arousal could be inferred from their characteristic patterns of EEG activity. It had been suggested that emotional states correspond to a relatively high degree of arousal. It also had been proposed that behavioral efficiency is greatest at intermediate levels of arousal, being low both at the relaxed-sleep end and at the excited or emotional extreme. Evidence accumulated showed that, to a substantial degree, arousal of the cortex (outer surface) of the brain was a response to activity in the ascending fibres of a structure (the reticular formation) found in the brain stem. Stimulation or destruction of these reticular activating tissues was found to produce signs of arousal or of a decline in the bioelectric evidences of cortical arousal. Although some data were obtained empirically that behavioral performance was more effective at intermediate levels of arousal than at the extremes, this evidence was neither very clear nor unequivocal. It further was posited that intermediate levels of arousal would be perceived by the individual as desirable or pleasant, so that both increases from low levels of arousal as well as decreases from high levels of arousal could be rewarding. The incorporation of the idea that increase in arousal would be a rewarding state clearly seems inconsistent with theories that hold all reinforcement in learning to require drive reduction or tension relief.

Empirical discrepancies with drive theory

Meanwhile, a number of findings were being reported that were interpreted as posing difficulties for homeostatic drive theory. One of these groups of data provided evidence for curiosity as manifested in the behaviour of a variety of animals. Thus, animals in the laboratory were found to explore their environments, to manipulate objects, to evince orientation and attention to novel stimuli, and to exhibit preferences for situations and objects that the observers judged to be novel or unfamiliar. In all of these cases, there seemed to be no obvious way in which these behaviours could be attributed to the operation of the postulated primary (or biological), homeostatic drives. Further, it was shown that animals would learn when the reward seemed to be simply an opportunity to explore or to encounter novelty, a kind of reinforcement that is not readily attributable to tension reduction. Direct, experimental investigation also showed that moderate increases in the degree of stimulation often have reinforcing effects. One consequence of this work was that postulated new drives such as curiosity and exploration were added to the list. Theorists also were led to question whether reductions in drive and tension were necessary conditions for reinforcement in learning.

An additional set of evidence was obtained from laboratory work with human subjects who had volunteered to be confined under conditions of sensory deprivation. For example, they might stay in a small, sound-deadened room, wearing goggles and special cuffs over their limbs to further reduce the amount of sensory stimulation they received, and they were given nothing in particular to do. Environmental stimuli were substantially eliminated under these conditions, and, according to some interpretations of drive theory and psychoanalytic motivation theory, it might have been expected that the subjects would enjoy this ostensibly tension-free situation. (Indeed, typically they were being paid well for their service as subjects.) Yet, after a time (usually within a day or so), most of them had found the situation intolerable to the point that they urgently left to find higher levels of sensory stimulation. Many said they experienced hallucinations and showed other abnormal symptoms while confined under these circumstances. The results were interpreted to mean that people have a need for stimulation, perhaps at some optimal level.

Other findings during the 1950s that did not seem to accord well with a tension-reduction view of reinforcement included the discovery of so-called pleasure centres or reward centres in the brain. Electrodes were permanently implanted in various parts of the brains of such animals as rats, and stimulation from the electrodes in some of the tissues served to reinforce and maintain behaviour on the part of the animals. A rat would work to the point of exhaustion to obtain the stimulation. Other parts of the brain that seem to produce unpleasant or aversive responses under electrical stimulation also were identified.

While there is a similarity between the concept of states of arousal and the nonspecific drive as posited in drive theory, arousal theorists proposed a much more inclusive set of antecedent conditions. According to drive theory, on the other hand, time of deprivation is the chief factor for manipulation in eliciting motivated behaviour. There is an unquestionable degree of vagueness in arousal theory, since the state of arousal is not tied to specific antecedent circumstances. Nevertheless, arousal theory in some form has had a wide appeal and serves to provide a popularly acceptable basis for interpreting stimulation and information seeking, curiosity, and the like as motivational factors. It stresses factors that are external rather than internal, and to the extent that the theory is homeostatic, stimulation is the variable that is held to be

maintained at an optimal level. Arousal does seem to be offered as a general, energizing state, rather than as acting to direct the organism's behaviour.

Incentive theory. It long has been observed that incentives (that is, goal objects such as food or sexual partners) exert a powerful influence in behaviour. For example, laboratory rats were found to perform better in maze learning when particular incentives were used than when others were offered (efforts being made to hold drive constant). Selective preferences among food substances are manifested by these animals, even when they are not judged to be hungry.

Incentive theory came to prominence in the 1960s on the basis of evidence that appeared to be embarrassing to drive theory. Though major support for drive theory had been found in an observed increase in an animal's general activity as the duration of deprivation was prolonged, clear evidence now available indicates that the relationship is the result of functions that are neither automatic nor innate but learned.

In the animal experiments conducted through which the relationship was demonstrated, feeding occurred at predictable points in time. When animals are deprived of food for the first time (say, for three days) activity does not appear to increase substantially; but when subjects are fed regularly every 24 hours they appear gradually to become active in anticipation of feeding time (*i.e.*, apparently they learn to anticipate the appearance of the goal object or incentive).

Other evidence that clearly fails to conform to drive theory and that points to external incentives comes from studies made of such hormonally based activity as sexual behaviour. For example, male rats show no sexual behaviour (even though they are hormonally primed for it) in the absence of a female in heat. Even in her presence a considerable amount of mutual stimulation seems to be necessary before the threshold for actual copulatory behaviour is reached. Consummatory behaviour and crossing an electrified grid to reach an incentive require the presence of exhibited incentives. Such behaviour is not reliably elicited when the incentive is obstructed or screened from view. Specific hungers (*e.g.*, under calcium deprivation) also do not function as automatically and as efficiently in regulating dietary imbalances as earlier research had suggested. There is evidence for apparently inappropriate choice of foodstuffs under some circumstances. Periodicity in feeding, reported for groups of animals when food is always available, has been demonstrated to hold only when the average behaviour of a group is considered. Episodes of eating and noneating in individual animals were observed to be ordered randomly rather than periodically.

Incentive theory in the 1970s was not fully developed. It relied on a general process of arousal or activation that was seen to be instigated by incentives. The U.S. psychologists Charles N. Cofer and Mortimer H. Appley suggested two mechanisms in describing such instigation. In their anticipation-invigoration mechanism (AIM), animals learn to be aroused by the presence of or the imminent occurrence of a given incentive. In their sensitization-invigoration mechanism (SIM), which applies to behaviour not obviously learned, arousal appears (as in sexual behaviour) in the presence of an incentive and a suitable hormonal condition. In some way not yet understood, the latter is held to sensitize the animal to be responsive to the incentive object without learning. Cofer and Appley identified the motivational aspect of the incentive mechanism as being the general arousal, holding that direction of behaviour is supplied by the stimulus conditions of the situation through learning or innate relationships. (These conditions include the incentive itself.) External factors possess an obvious importance in this analysis and, in general, some optimal level of arousal would be postulated.

Hedonic theory. Hedonic theory is similar to arousal theory and incentive theory in rejecting the notion of drive and in emphasizing external objects and situations. It differs from them in specifying that the arousal is affective: pleasant, unpleasant, or neutral. In hedonic terms, behaviour is conceived as being ordered on a continuum of approach and withdrawal. Approach is described as being manifested toward goals, situations, or outcomes that are anticipated by the organism as pleasant, avoidance being shown when the individual's anticipation of contact with the incentive condition is that it will be unpleasant. The theory reinstates the principle of hedonism, long considered significant in the control of conduct (*e.g.*, Epicurus).

The idea was offered that pleasantness and unpleasantness represent departures from a level of adaptation to prevailing conditions, small deviations being perceived as pleasant and large discrepancies being experienced as unpleasant. The concept formalized as adaptation level (AL) was interpreted to mean that at any given time either an individual has adapted or must adapt himself to past and prevailing levels of stimulation. Movement toward or away from specific objects is held to change the individual's AL. Given an AL for some situation, then, one view was that motivation of approach or avoidance will arise on the basis of one's anticipated deviations from that level if he moves with respect to the objects.

This hedonic theory represents a break with drive theory; it was distinguished by its proposition that all motives are learned and by its reintroduction of affective (emotional) processes as factors in motivation. The theory did not generate a great deal of empirical work and seemed difficult to test unequivocally. One problem was with its definition of the conditions for pleasantness and unpleasantness; it seemed likely that these reactions are not ordered well by deviations from AL alone, but that they are determined by favourable or unfavourable outcomes, whatever the deviation from AL. Hedonic theory, however, clearly identified a learned, general arousal state, instigated by anticipation of contact with incentives. It was homeostatic in the sense that the AL is a homeostatic concept; in this respect hedonic theory resembled arousal theory.

Conclusions with respect to theory. Theories of motivation, as the foregoing survey reveals, were in a rather unsettled state in the 1970s. Drive theory seemed to have fallen, leaving no single, systematic alternative that had yet risen to replace it. The trend, however, seemed to be in the direction of emphasizing external factors (or their surrogates), toward reliance on some notion of arousal, and toward supporting a postulate that there is an optimal level of arousal. In addition to this incomplete picture of alternatives, further suggestions of developments will be made below.

In passing, it may be mentioned that psychoanalysis had changed, too. The direction of the change was to provide for ego functions that, while they may be said to use the energy of the instincts, are, or may become, independent of the conflicts originally proposed as the reasons underlying their development. It would no longer be pertinent to say, as some psychoanalytical theorists once did, that a conception of reality would never develop without conflict and frustration. Ego structures, it now seemed, would develop anyway.

HUMAN MOTIVATION

With the exception of psychoanalysis, the viewpoints surveyed in the preceding sections constitute general theories. These systems are applicable in principle both to humans and to the behaviour of other animals, taking work with both human and infrahuman subjects as their empirical base and, in several cases, relying more on the latter than on studies of humans. Since it shares common elements with other theories (*e.g.*, drive theory), and, in view of its historical importance, psychoanalysis was included in the discussion above. Attention must also be given to formulations developed largely in conjunction with human beings, ideas that do not seem to have the general sweep of the theories already reviewed.

In general, these more limited notions may be divided into two classes; one focusses on specific human motives or needs; the other looks to situational factors or to their interaction with need factors as providing motivational support for human behaviour.

Needs and motives. In 1938 the U.S. psychologist Henry A. Murray published a catalogue of human motives, or what he called needs. These needs were distinguished as being either primary (biological, innate) or secondary (learned, acquired); they were seen as impelling behaviour and giving it a goal-related cast. Among the motives that Murray discussed were what he identified as needs for achievement, affiliation, power, dependence, and succor (the need to be taken care of), these being selected from among many more in the list.

Work on the needs listed by Murray largely has been limited to those called achievement, affiliation, and power; inquiry also has been directed selectively to other motives such as anxiety and aggression. Anxiety has been studied both in its own right and as a factor that might underlie other apparent needs, such as affiliation. Brief accounts will be given here of needs that Murray labelled achievement, anxiety, and aggression.

Achievement. By the 1970s achievement had been explored more fully than any of the others had been. The study of the need for achievement (abbreviated as nAch) can be considered under three headings: measurement, relationship to behaviour, and development.

Measurement: A kind of quantification of nAch was accomplished by having people write a story in response to each of several pictures (selected from Murray's Thematic Apperception Test, or TAT) and scoring the stories for imagery of achievement and for achievement-related themes. Achievement was scored when a story gave evidence that the person felt concern over competition against some standard of excellence. The scoring system was empirically based; *i.e.*, the elements to be scored were those features that had been found to distinguish stories written under experimental conditions designed to arouse competition with a standard of excellence from stories produced under relaxed circumstances or other competition-free conditions. The scoring system is not perfect from a psychometric standpoint but does seem to have acceptable reliability. (That is, measures obtained from the same story by different scorers tend not to differ markedly.)

Thematic Apperception Test

Relationship to behaviour: People vary in their nAch scores under ordinary conditions; *i.e.*, when no attempt to arouse achievement motivation has been made. Scores reflecting nAch have been found to be related positively to performance on an anagrams test and an arithmetic test, although the degree of relationship has not always been high and the evidence has not always been consistent. Similar findings have been reported for the validity of nAch scores in predicting college grades, recognition times for achievement-related words presented for brief visual intervals, estimates of the individual's level of aspiration, ability to recall the names of tasks that one has and has not successfully completed, and measures of risk-taking behaviour (as in gambling games or tests of skill). Some of this work has shown that the nAch score reflects elements of positive achievement as well as those relating to the person's fear of failure. In risk-taking studies, for example, people with high nAch scores have been found prone to take moderate risks when their chances of success are reasonable; such people do not seem to be interested in choices for which the chance of accomplishment is practically assured or virtually nonexistent. Risk-taking behaviour among subjects who are judged to fear failure differs from this pattern; such people tend to choose alternatives for which success appears certain or for which success seems so unlikely that they can hardly be blamed for failure.

One outcome of efforts to relate nAch scores to different kinds of behaviour is the conclusion that such relationships have complex sets of determinants. For example, the need for achievement is not always manifested; the situation apparently has to be identifiable by the individual as one involving competition with a standard before the need will be aroused.

Development: Evidence has been offered that the need for achievement is learned, and its origins have been identified in a general emphasis on independence training in the early childhood of those who make high nAch scores. There remained some uncertainty in the 1970s, however, as to exactly which prior experiences are related to the development of nAch.

Anxiety. Anxiety has been given considerable importance in some of the theories of motivation discussed earlier, but it has been given extensive empirical attention in research conducted only in a limited number of contexts. One of these involves anxiety as a basis for affiliative behaviour. It has been reported that when subjects were threatened with electric shock certain female subjects (those who were first-born children or only children) were prone to seek an opportunity to spend a waiting period in the company of others; later born subjects were markedly less likely to show this affiliative behaviour. Thus, at least among some people under given conditions, fear or anxiety seems to motivate such forms of social behaviour, and it does not appear necessary to postulate an affiliation motive in such cases.

In other contexts anxiety has been studied by making quantitative estimates of the levels of anxiety among people and then comparing the performance on selected tasks of highly anxious subjects with the task behaviour of those found low in anxiety. One measuring instrument for this purpose is the Manifest Anxiety Scale, which includes true-false statements concerning symptoms of anxiety; *e.g.*, My mind wanders; I have many headaches; I am easily embarrassed; I have frequent nightmares. Proponents of the view that anxiety is a drive have been especially prone to use this scale. These investigators tested people with the Manifest Anxiety Scale and then had them perform in other situations (such as classical conditioning routines and complex verbal learning tasks) in the expectation that level of anxiety would affect performance in accordance with theoretical predictions. The derivation of the predictions is technical and complex, but it can be noted that theoretical expectations usually have been verified in classical conditioning situations; results have been less consistent for the more complex tasks.

Another instrument, the Test Anxiety Questionnaire, also has been used to quantify anxious reactions under selected conditions. People judged to be highly anxious on the basis of this scale are reported to compare poorly with others on tasks in which they experience feelings of failure or threats of failure. On the other hand, the anxious subjects were observed to do as well as or better than did the so-called low-anxiety people in situations where there seemed to be no threat or in which they had experienced success.

Aggression. Aside from Freudian theory (and later contributions by such ethologists as Konrad Lorenz), most theorists have regarded aggression as a learned pattern for getting one's way. Hence, these theorists either do not consider such behaviour as expressing an aggressive drive or they have conceived of aggression as a learned drive expressive of anger or frustration. The frustration-aggression hypothesis (that frustration leads to aggression) is difficult to entertain as an accurate generalization since its validity is not always supported by observation. It appears that frustration does not lead to aggressive behaviour when the frustrating condition is not perceived by the individual as threatening, as emotion arousing, or as arbitrary. Neither is aggression the only consequence of frustration, and alternative responses among frustrated people have been observed. If aggressive behaviour has instrumental value in a situation or if it is frequently rewarded, it seems more likely to follow frustration. There are wide individual differences, however, and some people are said to have considerable tolerance for frustration, while others seem to have very little ability to resist frustrating stresses. Aggression may arise in a context of anger that is instigated by annoyances, insults, or attacks from others but tends to be short-lived under these conditions. Aggressive behaviour can be suppressed by punishment and augmented by permissiveness. Clear demonstrations of displacement of angry aggression (*e.g.*, the angry man who kicks his innocent dog) and of catharsis (emotional release) through angry aggression have yet to be made in the laboratory, although Freudian theory predicts that both should occur.

The psychological literature of aggression indicates the difficulty, as it does also for other human motives, of separating aggressive acts held to arise from anger from those said to occur because they are instrumental in reaching specific rewards. Anger-instigated aggression, however, would be the one of motivational significance as far as the possible existence of a motive of aggression is concerned. Yet, it is remarkable that the research literature offers only marginal support for the notion of an enduring motive of aggression based on anger, despite the heavy emphasis on aggressive expression in popular accounts of war, crime, and other social disturbances.

Aggressive behaviour, of course, need not require a specific motive for its appearance. Modern neurophysiological work may show (see EMOTION) that there are neural centres or more or less diffuse neural systems that mediate aggression. If their existence is substantiated, it could turn out that activation of aggressive behaviour depends less on anger than on properties of stimuli and situations that engage these structures in the nervous system.

A hierarchy of human motives. Three posited human motives, those to which major research attention has been devoted, have just been discussed. They have been considered, however, independently of one another. The question remains: are such motives organized in any way? The U.S. psychologist Abraham H. Maslow (1908–70) suggested the existence of a hierarchy of motives. In his formulation, the so-called physiological needs (such as hunger) were at the bottom of the hierarchy. Above them were the needs for safety and higher still were needs for belongingness and love; needs for esteem came next. At the top of the hierarchy he placed needs for self-actualization and for various cognitive and aesthetic goals. Maslow suggested that self-actualization was the ultimate and desirable goal but that it could not be reached until what he called lower needs were relatively well satisfied. Such lower needs were seen to dominate conduct unless they were satisfied or unless the individual felt no concern over their satisfaction, and it was held that higher needs cannot be expressed adequately in the unsatisfied presence of dominant lower needs. Of the motives discussed above, achievement probably should be classified as one of the esteem needs, and anxiety is probably to be related to the safety needs—that is, to a lack of safety. Maslow did not regard aggression as an innate need, as he did the others, holding it to be a reaction to frustration and to other conditions.

Maslow's hierarchy has had much appeal, but, although there are some data that support the idea that lower needs may dominate conduct at the expense of higher needs, there is little direct evidence for the notion. Maslow perceived the human personality as typically being caught in conformity to the demands of society. He wrote that such conformity is not necessary, proposing the idea of self-actualization as a condition that can be reached by everyone if he eschews the conformity brought upon him by the dominance of those needs at the lower levels of the hierarchy.

Situational and interactional factors. So far, this discussion has emphasized internal needs or motives (except in the case of incentives); this inner orientation has pervaded thinking about motivation. It is possible, however, to conceive motivation not as involving a list of stable motives but, rather, as embracing needs that arise from situation to situation. The theorizing of the German-U.S. psychologist Kurt Lewin (1890–1947) to a large extent represents this way of looking at motivation, and investigators that he influenced have stressed cognitive aspects of motivation. Lewin's orientation stemmed from Gestalt psychology, a movement in which the prime concern was with the process of perceiving.

Gestalt psychologists regarded perceptual experiences as the outcomes of a stimulus field in which there are stresses. An unbalanced visual figure, for example, was said to contain dynamic forces or stresses that would tend to make it be perceived or remembered in a regular, symmetrical, or balanced way. This is a dynamic theory, but the forces were posited not as arising from internal agencies (*e.g.*, the Freudian id) but from the stimulus field as imposed upon brain processes said to correspond (to be isomorphic) with the organization of the external, environmental field. In a sense Lewin applied these interpretations to the dynamics of action.

Lewin conceived of the person as existing from moment to moment in a life space with regions that could be in states of tension; when these states occurred, forces were supposed to be created that instigated symbolic or actual movement or attempts to move from one region to another. While the tension was depicted as being internal, it was held often to be created by the circumstances in the field at the moment, in interaction with the goals that the individual saw the situation as permitting him to reach. Thus, task completion was offered as one such goal, and failure to complete a task was expected to lead to a continuing tension that would make it easier to remember incomplete tasks than would be the case for completed tasks (this is the so-called Zeigarnik effect, in honour of its discoverer). It also was predicted that there would be a tendency for work on the uncompleted tasks to be resumed at the first opportunity. Lewin also pointed to the importance of the person's cognitive evaluation of any given situation. Much study was devoted to the individual's level of aspiration for performance on a number of tasks, the subject being asked to state at what level he hoped to achieve the next time he attempted to perform a task. It was found that subjects do not typically give perfect performance as their expectation; yet, if achievement of success is a major goal, chances are that they should do so. These empirical findings suggested to Lewin that goal setting is a complex process, being affected not only by the positive valence of success but also by the negative valence of failure. He proposed that subjects are affected by how they estimate the probability level at which they expect to succeed or to fail in performing at a given degree of effectiveness. Lewin added that the valences of success and failure are weighted by these subjective probabilities in determining the actual goal the person says he will strive to achieve.

To an extent, then, Lewin saw the tasks themselves as creating motivation, and he characterized behaviour as occurring because of imbalances in the stimulus field. He also depicted cognitive evaluations as interacting with valences in determining the behavioural outcomes in a given situation. Lewin was especially influential in outlining presumed conditions of conflict and in describing methods held likely to resolve them. His interest in conflict was an understandable one, since conflict of goals, for example, was depicted as creating imbalance in the field, the resulting behaviour tending to restore equilibrium.

A number of theories, especially in social psychology, appear to have been influenced by Lewin's conceptualizing. At any rate, these theories have little to say about standard motives but rather look to motivation as arising from imbalances in a situation. Balance theory and congruity theory are among the major viewpoints of this type. Discussion here will be limited to the most well-known system in this group: the theory of cognitive dissonance. In its career since the U.S. psychologist Leon Festinger proposed it in 1957, the theory of cognitive dissonance has stimulated a great deal of investigation. Although it is not without problems, this theory as presented here will serve to illustrate another way of looking at motivation.

Festinger proposed that a tension state (that seems to be much like a drive) would be created when two or more cognitive elements exist but are dissonant with one another in that they are perceived to be incompatible or inconsistent. For example, if it is supposed that one regards himself as an excellent golfer but that on a given day he earns an unusually poor score, there are two cognitive elements: the man's high estimation of his skill and the conflicting evidence of his relatively inept performance. The two elements are said to be dissonant, and dissonance was depicted as a state of tension that must be resolved. This situation assumes that the elements are related to one another, as the two in the example are, and that they are perceived to be important, as the elements involving golf would be in one who prides himself on his skill at playing this game.

Festinger applied his theory to dissonance that arises after one has made a decision. He saw the resolution of the dissonance in the individual's behaviour relative to the conflicting elements. If it is supposed one has just bought a new automobile after carefully considering two models and having perceived both of them to have positive and negative features, the positive features of the rejected alternative are said to be dissonant with the fact that it was rejected, the negative features of the vehicle that was purchased being dissonant with the fact that it was selected. A prediction that follows from the dissonance formulation is that, after the car was bought, one would be prone to read advertisements that stress its positive features and to avoid advertisements for the rejected automobile, and to do both for the same reason. Partial support was obtained in a study conducted to test this prediction. In another experiment, college girls were invited to participate in group discussions concerning sexual behaviour. As a precondition for participation, some of these girls underwent what was designed to be an embarrassing interview; another group of them were interviewed under comparatively bland conditions. Both groups then listened to a taped recording of what most people would judge to be a dull discussion of sexual behaviour among subhuman animals. The girls who had gone through the embarrassing preconditioning tended to rate the discussion as much more interesting than did those who had been subjected to the neutral preparatory interview. The cognitive elements here are those of the embarrassment of the initial conditioning procedure and the dullness of the ensuing discussion; they are interpreted as being dissonant. According to Festinger's theory, the embarrassed girls evaluated the discussion favourably to resolve the cognitive dissonance.

A great deal of research that has been generated by the theory of cognitive dissonance has tended to support the principles that Festinger enunciated. The lustre of this theory, however, seemed to fade in the years following the first surge of interest it attracted. This dimming may be attributable to difficulties that arise in defining and measuring dissonance and in the rather elaborate and contrived experiments that had to be undertaken in efforts to test the theoretical formulations. Results of a number of the experiments failed to replicate earlier findings. Although the theory was not regarded as favourably in the 1970s as it once was, it clearly was most provocative and stimulated a wealth of empirical work. It exemplified an approach to the study of human motivation that is rather different from those theories in which there is a dependence on the controversial existence of relatively permanent or stable motives or needs.

LATER DEVELOPMENTS

A discussion of motivation written in 1950 in all likelihood would have been more positive in its evaluation of the status of the field and of the systematic integrity of its concepts than one written in the 1970s. By 1970 motivation theory was in a state of flux, partly in view of a decline in the cogency and acceptability of the notions of drive and acquired drive. There was a growing uncertainty that motivational concepts were needed at all. For example, it has been suggested that notions of motivation may not be needed in accounting for behaviour; that the phenomena popularly listed under the rubric motivation can be encompassed readily by considering the operation of the general phenomenon of learning called reinforcement.

Reinforcement refers to an increase in the probability that an organism will exhibit a given kind of behaviour (in a situational context) through some consequence of the behaviour itself. It is well known that laboratory animals will learn mazes, solve puzzles, and develop the ability to discriminate among sensory stimuli, for example, if they are rewarded (reinforced) for such behaviour with food if they seem hungry, with water if they appear thirsty, and so on. Learned rewards (*e.g.*, approbation) also have been found to be effective and may be supposed to function as the typical reinforcers for human learning. It was tentatively suggested that close analysis of the conditions under which reinforcement occurs will reveal these conditions (rather than motives or incentives) to be the factors that control the behaviour. It was too early at that time for a rigorous evaluation of this notion, although a number of investigators were not optimistic concerning its chances for success. Reinforcement itself is a complex concept and is not well understood; it becomes a nice question whether some condition of motivation is necessary for many reinforcers to be effective.

Another trend in the 1970s was seen in the re-emergence of the term instinct in the writings of ethologists. The ethologists distinguished appetitive behaviour (*e.g.*, copulation) from instinctive acts (*e.g.*, the courtship of some birds), a distinction somewhat like that made by Woodworth between consummatory and preparatory behaviours. They pointed to these species-specific, innately generated acts and to the stimuli that apparently can release such species-specific behaviour (or instinctive acts) without the need for learning. Since much of the work of ethologists concerned birds and lower mammals, the extent to which their conclusions could be generalized to other species was uncertain. They emphasized such factors as territoriality (as observed in birds that drive interlopers away) in the determination of behaviour. While some ethologists seemed willing to extrapolate their conclusions to human beings, others were more conservative. In general, however, it did seem that stress on innate determiners of behaviour had grown and received acceptance since about 1960, in contrast to the radical environmentalism of earlier decades. It should be added that the evidence offered for innate factors in behaviour was much more specific and was obtained under much more carefully controlled conditions than was the case in McDougall's time during the earlier period of the dominance of the instinct concept. The instincts postulated in the 1970s also were to be understood as behavioral acts rather than as urges or purposes. It seemed more modest or objective to speak of unlearned acts and their innate determinants than to refer to such activities as reflecting instincts.

PRACTICAL APPLICATIONS

Despite the uncertainties that have been indicated concerning definitions and the theoretical status of motivational concepts, what might be called a technology of motivation has developed. There is, however, little theory that may be extended in socially or economically useful ways. Indeed, applied efforts that are successful may turn out to have desirable effects for reasons that are not motivational, despite their having been undertaken in the name of motivation.

In the control of animal behaviour, for example, it is clear that depriving an organism of food is a powerful means for accomplishing reinforcement. Appropriate use of food under these circumstances is an effective procedure for shaping an animal's behaviour, maintaining it, and controlling the rate of its occurrence. This is not to say that any act can be controlled by such procedures; such methods, however, are widely applicable. Likewise, it is clear that animals have preferences (within, for example, the range of foodstuffs) and that their behaviour can be controlled with relatively greater effectiveness by the proper selection of preferred substances for use in training.

In many cultures, deprivation cannot be used so readily with human beings as it can be with other animals, although there are many human examples. Thus, some success has been reported in effecting desired behaviour in the classroom by depriving children of some of their recess time when they behave in ways deemed undesirable by the school authorities. Economies based on the use of tokens (*e.g.*, poker chips) have been set up in wards in psychiatric hospitals and in institutions for retarded people. The result typically has been an improvement in the inmates' behaviour and personal care and in the ease with which they may be managed. In such economies, tokens can be exchanged for privileges and commodities (*e.g.*, candy, toys). The individual's ability to obtain tokens is made contingent on socially desirable acts, such as mak-

ing beds, being personally clean, being cooperative, and generally acceptable to others. There have been reports of marked improvement in scholastic achievement among institutionalized juvenile delinquents who have been placed in such token economies.

The effectiveness of these and similar procedures has been most easily demonstrated in institutions, in which the situation permits a great deal of control over the inmate's conditions of life and over his activities. In society at large, of course, this degree of control is effectively not feasible. There also are widely endorsed moral or ethical concerns about the desirability of instituting such control even if it were possible. The use of particular kinds of motivational devices in the control of human behaviour seems to many to be incompatible with the ethical idea of personal freedom and to be fraught with potentials for immoral misuse in the hands of those who seek to manipulate others for ends that are politically or socially conformist. To some, the disarray of motivational concepts and the constraints that limit the effectiveness of such motivational devices as advertising and political propaganda in society at large come as welcome antidotes to rising concern. Hopefully, the development of society, as grimly outlined in George Orwell's novel *Nineteen Eighty-Four* or in Aldous Huxley's *Brave New World*, may not come to pass as readily as many people have feared.

On the other hand, it often is observed that many of society's problems are motivational. This observation usually means that the goals and values of economically affluent groups in the Americas, Europe, and Asia are not shared by members of deprived urban ghetto populations or by millions of poor people in industrially less developed countries. Or, it may mean that the goals of those who own or control profit-seeking enterprises (to make a product or deliver a service for the investors' profit) are not shared by workers below the level of middle management. Many techniques have been tried both in capitalistic and socialistic forms of business and industry to effect so-called motivational involvement with production on the part of ordinary employees. Some of them have had success, in Soviet tractor factories as well as in West German automobile plants. Incentive systems, employee participation in company planning and decisions, and human-relations training exemplify the procedures used. A substantial corps of specialists throughout the world provides programs to industry designed to improve the motivation, morale, and satisfaction of workers at all levels. Although these programs have wide acceptance, most of them have received very little objective evaluation. All of them must, of necessity, be designed ultimately to orient the worker to identify his interest, goals, and values with those of management. As complex a process as such identification is, it no doubt contains motivational elements.

BIBLIOGRAPHY. The most comprehensive treatment of the topic is the one by C.N. COFER and M.H. APPLEY, *Motivation: Theory and Research* (1964), although R.C. BOLLES, *Theory of Motivation* (1967), gives an excellent historical introduction and offers a thorough analysis of most of the theoretical issues. R.S. PETERS, *The Concept of Motivation* (1958), presents a rationalistic approach to motivation; V.H. VROOM, *Work and Motivation* (1964), surveys the industrial applications. R.A. HINDE, *Animal Behavior: A Synthesis of Ethology and Comparative Psychology*, 2nd ed. (1970), summarizes the ethological position; B.A. CAMPBELL and J.R. MISANIN review recent literature on special topics in "Basic Drives," *Ann. Rev. Psychol.*, 20:57–84 (1969). Work on token economies is indicated by T. AYLLON and N. AZRIN in *The Token Economy: A Motivational System for Therapy and Rehabilitation* (1968). Specific human motives are discussed in J.W. ATKINSON and N.T. FEATHER (eds.), *A Theory of Achievement Motivation* (1966); and by L. BERKOWITZ in *Aggression: A Social Psychological Analysis* (1962). A comprehensive source for balance theories, including the theory of cognitive dissonance, is *Theories of Cognitive Consistency: A Sourcebook* (1969), ed. by R.P. ABELSON *et al.*

(C.N.C.)

Motor Sports

Motor sports cover a multitude of different types of events, the majority of which are races. The basic idea of a race is, of course, simple: drivers of automobiles or motorcycles run as fast as they can over a given course. The man and machine combination that has covered the prescribed distance in the shortest time (or covered the greatest distance within the prescribed time) wins.

Rules and regulations for competitions are made by international governing bodies, national sanctioning bodies, or both, and supplementary regulations are issued by the organizers of a given individual event. The rules spell out in great detail the type of vehicle eligible to race so as to assure equality among the competitors. There may be restrictions on engine type and size, body dimensions, fuel supply, number of driving wheels, and number of vehicles built to the same specifications. Though automobile racing and motorcycle racing are under completely different rules, passed by different organizations, both types of vehicles often race at the same places, and occasionally, in competition with each other.

Types of races and other motor-sports events

There are several types of races held on a great diversity of racing sites, or venues. Road races are staged on public roads (closed to the public for the occasion) or on simulated road circuits. Other races are held on speedway tracks, mostly oval, with or without banked turns. Acceleration races are held on straight and level strips called drag strips. Certain kinds of races are held in rough terrain (off-road racing). Speed-record events are held on beaches or the beds of dry lakes.

Motor-sports events that are not races include rallies, of which there are numerous categories. Driving-skill tests (gymkhanas) and historical car meets are also considered motor-sports events.

For speed records and records of principal automobile and motorcycle-racing events and championships see also SPORTING RECORD in the *Ready Reference and Index*.

Popularity. Motor sports constitute one of the most popular spectator sports in the world. Crowds of 300,000 and more have been claimed for such automobile racing events as the Le Mans (France) 24-hour Grand Prix d'Endurance and the Indianapolis (Indiana) 500 Miles Sweepstakes. In the United States, statistics show auto racing to have higher attendance than football and baseball, running second in popularity only to horse racing. Participation in motor sports has increased to new proportions since the end of World War II, and enthusiasts worldwide have formed clubs and stage local events.

The racing calendar is overcrowded. Ranking international events compete for the fixed dates of their choice, and sanctioning bodies have a hard task trying to accommodate all interests. On any given weekend there are countless minor car and motorcycle races, rallies, and other events. Many of these do not cater to spectators but are organized for the participants alone.

This article is divided into the following sections:

I. History of motor sports
  Motoring and auto racing
    Development of Grand Prix racing
    Development of speedway racing
    Development of endurance racing
    Evolution of hill climbs
    Evolution of drag racing
    Evolution of rallies
  Motorcycling
II. Modern automobile sports
  Racing
    Organization of the sport
    Classes and events
  Rallies
  Hill climbs and trials
    Hill climbing
    Hill trials
  Speed-record trials
  Other events
    Gymkhanas
    Historic (Classic or Antique) car meets
III. Modern motorcycle sports
  Motorcycle racing
    Governing bodies
    Classes and events
  Other events
    Rallies
    Hill climbs
    Speed-record trials

I. History of motor sports

MOTORING AND AUTO RACING

First automobile races

Motor sports began at 8 AM, July 22, 1894, when 21 cars set off from Paris to Rouen, France, in a reliability test run organized by the *Petit Journal* newspaper. The first true automobile race was held on June 11, 1895, over a 732-mile (1,178-kilometre) route from Paris to Bordeaux, France. Town-to-town French races on public roads became the norm as other events followed: Paris–Marseille–Paris, Paris–Dieppe, Paris–Trouville, Paris–Amsterdam–Paris. The Course de Périgueux, the first race held on a closed-road circuit, took place on May 1, 1898. The circuit was 90 miles (145 kilometres) long, and competitors covered only one lap.

Other organizers took up the idea and circuits were improvised in the Ardennes Mountains and in the Auvergne Mountains. Town-to-town events dominated the sport until 1903, when the high number of accidents occurring during the 1903 Paris–Madrid race caused it to be halted at Bordeaux. After that, road racing became confined almost entirely to closed circuits.

The Automobile Club de France defined racing rules and the classification of cars and functioned as a governing body for the sport in its formative period, but the man who did most to make motor racing international was an American, James Gordon Bennett, owner of the *New York Herald* newspaper. Automobile racing in North America started on November 28, 1895 with a contest over a 54-mile (87-kilometre) road course near Chicago, organized by the *Chicago Times Herald*.

In 1899 Bennett presented a trophy for a race, not between the drivers or the manufacturers but between representatives of the national automobile clubs, using cars built entirely from parts manufactured in the country represented. Each nation could enter a maximum of three cars. The Automobile Club de France was to organize the first Gordon Bennett Cup Race. As other countries became more active in motor racing, the Gordon Bennett went to the Circuit of Ireland in 1903, the Taunus Rundstrecke in Germany in 1904, and to France's Circuit d'Auvergne in 1905.

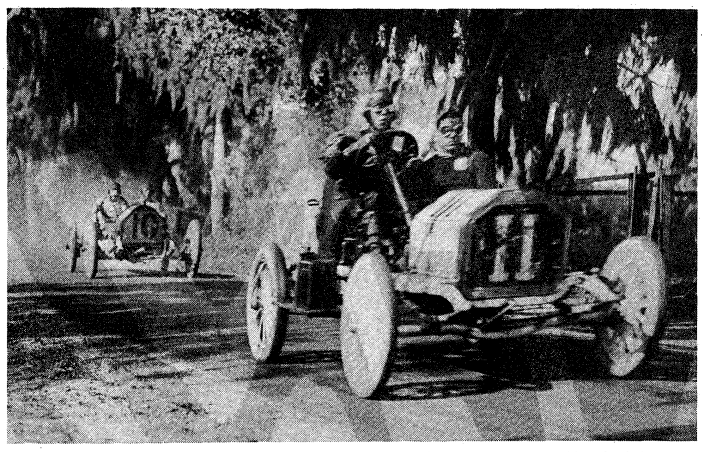
By 1906, the French automobile industry was no longer willing to be restricted to only three cars as was required by the Gordon Bennett races, and a new type of event with no limitations on the number of cars that might be entered from any nation was inaugurated. The new event was known as a Grand Prix (see below) and the first Grand Prix race was held that year at Le Mans.

During the 1906–14 period, the French Grand Prix was the main event of the year, and no other nations held Grand Prix races during this period.

The Vanderbilt Cup and racing in America

William K. Vanderbilt, Jr., was the next American to offer a motor-racing trophy. The first Vanderbilt Cup Race was held on Long Island, New York, in 1904. The rules were patterned on French practice; and America became an important factor in road racing, although at first most competing cars were imported.

PHOTOWORLD—FPG



Two of the competitors in an early U.S. road race, the Vanderbilt Cup Race of 1906.

The Vanderbilt Cup Race was held annually on Long Island until 1911, when it was moved to Savannah, Georgia; then to Milwaukee, Wisconsin; then to Santa Monica, California. The 1915 race took place in San Francisco, and the 1916 event was back at Santa Monica. Then the cup was retired and not revived until 1936.

Italy moved into the forefront of international racing in 1906 when the first Targa Florio was organized in Sicily by Conte Vincenzo Florio. It was a 90-mile course in the hills of sparsely populated Sicily where crowd control was less of a problem than near the big cities. It was, and is, run on public roads (theoretically closed off) and has been renewed annually (with interruptions) to this day.

Up to World War I the same type of car was used in most races. The racing car began as a prototype for the following year's production model. Later, racing cars became too specialized for quantity production. At the same time, many high-performance touring cars were turned into racers by the removal of the bodywork and the fitting of special seats, tanks, and extra tires. After World War I, Grand Prix racing became a tightly regulated sport with strict rules, and sports-car racing and endurance racing gained in popularity. Other forms of racing also sprang up, including production car (stock-car) racing, maximum-speed or maximum-acceleration events (drag racing), hill climbs, and others, each catering for a particular type of vehicle.

Development of Grand Prix racing. In all of automobile racing, Grand Prix events are the most elite. Grand Prix automobiles are commonly the ultimate in design, speed, and horsepower and are regarded as the purest racing cars. They are painted in national colours (British green, Italian red, French blue, German white, U.S. blue and white, etc.). They are entered in events by their manufacturers or by private owners as teams, usually consisting of two or three cars, and driven by professional drivers. Races are held on closed circuits, usually three or four miles (five or six kilometres) to the lap, with the total distance varying from 150 to 400 miles (250 to 650 kilometres). All modern Grand Prix cars conform to one of various formulas or sets of specifications, laid down by the Fédération Internationale de l'Automobile (FIA), to govern the size, weight, and other specifications of the vehicles that may compete.

Grand Prix racing, interrupted by World War I, resumed in 1920 with an engine-displacement limit (intended to equalize power and promote technological advance). Engines became smaller and lighter, and cars became lower and faster. All leading industrial countries produced Grand Prix racing cars, though Italian and French makers dominated until 1934.

Up to 1925 a riding mechanic had occupied the passenger's seat. He had been needed in earlier days to pump oil, help change tires, and so on; but in 1925 the riding mechanic was outlawed.

European Grand Prix circuits and speedways

Most western European countries soon began to organize Grand Prix races on road circuits. Some built speedways. The Belgian Grand Prix was first held at the triangular Spa–Francorchamps course in a wooded Ardennes valley in 1925. Italy built its Monza track (outside Milan) in 1922, and France followed with Montlhéry (outside Paris) in 1924. Italy also organized Grand Prix races at Mellaha Circuit in Tripoli, Libya, from 1925 to 1940.

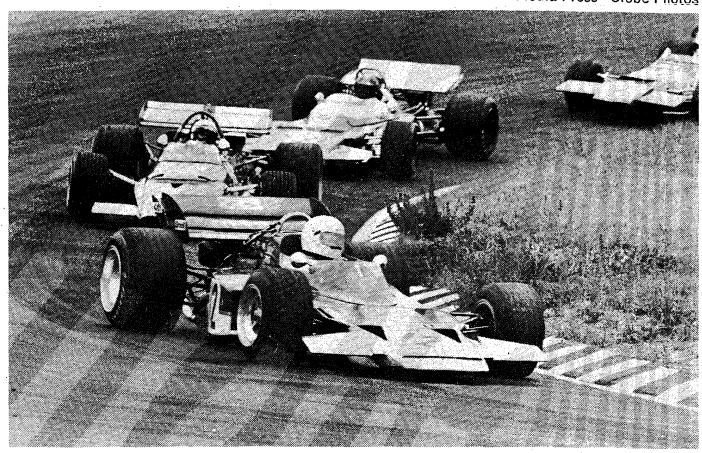
Germany created the Nürburgring road course in the Eifel Mountains as a make-work project to relieve unemployment in 1927. The circuit is 14.2 miles (22.8 kilometres) long and has an altitude span of 2,400 feet (about 700 metres). Germany also built a speedway between Potsdam and Berlin in 1926. It was called Avus (Automobil-Verkehr-und Übungs-Strasse, automobile test and training road).

Many local clubs organized races for Grand Prix cars through city streets. The most famous is Monaco (Monte Carlo), which began in 1929. Nice and Marseille followed. Others were Albi, Pau, Rome, San Remo, Turin, Cremona, and Florence.

The displacement limit, having proved detrimental to competition, was abandoned in 1928, and only a weight

limit (550 kilograms [1,200 pounds] minimum, 750 kilograms [1,650 pounds] maximum) was left in effect. The maximum-weight limit was abandoned for 1929–30, and the minimum raised to 900 kilograms (2,000 pounds).

Media Press—Globe Photos



Cars rounding a curve during the 1971 Grand Prix of The Netherlands

During the 1931–33 era, there was a free formula (no restrictions). From 1934 through 1937 there was a 750-kilogram maximum-weight limit that produced some of the fastest, most spectacular cars ever built. During the 1934–39 period German cars (Auto-Union and Mercedes-Benz) dominated completely. The 1938–39 formula reverted to a displacement limit. Racing resumed after World War II under a free formula. For the 1947–53 period, there was a new and still lower displacement limit.

The World Championship for drivers

The World Championship for drivers was instituted in 1950; it is based on points earned in major Formula I events (those open to the biggest, fastest cars) each year. The qualifying events, called *grandes épreuves*, are customarily the Grand Prix of Monaco, Belgium, The Netherlands, France, Great Britain, Germany, Italy, Mexico, South Africa, Canada, and the United States (Watkins Glen; the Indianapolis 500—for track-type cars—no longer qualifies for championship points). The first world champion was Giuseppe Farina (driving with the Alfa Romeo company team). An Argentinian, Juan Manuel Fangio, was champion five times (1951 with Alfa Romeo; 1954, Maserati and Mercedes-Benz; 1955, Mercedes-Benz; 1956, Ferrari; and 1957, Maserati). For the 1954–60 period, displacement limits went up for unsupercharged engines, and superchargers were further penalized. (From 1922 through 1937 no distinction had been made for superchargers; a sliding scale was introduced in 1938 and modified for later formulas.)

The 1961–65 Grand Prix formula again reduced the displacement limit. Speeds fell but racing-car design was revolutionized. In the course of this five-year period all racing-car builders who had not yet placed the engine in the middle of the chassis (behind the driver) did so. Beginning in 1966, engine displacement limits were doubled, and cars became larger and heavier again.

Development of speedway racing. The first speedway especially for automobile racing was constructed in 1907 at Brooklands, near Weybridge, Surrey. The Brooklands track consisted of a 2.766-mile (4.450-kilometre) circuit with a minimum width of 100 feet (30 metres) and two curves banked up to one in 1.6 and a height of 27½ feet (8½ metres). Brooklands became the scene of sprint races, relay races, endurance races, handicap races, and long-distance record runs. A 24-hour race was held in 1929, another in 1930, and another in 1931. These were followed by a 1,000-mile (1,600-kilometre) race in 1932.

The importance of Brooklands can be fully assessed only when it is realized that for many years it was the only place in England where any form of motor racing was legal. Road racing was barred from England until 1932 when a circuit was opened at Donington Park. Brooklands closed in 1939.

Oval, banked speedways on the European continent, such as Monza (Italy) and Montlhéry (France), were attached to road circuits and mostly used for Grand Prix racing, with the cars completing only half of the oval between laps of the road course. Montlhéry was also the scene of many long-distance international speed records.

Possibly the best known speedway in the world is a 2½-mile (four-kilometre) track at Indianapolis, Indiana, scene of a 500-mile sweepstakes. The Indianapolis Speedway opened as a dirt track on August 19, 1909. The first 500-mile race was run on May 30, 1911, after the track had been paved with bricks.

The Indianapolis 500

The Indianapolis races did not follow Grand Prix formulas, but the rules also restricted engine displacement, though far larger engines were admitted. The Indianapolis rules closely followed Grand Prix displacement limits between 1921 and 1930, and the result was total domination by two U.S. makes: Miller and Duesenberg. This led to a radical change in the formula. A so-called junk formula took effect in 1930; it encouraged the use of stock-block engines, and superchargers were banned. In 1938 a new displacement-limit formula was adopted, coinciding with the contemporary Grand Prix formula. The entry was limited to 33 cars in 1934, and this rule still stands. Helmets have been required since 1935. Fixed roll bars to protect the driver of an overturned car became mandatory in 1959.

Part of the track was resurfaced with asphalt in 1935, and in 1962 the whole track was asphalt-covered except for a 36-inch (91-centimetre) strip of bricks at the starting line. In 1957 a new formula restricted engine displacement to 255 cubic inches (4,180 cubic centimetres) without supercharger and 171 cubic inches (2,800 cubic centimetres) with supercharger. This formula was in effect through 1968.

American-built Indianapolis-type cars are not produced by the major auto manufacturers but by small specialist establishments. Established European racing-car builders have also competed with success at Indianapolis.

Beginning of stock-car racing

Oval-banked board tracks were popular in the U.S. during the 1920s. Rules generally catered to the Indianapolis type of cars, although some 24-hour events were staged, mainly for stripped-down touring cars. Stock-car racing, limited to American production-model passenger cars with suitable modifications, began at Langhorne, Pennsylvania, on July 4, 1939. Langhorne is an oval track, and stock-car racing has been primarily identified with oval speedways although stock cars regularly appear at road-racing circuits in various parts of America.

Stock-car racing did not become organized until the formation of the National Association for Stock Car Auto Racing (NASCAR) in 1948. The first NASCAR race was held on the Daytona Beach (Florida) road course on February 15, 1948. A Grand National division was started in 1949, and the first Grand National race was held on the Charlotte (North Carolina) speedway. There were nine Grand National races that year, including Heidelberg, Pennsylvania; North Wilkesboro, North Carolina; and Hamburg, New York.

Racing, conducted in several classes—grand national, modified, sportsman, and others—depending upon the type of automobile and the degree of modification allowed, has flourished, with annual attendance in the United States alone estimated at 10,000,000. Premier events are held at Charlotte and Daytona, and at Atlanta, Georgia, Darlington, South Carolina, Talladega, Alabama, and Ontario, California.

Development of endurance racing. Endurance racing started at Le Mans in 1923 when the Automobile Club de l'Ouest organized its first 24-hour race. The event took place on a road course of 11.4 miles (18.3 kilometres), partly paved, partly dirt, outside the city of Le Mans in western France. Racing went nonstop from 4 PM on Saturday to 4 PM on Sunday. Only cars in series production were admitted. Four seats were required (except for small cars with engines under 1,100 cubic centimetres), and the seats had to carry ballast to simulate a full complement of passengers. Cars were disqualified for failing to maintain minimum average speed, based on their engine size.

In the second year, convertibles were required to stop on the 12th lap and put up the top, run a minimum of 20 laps, then fold it again. Self-starters and headlights were required.

The rules were relaxed, year by year, in the interest of obtaining more entries, faster cars, and bigger crowds. By 1931 sports cars with two-seater bodies were eligible. The circuit was modified over the years (straightened, widened, resurfaced, and shortened). Since the resumption of this event after World War II the circuit has been 8.4 miles (13.5 kilometres) per lap. The most successful cars over the years have been, in succession, Bentley, Alfa Romeo, Bugatti, Jaguar, Ferrari, Ford, and Porsche.

Le Mans running start

The so-called Le Mans start requires the drivers to run across the road at the drop of the flag, get into their cars as lined up in order of engine size along the pit front, start them, and drive off as fast as feasible. This rule was not changed until 1970. Now the drivers sit inside their cars, with harness fastened, but are not allowed to start the engines until the fall of the flag. Race average speeds have gone from under 60 miles (100 kilometres) per hour to over 120 miles (200 kilometres) per hour for the full 24 hours, with no time out for refuelling or tire changes.

Belgium started its own 24-hour race in 1924, on the Spa–Francorchamps road course. It was an annual event, with average speeds rising from under 50 miles (80 kilometres) per hour the first year to over 90 miles (145 kilometres) per hour by 1953. The race was cancelled for 1954 and was later reorganized as a touring-car race.

The first step toward endurance racing in Great Britain was the Tourist Trophy (TT), the premier motor racing event until Britain got its own Grand Prix in 1948. The first Tourist Trophy race was held in the Isle of Man in 1905. It was a 205-mile (330-kilometre) road race for touring cars. Distance went to 338 miles (544 kilometres) in 1908 and to 600 miles (965 kilometres) in 1914. The TT was suspended between 1914 and 1922 and from 1923 to 1928. For 1928 it was moved from the Isle of Man to Ards, east of Belfast, in Northern Ireland. Distance at Ards was 410 miles (660 kilometres), except for 1933 and 1934, when it was 478 miles (769 kilometres). The series terminated in 1936, not to be revived until 1950.

In 1927 Italy revived point-to-point road racing with a 1,000-mile (1,600-kilometre) event called the Mille Miglia, starting at Brescia, running down to Rome and back again over another route to Brescia. The first race was won at 48.27 miles (77.67 kilometres) per hour. The 1955 race-winning average was just short of 100 miles (160 kilometres) per hour. The race was open to sports cars of all descriptions, with full road equipment. No codriver was required, but most drivers chose a navigator or driving partner. The Mille Miglia was entirely dominated by Alfa Romeo throughout the 1930s and by Ferrari from 1948 to the end in 1957.

Opposition to the race (for its lack of public safety precautions) had been rampant ever since Premier Benito Mussolini had closed it down after the death of 23 spectators in 1938 and directed it to be held on a shorter, closed circuit. In 1957, a single-car accident took the lives of 15 spectators, and the Mille Miglia was outlawed.

The Sports Car World Championship

The Sports Car World Championship, which is awarded not to a driver but to a make of automobile, started in 1953, adding up points for each make on the basis of results in the premier endurance races across the world: 12 hours of Sebring, 1,000-kilometres at Nürburgring, 24 hours of Le Mans, Mille Miglia, Tourist Trophy, Targa Florio, Carrera Panamericana de Mexico, the 24-hour race at Spa, and the Grand Prix of Sweden. (When the term Grand Prix is used in this context it refers, not to the type of car raced but to the fact that the race is a major automotive event in the nation in which it is held.) The list of qualifying races has changed from year to year, as some events are cancelled and new ones come into prominence.

The Mexican road race was from Tuxtla Gutiérrez in the south to Ciudad Juárez on the U.S. border, held from 1952 through 1954. The Grand Prix of Sweden was taken off the list in 1955 and has not been on the calendar since 1957. The 24 hours of Daytona was added in 1964, using part of the speedway and a flat, simulated road course in its infield. Ferrari was a consistent winner in the 1950s (except for Mercedes-Benz in 1955 and Aston Martin in 1959). The Ferrari domination lasted until conquered by Porsche in the late 1960s.

Evolution of hill climbs. Hill climbing started in France with the Nice–La Turbie event of 1897, covering four miles (six kilometres) of hard-surface public roads. Britain's first hill climb was the 1905 Shelsley Walsh, a rough, dusty or muddy, narrow track running steeply uphill through wooded terrain for 992 yards (907 metres).

On the Continent, hill climbs tended to have longer distance, such as Austria's Klausen (13.3 miles [21.4 kilometres]) and Semmering (6.15 miles [9.90 kilometres]), and they became a substitute for point-to-point road racing. France's Mont Ventoux (13.5 miles [21.7 kilometres]) ran from 1902 to 1936 and Grand Prix cars were eligible. Italy had a variety of hill climbs, starting from the northern lakes and running as high up in the Alps as the roads went (Stelvio, Susa-Moncenisio, Aosta-Great Saint Bernard), and a number of shorter events in the Apennines. The most famous hill-climb scene in the New World is Pikes Peak in Colorado, an unpaved hill of 18 miles (29 kilometres) length.

Evolution of drag racing. Drag racing as an organized sport started on the dry lake beds of southern California. Unofficially, drag racing had been going on in the streets of California and other American states for several years. A drag race might take place whenever young men driving cars especially tuned for acceleration met at traffic lights and gave each other the signal to "drag" (*i.e.*, race to the next stoplight). The first moves to get the sport off the street and channelled into safer surroundings were made with the foundation of the National Hot Rod Association (NHRA) in 1951 and the scheduling of races at a drag strip in Madera, California.

Drag race safety regulations

Safety regulations are strict. Steel safety shields completely surrounding flywheel housings became mandatory in 1955. Parachute braking devices were first used in 1959 and were soon made mandatory for all vehicles running over 150 miles (240 kilometres) per hour. Twin V-8-powered dragsters became popular in 1961. Fireproof clothing became mandatory for drivers of fuel dragsters (cars that burn mixtures of alcohol-benzol [or methanol] and nitromethane) in 1963 and later extended to drivers of all supercharged open-bodied machines. On-board fire extinguishers became mandatory for all vehicles in 1971.

NHRA sanctioned its first drag strip in Europe in 1961, but drag racing has not reached any degree of popularity outside the U.S.

The American Hot Rod Association, started in 1954 as a rival to NHRA, sanctions drag races on a national scale.

Evolution of rallies. Rallies are thought to have originated as a competitive version of motoring for pleasure, a popular pastime early in the 20th century. The basis of rally events is the attainment of regularity, by means of an average speed that has to be maintained not merely over the route as a whole but on every section of the route between control points. Legal speed limits have to be observed, high speeds being achieved at specified stages on unrestricted roads and on special sections (private land) and race circuits. Special tests of driving skill are held in difficult terrain en route. Consistent driving is needed as these tests and others at the finishing point are made against the stopwatch; and the results are used, in combination with a penalty marking system for general timekeeping, in assessing the winners in the different classes. In the United States rallyists follow a more or less complicated course—unknown to them until the start of the run—on public roads and under prevailing traffic regulations, the object being to maintain specified precise average speeds between checkpoints. All cars entered in a rally must pass careful inspection. Ordinarily each car is driven by a team consisting of a driver and a navigator who follows the directions for the route and calculates elapsed time, mileage, and speed.

Early tours and rallies

Not counting the Paris–Rouen event of 1894, the first rally was England's 1,000 Miles Trial of 1900. It was a tour from London to major cities such as Bristol, Manchester, Edinburgh, and Leeds, and return to London, with one-day exhibitions of the competing cars at these stops. Of 65 starters, 51 finished.

In 1904 the American Automobile Association sponsored a tour from New York City to St. Louis, Missouri, to visit the St. Louis world's fair. Charles Jasper Glidden, an enthusiastic participant, offered a $2,000 trophy to the AAA. Mr. Glidden then decided to sponsor a tour annually. The first Glidden Tour (1905) went from New York City through the New England States. The American Automobile Association continued to sponsor the Glidden tours until 1913, when the series ended.

Many early competitions that would now be called rallies used the name trials while the word trials today has a different meaning.

The first trial was staged at Edinburgh in 1904, including a road rally from London and back, and a hill climb in Scotland. This event was followed by two English events, the Land's End Trial in 1908 and the Exeter Trial in 1910. A form of rally more oriented to speed and regularity started with the Herkomer Trials of 1905–07 in Germany. The results were not determined on a basis of speed but of correct-route following and regularity. These events, sponsored by a wealthy painter, Hubert von Herkomer, led the Royal Automobile Club of Prussia to institute the Prince Henry Trials in 1908, held annually through 1911. Eligible vehicles were touring cars of certain minimum weights and dimensions, and the trial included special speed events along the way. Prominent cars were Benz, Mercedes, Adler, Opel, Austro-Daimler, and Laurin-Klement. The Monte-Carlo Rally started in 1911 and soon became the premier event of its type. Staged in midwinter, with starting points as far afield as St. Petersburg (Leningrad), Oslo, Athens, and Lisbon, it introduced the realities of extreme differences in climatic conditions to motor sport. The French Hotchkiss dominated the Monte-Carlo Rally in the 1930s and 1950s, giving way to Sunbeam, Austin, Citroën, Lancia, Panhard, Porsche, and Renault in later years. Smaller-scale rallies also became popular after World War II and led to the creation of an international rally championship, under Fédération Internationale de l'Automobile rules.

Long-distance tours

Extreme long-distance tours such as the Peking–Paris of 1907 and New York City–Paris of 1908 were races rather than rallies (with the outcome more dependent on road conditions and reliability than on speed potential). Such events were revived in 1953 with the staging of the Coronation Safari in Kenya. This became an annual event as the East African Safari Rally and developed into an extremely tough event with heavy emphasis on speed. In the mid-1950s Australia had its long-distance RedeX Round-Australia Trials (since discontinued). The most recent events have been the London–Sydney Rally of 1969 and the London–Mexico City Rally of 1970.

MOTORCYCLING

The development of motorcycling largely paralleled and often coincided with the development of automobile sports. There was a class for motorcycles in many of the old town-to-town automobile road races: Paris–Vienna, Paris–Nice, Paris–Cabourg (France), and Paris–Madrid. The de Dion tricycle dominated the sport in 1897, but two-wheelers like the Werner soon set the stage for an entirely different form of racing. Prior to 1901 the Werner carried its engine above the front wheel, but later versions had the engine mounted low in the frame, between the two wheels, with belt or chain drive to the rear wheel.

The international cup and the TT

After the turn of the 20th century, other makes came to the forefront: Griffon, Buchet, Sarolea, Adler, and Laurin-Klement. In 1904 the FIM (Fédération Internationale du Motocyclisme) created the international cup, uniting five nations: Austria, Denmark, France, Germany, and Britain. Each nation was to enter a team of three machines manufactured entirely in the country they represented, as in the Gordon Bennett automobile races. The first international cup race took place in 1905 at Dourdan, France.

The race for the Tourist Trophy (TT) was inaugurated in 1907 and was to become the most famous of all motorcycle races. The first TT race took place in the Isle of Man, and the winning machine was a British Norton powered by a Peugeot (French) engine. The victory gave a new impetus to Norton and the entire British motorcycle industry. Norton was in the front rank of motorcycle racing until 1951, when the Italians (Gilera, MV Agusta, Moto Guzzi) and Germans (NSU) took the lead. The 1960s were years of Japanese domination (Honda). The TT is still held every year in the Isle of Man, and a victory there carries more weight than any other motorcycle race.

Motorcycle racing in North America began in 1903 with the formation of the Federation of American Motorcyclists in New York City. By 1924, this society evolved into the still-active American Motorcycle Association, with headquarters in Chicago.

In the early days, motorcycles raced at the same tracks as automobiles: Elgin, Illinois, Savannah, Georgia, La Grande, Oregon, and Marion, Indiana, were among the most important.

Since 1937 the Daytona 200 miles (320 kilometres) race has been the leading race in the U.S. The race is held on the same road circuit used for the 24-hour Daytona auto race. Grand Prix racing (in the sense of being a major event in the country in which it is held) for motorcycles started soon after auto racing resumed after the end of World War I.

The Belgian motorcycle racing Grand Prix dates from 1921 when it was held at the Francorchamps circuit. The first event covered 188 miles (302 kilometres) on roads consisting of stones rolled into a mixture of sand and clay. The roads have been improved, but, except for one meeting held on a circuit near Dinant in 1923, one at Floreffe in 1936, and the interruption occasioned by World War II (1939–46), it has remained as an annual event at the traditional Francorchamps site. Germany began motorcycle Grand Prix racing in 1925, at the Avus track on the outskirts of Berlin. The event was moved to Nürburgring, back to Avus, and then to Sachsenring where it remained from 1934 through 1939. Resumed in 1949 at Nürburgring, it is now held at Hockenheim.

The Dutch Grand Prix started in 1925, for Dutch riders only, on a circuit near Assen in the northern part of The Netherlands. In 1927 the meeting, held on a different course in the area, became international. Switzerland's motorcycle Grand Prix started in 1934 on a short circuit outside Geneva (Genève) and moved to Bremgarten in 1936. The last Swiss Grand Prix for motorcycles was held in 1954.

There are about as many forms of motorcycle racing as there are of automobile racing—trials, speedway, hill climbs, rallies, moto-cross, drag racing, and ice racing. Trials date back to pre-World War I days, and speedway racing (dirt track) came into being in the early 1920s in Australia. Moto-cross—cross-country races consisting of a specified number of laps over a closed course laid out on rough, natural terrain—was successfully introduced in Great Britain and on the Continent in the late 1940s and early 1950s and by the 1960s had become a popular spectator sport. It was demonstrated in North America in the 1960s and the first regularly scheduled series of moto-cross events there was held in 1970.

Also beginning in the late 1940s, increased interest developed in events known as scrambles, which are similar to moto-cross but less regulated.

Motorcycle drag racing is an American phenomenon, dating from about 1950. Motorcycle ice racing started in Scandinavia in the 1930s and has spread to other temperate-climate countries. (For history and description of both of these events see below.)

II. Modern automobile sports

RACING

The various participants in automobile racing have different reasons for competing. Racing drivers are motivat-

ed by competitive spirit, and the promise of wealth, fame, and glory. Racing-car builders compete because the sport offers (a) publicity for all products bearing the same name, and (b) the acquisition of technical knowledge, trained specialists, and facilities that can be utilized for product improvement and associated commercial purposes.

Reasons for racing

Organization of the sport. There are two principal categories of race organizers: national and local clubs, which are nonprofit organizations, and private promoters, who are motivated by financial gain. The clubs participate partly for reasons of prestige, partly to gain power and weight in the organizations that govern motor racing.

Governing bodies. The world governing body of automobile racing is the International Automobile Federation (Fédération Internationale de l'Automobile, or FIA), a voluntary association of the national automobile clubs of 17 countries. It was founded in 1904 as the AIACR (Association Internationale des Automobile Clubs Reconnus) and acquired its present title in 1946. The motor-racing department of the FIA is the Commission Sportive Internationale (CSI). The CSI has delegates from 12 countries with permanent members from England, France, Germany, Italy, and the United States. The other seven members are elected annually from other FIA member countries. The CSI is responsible for the international sporting code, the international competition calendar, and various international championships.

The first duty of the CSI after its creation in 1922 was to establish an international sporting code for motoring sport. The first such code, developed, with modifications, from a draft code of the Automobile Club de France, was adopted in 1926. Since that time it has been continuously amended and brought up to date.

The Automobile Competition Committee for the United States (ACCUS), established in 1957, is the permanent representative of the CSI and the FIA in the U.S. Its board of directors includes representatives from the Sports Car Club of America (SCCA), the United States Auto Club (USAC), the National Association for Stock Car Auto Racing (NASCAR), and the National Hot Rod Association (NHRA). The primary responsibility of ACCUS centres on coordinating the schedule of international races held in the United States.

Finance. The question of finance has two parts: (1) How the race organizers finance their events. (2) How the makers (*i.e.*, builders or manufacturers of automobiles) finance the building and operating of racing machinery.

The functions of race organizers

The race organizer is traditionally a major or national club but increasingly of late has been a company formed to operate a racetrack. This is big business: the construction of a superspeedway with various ancillary facilities may cost $25,000,000, and to operate it takes a permanent staff, promotion and public relations officers, maintenance staff and crew. The organization and management associated with any major event requires large numbers of specialized personnel in a variety of branches. Some are volunteers, some paid by other organizations, but fees and wages can run into very important figures.

To attract a competitive field, the race organizer must offer starting money or prize money, or both. Starting money is a sum paid to an entrant when the machine and its driver are presented at the starting line. This sum can run from $1,000 to $20,000. Prize money paid by the race organizer can run to $750,000 in a single event, with an individual entrant (winner) collecting up to $125,000. The biggest purses are paid at Indianapolis Motor Speedway and in the Canadian-American Challenge Cup races.

The organizer's income is from ticket sales, fees, or profits from food, refreshment, or other concessions, advertising on the track, program sales, and advertising in the program. Most racing-car constructors are also entrants. The manufacture of a competitive Grand Prix or Indianapolis car may cost up to $100,000. Builders have found various ways to finance the building and operation of a team of cars: (1) If the constructor also makes touring cars, some of the racing-car development costs can be billed to the touring cars (and built into their prices). (2) The maker may produce racing cars for sale to private teams and professional stables. (3) He may seek a sponsor. This amounts to using the race car as an advertising medium, and painting the sponsor's name (or that of his product) on the cars that are raced officially by the constructor. Such sponsors may belong to associated industries (tires, oil, gasoline, spark plugs, antifreeze) or completely unrelated businesses (tobacco, beer, liquor, cosmetics, banks and credit institutions). (4) Starting money and prize money.

Motor-industry associations have occasionally given financial assistance to a constructor in order to extend the benefits of favourable publicity to the entire membership. National governments have rarely subsidized motor racing, but when they have, it has been with the double objective of increasing national prestige and to bring about a rise in national export trade.

Race management. The primary requirement for a race is that the track or site be approved by the sanctioning body in respect to safety for both participants and spectators. A substantial staff is needed for the organization and management of a major event: (1) The clerk of the course is responsible for conducting the event in accordance with the program. (2) The stewards are responsible for enforcement of the international sporting code. (3) The secretary is in charge of the race preparation and administration. (4) Scrutineers are responsible for vehicle inspection. (5) Timekeepers are responsible for the lap scoring and timing. (6) Pit observers and road observers are agents of the clerk of the course. (7) Flag marshals are in charge of signals to be given to the drivers (to warn of accident ahead, oil on the track, other dangers, or disqualification). (8) The starter gives the official starting signal (by flag, or signal lights).

A host of other services are required for various purposes (crowd control, safety, plant protection, communications, revenue). In addition to the presence of a regular police force, most events call for a special security force. With 250,000 spectators, a staff of 500 security men may be needed. Ticket and program sales require cashiers, supervisors, and assistants. A telephone-telegraph service requires operators; official communications depend on messengers; and the public address system must be staffed. Emergency medical service may call for ambulances, doctors, and nurses, with possibly a helicopter or a mobile hospital unit standing by.

Safety. Over the years, hundreds of drivers and spectators have been killed in racing. The risks are implicit in the nature of racing. Accidents are caused by pilot error, mechanical defects, and random hazards. They will continue to happen. The problem is to protect the drivers and spectators when they do occur.

Safety measures for drivers

The wearing of crash helmets has been common since 1950 and mandatory in most events since 1954–56. Seat belts or harnesses, pioneered by USAC and the SCCA about the same time, are now universally required.

Fireproof fuel tanks came in the early 1960s; automatic fire extinguishers in the late 1960s. Roll bars have been mandatory in SCCA and USAC events since the 1950s, in Grand Prix racing since 1961. The Motor Racing Safety Society, formed in New York in 1969, is internationally active as an advisory and information-collecting organization, with the goal of reducing injuries and fatalities in racing accidents.

The best protection for spectators is distance. Modern crowd-control methods place them in public enclosures, behind barriers or fences, separated from the road, track, or speedway by an open area. Guardrails are installed if distance is judged inadequate.

Classes and events. *Grand Prix.* Grand Prix racing comprises cars built under a variety of formulas, the most important of which is Formula I. Supporting formulas (such as Formula II, Formula III) were created to form a training ground for drivers prior to their acceptance into Formula I racing. All Grand Prix racing is for open-wheeled single-seater racing cars. Engine size and many other dimensions are limited by FIA rules.

Formula I limits engine displacement for supercharged and unsupercharged engines, specifies minimum weight, and prohibits aerodynamic devices that are not part of the body structure proper. Specifications cover camshafts, valve gears, and other components for various types of engines. Engines with more than 12 cylinders were outlawed in 1972.

Only FIA graded drivers of "A" standing are permitted to enter Formula I races and also have the right to participate in all international events listed on the FIA calendar. The qualifications for a new driver to obtain "A" standing include finishing at least twice in the top three in a race for the Manufacturers' Championship or winning the European Trophy for Formula II drivers (in addition to winning at least three nongraded events). In the 1970s there were only about 30 "A" graded drivers in the world.

The Formula I Grand Prix races

The countries that have their own Grand Prix (Formula I) races are the traditional ones: France, Italy, Germany, Spain, Monaco, The Netherlands, Belgium, and England. In addition, South Africa, Canada, and the United States are now included. Mexico was excluded after 1970 because of the lack of safety precautions. France's Grand Prix changes its scene from Reims to Rouen, Circuit Paul Ricard at Le Castellet to Clermont-Ferrand. Italy's Grand Prix rarely leaves Monza. Germany usually stages its Grand Prix at Nürburgring but occasionally goes to Avus. Spain alternates between Jarama (Madrid) and Montjuich (Barcelona). The Monaco circuit is the same as when it began, through the city streets. The Dutch Grand Prix is held at Zandvoort in the sand dunes on the coast outside Haarlem. The Belgian Grand Prix has its home at Spa-Francorchamps (but was banned in 1971 by the Grand Prix Drivers Association as unsafe). England's race goes from Brands Hatch to Silverstone and back. The South African Grand Prix is held near Cape Town. Canada uses Mont Tremblant or Saint-Jovite. The U.S. Grand Prix is firmly embedded at Watkins Glen in New York state. The Mexican race, when it was held, took place in the streets of Mexico City.

Formula II also limits engine displacement, but, since 1972, minimum vehicle weight has depended on the number of cylinders. Both the motor block and cylinder head must be derived from an engine used in a touring car or a sports car manufactured in a quantity of at least 1,000 units during 12 consecutive months. No restrictions are imposed on type of cylinder head or number of camshafts. Four-wheel drive is not allowed, and no more than five forward speeds are permitted.

Formula III also limits engine displacement. The number of cylinders is restricted to four. Engine block and cylinder heads must be those of a series-production car manufactured in quantities of at least 5,000 units during 12 consecutive months. The intake manifold must be fitted with a throttling flange through which all air fed to the engine must pass. Minimum wheelbase, minimum track (width between wheels), and minimum weight are specified.

Formula A was devised to encourage competition between the American and European continents. Formula A/5000 calls for open-wheeled single-seater cars using stock-block engines with pushrod valve gear of up to 5,000 cubic centimetres (305 cubic inches). Chevrolet and Ford-based units are preferred by most makers. Formula A has gained popularity in the U.S. and dominates the Tasmanian series (run during the Northern Hemisphere off-season). Formula A also enjoys a great following in Argentina, where the cars compete against Formula I cars on even terms (as they also do in the Questor Grand Prix at Ontario, California). There is also a supporting Formula A with a displacement limit of 1,600 cubic centimetres (100 cubic inches). Only production engines are admitted. This category is seen as a low-budget way of entry into Formula II racing.

Two championships are in existence for Formula A cars: the SCCA Continental Championship in the U.S. and the Guards Formula 500 championship organized by the British Racing and Sports Car Club.

Formula-V racing is for Volkswagens, and the engines and all major chassis components are basically those of the 1961–65 Volkswagen. The body is single-seater, and the wheels are open. No component of engine, power train, front suspension or brakes may be altered, modified, or replaced. Wheelbase and overall length are limited. Any size tire is permitted as long as it fits the standard VW wheels. Formula V is known as the lowest cost form of open-wheel racing.

Small-car formulas

Formula Super-V allows use of the Volkswagen "1600" engine and poses no restrictions on wheelbase. Still, no engine, drive line, brakes, or suspension components may be altered. Minimum weight is fixed, and wheel sizes are restricted.

Formula Ford was an English formula for cars using Ford of England four-cylinder production engines up to 1,600 cubic centimetres displacement (Cortina, Capri, Escort).

Formula 100 took the place of Formula Ford in the U.S. under regulations of the International Motor Sports Association. Cars powered by 1,600 cubic-centimetre Renault and Volkswagen engines can compete with the Ford-powered cars.

Formule Renault is an edition of Formule France, for cars equipped with 1,135-cubic-centimetre (70-cubic-inch) Renault engines, and a minimum weight of 425 kilograms (935 pounds). A number of French makers have built cars under Formule Renault rules, but the cars have not gained international recognition.

Manufacturers' Championship. The major races that count for the Manufacturers' International Championship follow closely on Formula-I events in importance. They are the 1,000 kilometres of Buenos Aires; the 24 hours of Daytona; the 12 hours of Sebring; the BOAC (British Overseas Aircraft Corporation) 1,000-mile race at Brands Hatch; 1,000 kilometres of Monza; 1,000 kilometres of Spa; Targa Florio; 1,000 kilometres of Nürburgring; 24 hours of Le Mans; 1,000 kilometres of Österreichring; six hours of Watkins Glen; and 1,000 kilometres of Canada at Saint-Jovite.

The championship racecourses

At Buenos Aires, the race is staged at an autodrome (speedway) near the city. The Daytona event uses part of the speedway plus a road course in the infield. Florida's Sebring is a track made up of airfield runways and access roads. Brands Hatch is a road circuit in rolling country in Kent, near London. Monza is a road course (with part of a speedway) near Milan, Italy. Spa-Francorchamps is a 5.6-mile (9.0-kilometre) road circuit in southeastern Belgium. Targa Florio is run on mountain roads in Sicily, as always. Nürburgring is a 14.2-mile (22.8-kilometre) course, with 176 curves, in western Germany. Le Mans permanent circuit is 8.3 miles (13.4 kilometres) long and has a straightaway of over three miles (five kilometres), on which some cars exceed 200 miles (320 kilometres) per hour. Österreichring, better known as Zeltweg, is a short 3.7-mile (six-kilometre) circuit in Austria. Watkins Glen is a fast, difficult road course in New York's Finger Lake region. Saint-Jovite is a relatively new facility with a good many interesting features, in the Laurentian Highlands in Quebec Province.

The eligible cars for the Manufacturers' Championship are grand touring cars, sports cars, and sports-car prototypes, as defined by the Fédération Internationale de l'Automobile. There are nine groups of cars, ranked according to production levels and purpose of operation. The groups admitted to the Manufacturers' Championship are 4, 5, and 6. Group 4 is defined as special grand touring cars with a minimum production of 500 identical models in 12 consecutive months. Group 5 is defined as sports cars with a minimum production of 25 a year and group 6 as prototype sports cars with no minimum production figure.

In this series of races, Group 4 cars enjoy unlimited engine displacement. Group 5 cars are restricted to 5,000 cubic centimetres (300 cubic inches), and Group 6 cars to 3,000 cubic centimetres (180 cubic inches). FIA specifies that the races must cover a minimum distance of 1,000 kilometres (620 miles) or have a duration of at least six hours. Some races exceed these requirements by a substantial margin. Supplementary regulations specify which parts may or may not be replaced during the race,

what tools may be used, conditions for replenishment of fuel and oil, and the number of mechanics who may work on a car in the pits.

Each driver is limited to four hours at a stretch in races over 12 hours or more. Time at the wheel is further restricted for shorter races. In a 24-hour race, maximum total driving time is 14 hours and minimum total rest time is seven hours.

The Canadian-American Challenge Cup. The Canadian-American Challenge Cup, popularly known as the Can-Am series, is an annual series of races in North America staged jointly by the SCCA and the Canadian Automobile Sports Club (CASC). Eligible cars are known as Group 7 (two-seater racing cars with no engine displacement limit and sports-car type bodywork without headlights and bumpers). Cars must run on commercial fuel and be equipped with self-starters (though the races are run with a "flying start" behind a pace car and not from a standing start). The races are short (from one hour and 20 minutes to two hours) and speeds are extremely high. Race averages range from 95 to 130 miles (150 to 210 kilometres) per hour, and the cars (with upwards of 600 horsepower) are capable of top speeds approaching 200 miles (320 kilometres) per hour. Minimum seating dimensions are specified by the rules, which also prohibit the use of movable body parts (airfoils, for instance), and set forth specific requirements for safety equipment (roll bars and belts).

The Can-Am series

The Can-Am series includes ten events. Points are awarded to the finishers in each race and added up so that the vehicle and driver with the highest number of points at the end of the season wins the cup.

American track racing. The United States Auto Club sanctions racing in various categories of single-seater open-wheeled racing cars on oval tracks from one-fourth of a mile to 2½ miles (one-half to four kilometres) in length, plus certain special competitions, such as the Pikes Peak Hill Climb. USAC's premier event is the Indianapolis 500 Miles Sweepstakes. The championship division of USAC governs the national championship for drivers on a points basis for a total of 21 events staged all over the United States.

Championship races cover a distance of 100 to 200 miles (160 to 320 kilometres). Most tracks are paved ovals, though some road courses are also used. Championship cars are subject to the same design measurements laid down for cars that compete in the Indianapolis 500. Engine displacement is restricted according to type of construction: overhead-camshaft power plants, without supercharging and supercharged; stock-block engines with pushrod overhead valves, unsupercharged and supercharged; diesel engines, unsupercharged and supercharged. Gas-turbine engines are outlawed. Minimum wheelbase and maximum width are specified. Front or rear-wheel drive is optional but four-wheel drive cars are barred.

The sprint-car division of USAC sanctions short-distance speedway races at a number of second echelon speedways throughout the U.S. Sprint cars are front-engined single-seaters of simplified design and construction compared with championship cars. They are heavier and slower.

Sprint and midget cars and races

The midget-car division of USAC governs dirt-track races with midget cars. Sprint cars are regulated by the same engine-displacement limits that apply to championship cars. Midget cars are scaled-down versions of sprint cars.

Touring-car racing. Touring-car racing takes different forms in different parts of the world, but the most important categories are run on European road circuits under FIA rules. The FIA has created an International Touring Car Championship for production-type touring cars of Group 1, Group 2 and Group 3. Group 1 cars are series-production cars manufactured in a quantity of at least 5,000 units of identical specifications. Group 2 cars are special touring cars manufactured in a quantity of at least 1,000 units of identical specifications. Group 3 cars are series-production grand touring cars manufactured in a quantity of at least 1,000 units of identical specifications.

The cars are further subdivided into 13 classes according to engine displacement. The use of optional equipment is restricted, and all parts (springs, wheels, tires, shock absorbers, and so forth) must conform to rules set forth for each category.

Championship races are usually four-hour events for Group 2 cars. They take place from March to October on circuits such as Monza (Italy), Brands Hatch (Kent), Jarama (Spain), Nürburgring (Germany), Spa-Francorchamps (Belgium), Zandvoort (The Netherlands), and the Paul Ricard Circuit (Le Castellet, near Toulon, France). Lesser events with cars of Groups 1, 2, and 3 are held throughout the season on other road circuits.

Touring-car racing in the United States takes two distinct and separate forms. One type of car is used for races under The National Association for Stock Car Auto Racing (NASCAR; see below *Stock-car racing*) and United States Auto Club (USAC) rules; another type is used in the Trans-Am (Trans-American) races sanctioned by the SCCA.

Eligible cars for the Trans-Am series are fixed-roof production cars built in a quantity of at least 2,500 units of identical specifications (except for the engine and transmission). Engine displacement and wheelbase are limited. At least 500 of the engines must have been built and installed. There are four classes, depending on engine displacement.

The Trans-Am series

This is one of the few forms of racing in which American, European, and Japanese production cars compete on equal terms. Trans-Am races are held from May to October.

Touring-car racing as performed in Japan, Australia, and South America is not subject to the same strict rules that govern the sport in Europe and North America. Competitors generally have carte blanche for modifications. Such races are not listed on the international calendar and are not included in any but local championships.

Stock-car racing. Stock-car racing is a unique American form of racing modified production cars. Stock-car events are sanctioned either by NASCAR or USAC.

NASCAR sanctions competition in five distinct categories of stock-car racing: grand national, grand American challenger series (formerly grand touring), late-model sportsmen, modified late model, and Pacific-coast late model. The so-called stock cars are anything but stock, being specially built and prepared in expert speed shops and supplied with special nonproduction components by the auto manufacturers. The external bodywork remains identifiable as a production-type passenger car. Grand national cars can be no more than three years old. Minimum wheelbase and minimum weight are specified. There are two classes based on engine size. A requirement that a restrictor plate be installed between the carburetor base and the intake manifold was discontinued in 1971, but carburetor-base openings were limited, depending on engine design and displacement. All engines are stock-block V-8s with four-barrel carburetors, and the restrictor plate has four lobes with size limits according to engine type and size. There is also a sliding scale for race-ready vehicle weight without driver and minimum dimensions concerning ground clearance.

Stock-car categories

The Grand American category was instituted for cars in the Trans-Am category. The same engine-size limits apply. Late-model sportsman cars are three to 13 years old, with minimum wheelbase and weight. Engine modifications are allowed, but only one carburetor can be fitted. Modified late-model stock cars have a minimum wheelbase and weight, but there are no restrictions on engine and chassis modifications. Re-bored blocks, superchargers, fuel injection, and any number and type of carburetors are authorized. The Pacific-class rules differ only in detail. Most NASCAR stock-car races are held on superspeedways such as Daytona (Florida), Talladega (Alabama), Atlanta (Georgia), and Ontario (California); but many older speedways are still in use, such as Darlington (South Carolina), and Charlotte and Rockingham (both North Carolina). The cars approach 200 miles (320 kilometres) per hour and race averages exceeding 150 miles (240 kilometres) per hour have been recorded.

The stock-car division of USAC, which sets different rules

than NASCAR, sanctions races in which greatly modified late-model American production-model passenger cars are eligible. A third series of stock-car races is sanctioned by the Automobile Racing Club of America, mostly on short oval tracks, both dirt and paved.

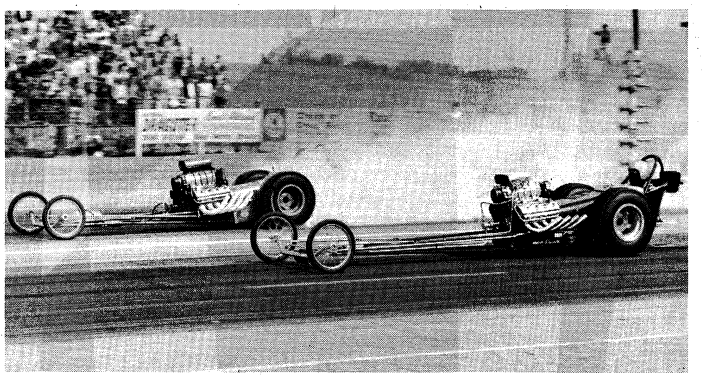
Sports-car racing. Races for production-type and modified sports cars are held all over the world. Sports-car racing is essentially an amateur sport, in contrast to the heavily professional manufacturers' championship races. Sports-car races are, of course, road-circuit races, often conducted concurrently with touring-car races. All major and many minor courses are in regular use throughout the season.

The most regulated form of sports-car racing takes place in the United States, where the SCCA directs a road-racing championship for production-type and modified sports cars. The SCCA lists a classification of all current or popular sports cars, grouping them in categories according to proved performance so as to accommodate all competitors on an equitable basis.

There are eight classes, from A through H, A being for the fastest machinery and H for the slowest. There are no engine-displacement or vehicle weight limits, but strict limitations govern modifications allowed from standard trim. The SCCA sanctions about 350 events per year, from drivers' schools, regional, and restricted-entry events, to national, interdivisional, national open, and international races.

Drag racing. Drag races are standing-start acceleration races over a quarter-mile with both elapsed time and terminal speed being recorded. Most events are match races between two vehicles, starting side by side and running parallel to each other on a straight and level roadway or drag strip. The loser of each race is eliminated, the winner returning to face other winners in his class. Starts are made by countdown-type signal lights, and performances are usually recorded electronically.

By courtesy of the National Hot Rod Association



"Slingshot" design cars at the start of their run for top gas eliminator title.

The largest drag-race sanctioning body is the National Hot Rod Association (NHRA), with an annual schedule of more than 3,000 events. The principal races are strictly professional and cover a multitude of classes, from those for special vehicles capable of covering the standing quarter-mile in about six seconds, with a terminal speed of about 230 miles (370 kilometres) per hour, to practically stock family sedans. The American Hot Rod Association (AHRA) is a rival organization, also sanctioning more than 3,000 events annually. Rules closely parallel those of the NHRA.

Off-road racing. Minor off-road events have been held all over the world, at least since the 1920s, but the sport has been organized and regulated in recent years with the formation of the National Off-Road Racing Association in America. NORRA stages two major annual events: the Baja 500, which starts and ends in Ensenada, Baja California, Mexico; and the Mexican 1000, which starts in Ensenada and ends in La Paz, Territorio Sur de Baja California, Mexico. Eligible cars fall in nine categories. Two-wheel-drive categories include: production two-wheel-drive passenger cars; production two-wheel-drive utility vehicles (pickup trucks, etc.); single-seat two-wheel-drive vehicles and two-wheel-drive vehicles with only one occupant; and modified and nonproduction two-wheel-drive cars with two occupants. Four-wheel-drive categories are: production-type four-wheel-drive vehicles and modified and nonproduction four-wheel-drive vehicles. Motorcycle categories are: motorcycles of 125 cubic centimetres (7.6 cubic inches) displacement and motorcycles over 125 cubic centimetres displacement. The ninth category is for "Baja Bugs," (modified Volkswagen sedans).

The Baja 500 and the Mexican 1000

The Mexican 1000 course includes long, straight stretches of open desert, some sudden detours, rock-strewn washes, and low, blind hillocks, boulder-covered rockbeds, sandy beaches, and tufted dunes.

A popular type of off-road event in Great Britain is the Rally-Cross, which is a short-distance cross-country race with production cars, with a minimum of rules. France has two main off-road races, the Rallye Infernal and Rallye des Cimes, which cover longer distances and allow production cars, dune buggies, and four-wheel-drive utility vehicles, as well as sports cars.

RALLIES

Rallies are point-to-point driving events taking place on public roads, with a common meeting point for routes from a number of different starting points, often in different countries. A rally is a reliability trial, not a race. The average speed is limited, and the highest speed attained under way is not a factor in the competition. Most rallies do, however, have special stages (divisions, or sections, of the route that follow a stretch of road that is closed to other traffic) over which each competitor races against the clock. The winner is decided on the basis of regularity on the road sections, speed on the special stages, and performance in extra tests held after arrival at the finishing point. The exact evaluation and allocation of points are set for each rally in the supplementary regulations issued by the organizer.

A secondary form of rally, officially called Touring Concentration, has as its aim the assembly of travellers at a certain point or points. Road stages have no speed schedules, and special stages are confined to additional trials (not speed tests) at the point of assembly.

Duration ranges from about six hours to more than a week, depending on the category of the event. Minimum distances are specified for all major international events to qualify for an FIA-sanctioned championship. The most important rallies are the events that count for the International Rally Championship for Constructors (points are given to the car, not to the driver).

Championship rally events

Eligible cars are specified by the FIA as series-production touring cars (Group 2), series-production grand touring cars (Group 3), and special grand touring cars (Group 4). The exact definitions are spelled out in the International Sporting Code.

The events are the Monte Carlo Rally, the Swedish Rally, the San Remo-Sestriere Rally in Italy, the East African Safari, the Rally of Morocco, the Austrian Alpine Rally, Greece's Acropolis Rally, France's Alpine Cup Rally, and Britain's RAC (Royal Automobile Club) Rally.

The rallies are usually organized by the national automobile clubs in the various countries, and the cars are entered either by the manufacturer or by a private competitor. Most drivers are professionals. The factory-team cars especially are extensively modified (within the rules of the class they are competing in). They may be equipped with larger fuel tanks; extra wheels with special tires, racing seats and steering wheels; a battery of fog lamps; special lights and auxiliary headlights and adjustable spotlights; plus reinforced suspension systems; super-tuned engines; heavy-duty cooling systems; and roll cages (steel tubes or bars to prevent body collapse in a roll-over accident).

There is also a European Rally Championship for Drivers, in which drivers and codrivers are allotted points for their performance in each event in the series, and the points are added up at the end of the series of 21 rallies,

including The Netherlands' Tulip Rally, Finland's Rally of the 1,000 Lakes, Tour de France, and Tour de Corse (Corsica). Eligible cars include all those allowed in the manufacturers' rally championship, plus sports cars and prototype sports cars, at each organizing club's discretion.

Rallies in America are governed by the SCCA. There are many types of rallies in the United States. Michigan's Press-On-Regardless Rally most closely conforms to the European rallies. The MG Car Club's International 1,000-Miles Rally stresses navigation and split-second timing. The Rebel Yell Rally is mainly a test of route-following ability. Other rallies are endurance events. Many local rallies are mere touring concentrations.

Long-distance events such as the 1968 London–Sydney and the 1970 London–Mexico City World Cup Rally are not true rallies in that they are not governed by FIA rules, nor do they conform to the traditional concept of multiple starting points and one finishing point. They were organized not by the national automobile clubs but by newspapers. Both routes included long sea voyages.

HILL CLIMBS AND TRIALS

Hill climbing. Hill climbs are short-distance races against the clock on mountain roads. Only one car races at a time, from a standing start to a flying finish. The roads are usually part of a public-road system, closed to traffic for the occasion. The most important hill climbs are the events that count for the European Mountain Championship: Austria's Semmering, Spain's Montseny, Germany's Rossfeld (Berchtesgaden), France's Mont Ventoux and Le Mont-Dore, Italy's Cesana-Sestriere and Trento-Bondene, Switzerland's Ollon-Villars.

The European Mountain Championship

Eligible cars are special touring cars (Group 2), special grand touring cars (Group 4), and sports cars (Group 5). The minimum distance is five kilometres (three miles), with a minimum altitude difference between start and finish lines of 350 metres (1,150 feet) and a minimum average gradient of 5 percent. Each competitor must run a minimum of ten kilometres (six miles), which means that he must make two runs on hill-climb courses shorter than that. In such cases, times are simply added and speeds averaged. The winner is determined solely by speed. Great Britain has its own hill-climb championship, with such venues as Shelsley Walsh and Prescott. The British events admit open-wheeled single-seater racing cars and sports cars but not touring cars.

Hill climbs in the United States are organized by local auto clubs, and entries range all the way from touring cars to open-wheeled single-seater racing cars. The U.S. number one hill climb is Pikes Peak, Colorado. Minor events are staged in all mountain areas (Rocky Mountains, Alleghenies, Poconos, and Adirondacks).

Hill trials. The word trials in this context has a very special meaning: a "trials" is a particular British event consisting of an off-road hill climb for two-wheel-drive sports cars. The "Trials Special" vehicle is usually a small, lightweight chassis powered by a mass-production engine, with thin front tires and fat rear ones. The Royal Automobile Club (RAC) has set rules for Trials Specials that demand that the distance between the front axle and the foremost spark plug can be no more than one-fifth of the car's wheelbase. The seating compartment must not extend beyond the rear axle. Maximum rear-wheel rim width is five inches (13 centimetres). Limited-slip differentials, chains, studs, and cleated tires are outlawed.

SPEED-RECORD TRIALS

Speed record trials do not constitute a popular spectator sport, as these record-breaking attempts usually take place far from population centres and last too long for public interest. The world land-speed record is a competition for the highest possible average speed over a measured mile, measured over two runs, one in each direction. There are no limitations on power plant, vehicle size and weight, fuel type or consumption. This kind of competition has progressed to the jet or rocket propulsion state (as opposed to mechanically driven wheels).

In the unlimited engine size class, a world land-speed record of 394.2 miles (634.3 kilometres) per hour set by Englishman John Cobb at the Bonneville Salt Flats, Utah, in 1947, stood until July 17, 1964, when Donald Campbell, son of a former British record holder, drove a jet-powered (rear wheels driven) "Bluebird" 403.1 miles (648.6 kilometres) per hour at the Lake Eyre salt flats in Australia. This record for wheel-driven vehicles was broken by Robert and William Summers of the United States on November 13, 1965, with 409.28 miles (658.53 kilometres) per hour at Bonneville. Driving pure-jet-thrust vehicles, Craig Breedlove of the U.S. became the first human to be officially timed at more than 400, 500, and 600 miles per hour, all at Bonneville. He set a record of 407.45 miles (655.59 kilometres) per hour in a three-wheeled jet in 1963, sanctioned by the Fédération Internationale du Motocyclisme. The FIA qualified jet-thrust propulsion for the land-speed record but still required that the vehicle have four wheels. In 1964 Breedlove in a four-wheeled jet was timed at 526.28 miles (846.78 kilometres) per hour (in 23 days during October 1964 the records were: October 2, Walt Arfons/Tom Green [U.S.], 413.20 miles [664.84 kilometres] per hour; October 5, Art Arfons [U.S.], 434.02 miles [698.34 kilometres] per hour; October 15, Breedlove, 526.28 miles [846.78 kilometres] per hour; and October 27, Art Arfons, 536.71 miles [863.57 kilometres] per hour). Art Arfons set the record at 576.553 miles (927.674 kilometres) per hour November 7, 1965; and on November 15 Breedlove pushed it past the 600 mark with an average of 600.601 miles (966.367 kilometres) per hour for two runs through the measured mile.

World land-speed record attempts

World land-speed record attempts can only be made one or two places in the whole world (Bonneville Salt Flats, Utah, and Lake Eyre, Australia) because of the tremendous speeds and distances required. There are also one-hour and 24-hour world-speed records for vehicles of unlimited engine size. International class records cover average speeds over distances from one kilometre to 100,000 kilometres for cars in ten different classes, based on engine displacement. International class record runs are usually made at banked speedways where nearly constant speeds can be maintained throughout, such as Montlhéry, Monza, and Miramas.

A variety of especially made cars compete annually at the Bonneville Salt Flats, during the Bonneville Speed Week, for top speeds on the dry lake. They follow NHRA classification (not FIA) since most active competitors come from the ranks of drag racing.

OTHER EVENTS

Gymkhanas. The gymkhana is a driving skill and manoeuvrability test with obstacles, usually including a selection of slalom S-bends, short straights, a reversing test, U-turns, and other elements of similar nature. Each car goes through alone and is timed. Gymkhanas are popular with local automobile clubs, one-make clubs, and other associations not involved with professional racing. There are no international rules, and most clubs set their own regulations.

Historic (Classic or Antique) car meets. Historic car meets are held all over the world, and competition exists on three levels—racing, road-driving regularity, and *concours d'élégance*. Very few events include racing. Most are route-following events or touring concentrations, with the main competition centred on the *concours*—*i.e.*, the condition (quality of restoration, authenticity, rarity, and overall desirability) of the vehicles.

Classifications of cars

Under British rules, all automobiles manufactured prior to or during 1904 are Veteran cars. Vehicles from the 1905–15 period are Edwardian, and 1915–30 cars are Classic. Under American rules, all cars manufactured prior to 1930 are Antique. Newer cars of special appeal are Classics, which overlap with the Special-Interest cars. Classics stop about 1948, while Special-Interest cars have no limit in time.

Most famous of the route-following events is the British Veteran Car Club's London to Brighton Run, which has been held annually (with interruptions for the war years) since 1896. Cars start at brief intervals from Hyde Park in London and congregate at the end of the 57-mile (92-

kilometre) trip on Madeira Drive in Brighton on the Channel coast. The only true historical car race is the Vintage Sports Car Club annual race meeting at Silverstone, England.

The FIVA (Fédération Internationale des Voitures Anciennes) sanctions nine rallies in Europe: the Westfalen-Rallye at Siegen; the Grand Prix Internationale de Restauration et Conservation at Verona, Italy; Automobilhistoriska Rallyt at Göteborg, Sweden; Gran Premio Barcelona-Alicante in Spain; Internationale Deutschlandrally at Travemünde; Rallye International des Ancêtres in Brussels; Rallye Monte Carlo des Ancêtres in Monaco; Tour du Luxembourg; and the Internationales Motorveteranenrally in Vienna. The Antique Automobile Club of America has a full calendar of events, topped by such functions as the AACA Spring Meet in Dearborn, Michigan; the AACA Glidden Tour; and the AACA National Fall Meet in Hershey, Pennsylvania.

III. Modern motorcycle sports

MOTORCYCLE RACING

Governing bodies. The International Sporting Code for motorcycle racing is the creation of the International Sporting Commission of the Fédération Internationale du Motocyclisme (FIM). The sporting code is continually revised and brought up to date. The International Sporting Commission has several subcommittees, such as the Standing Sub-Committee for Track Racing and the Moto-Cross Sub-Committee. FIM members are the motorcycle associations of more than 40 nations. Each country has only one national club that can be elected to FIM membership. These clubs sanction international, national, and regional events in their respective countries.

Classes and events. *Road racing.* Motorcycle road races are held on closed courses, parts or all of which may be public roads. The motorcycle road-racing world championships, established in 1949, have categories for both riders and machines. The machines are divided into two classes, A and B. Class A is for pure racing motorcycles, class B, for sports motorcycles. There are no formulas, such as in automobile racing, but Class A is divided into five groups, based on engine displacement, from 50 cubic centimetres (three cubic inches) and under, up to 350 to 500 cubic centimetres (20 to 30 cubic inches). Class A includes the Grand Prix motorcycles, which have supertuned engines and aerodynamic bodywork enveloping the handlebars, front wheel, and suspension system. Class B is limited to 500 cubic centimetres.

World-championship motorcycle road races

World-championship races are held on road courses or on speedways. There are 12 races that count for the championship with bonus points allotted to each machine and rider on a basis of performance in each event. The Austrian Grand Prix is held on Salzburgring; the German Grand Prix at Hockenheim; the Tourist Trophy (TT) on the Isle of Man; the Dutch Grand Prix at Assen; the Belgian Grand Prix at Francorchamps; the East German Grand Prix at Sachsenring; the Czechoslovakian Grand Prix at Brno; the Swedish Grand Prix at Anderstorp; the Finnish Grand Prix at Imatra; the Ulster Grand Prix at Dundrod; the Italian Grand Prix at Monza; and the Spanish Grand Prix at Montjuich (Barcelona). The Isle of Man TT, staged by the Auto-Cycle Union since 1907, is the world's oldest and its mountain road-racing course, site of the Senior (500-cubic-centimetre class) TT since 1911, is the world's most famous. It consists of well-surfaced highways running through the countryside and parts of two towns; hairpin corners and steep rises and falls abound, the highest point being 1,400 feet (425 metres) above sea level. The international events cover seven laps, or 264.13 miles (424.99 kilometres).

In addition, there are races with a Grand Prix title that do not count for the world championship, such as the Grand Prix of Canada, Grand Prix of Greece, and the Adriatic Grand Prix (Opatija, Yugoslavia).

Local clubs and organizations all over the world have their own international road-race meetings throughout the year. The FIM calendar shows over 100 such events sanctioned by national clubs and run in accordance with FIM rules and regulations.

Endurance and sprint races

Endurance races are road races with machines that are eligible in Grand Prix events, plus production-type sports models. Six races are selected as qualifying events for the FIM Endurance Cup: the Southampton 500 Miles Cup (England), six hours of Österreichring, 24 Hours of Montjuich, International six hours at Zandvoort (The Netherlands), Bol d'Or at Montlhéry, and the International six-hour race at Schwenningen (Germany).

Sprint races are short road races with Grand Prix type motorcycles. They do not qualify for any championship. The two principal sprint races on the calendar are both held in The Netherlands (Uden and Zandvoort).

Speedway racing. Speedway motorcycle racing is conducted on short, flat, oval dirt tracks for short distances. The machines are ultralightweight, with small tanks and small brakes.

The Speedway World Championship is not decided on a points basis but by elimination (as in drag racing). The finals decide the eventual winner(s). There are 11 qualifying events (one each in Germany, Italy, Yugoslavia, Hungary, East Germany, Sweden, Denmark, Poland, and Czechoslovakia, and three in England; but finals are held in a different country each year (which is then omitted from the following year's list of qualifying events). The national clubs sanction about 50 international speedway races organized by themselves or by local clubs under FIM rules. In addition, local clubs stage local events.

Grass- and sand-track speedway meetings are medium-speed events for machines of all types and sizes, with and without sidecars, competing within their respective classes. One-kilometre sand tracks (as used in harness [horse-and-sulky] races) are used in Scandinavia and Germany. The four World Sand Track Championship races are run on one-kilometre tracks. Shorter sand tracks and grass tracks are used for lesser international events.

Reliability trials. Motorcycle trials are slow, on-and-off-highway events of long duration, with machines in "production" trim; as in automobile rallies, speed is not the determining factor, but a time schedule must be maintained. There are classes for solo riders and machines with sidecars. The most famous are Britain's International Six-Days Trial, and the Scottish Six-Day Trial. A series of five two-day trials, each in a different country, forms the European Reliability Trial championship. The venues are Eschwege in West Germany, Zschopau in East Germany, Krkonosska in Czechoslovakia, Bergamo in Italy, Barcelona in Spain. The national clubs (under FIM rules) sanction a number of other international trials that do not count toward any championship. The Scottish Six-Day Trial is one of these.

Moto-cross. Moto-cross is a type of motorcycle race popular in Great Britain and Europe that takes place off the road, in rough terrain or open plains, of short distance and duration. Courses typically include steep up-and-down grades, sharp turns, straight stretches, bumps, stretches of loose sand or dirt, and mud holes. A race consists of a specified number of laps. The machines are equipped with knobby tires for off-road traction, high ground clearance, and extra-low gearing for pull power at low speeds.

Moto-cross racing has been formally organized in Europe since 1952. The moto-cross world championship is for motorcycles up to 500 cubic centimetres (30 cubic inches), and each participating nation has its own event that gives points toward the championship. There are 12 qualifying events: the Italian Championship Moto-Cross; the Austrian Moto-Cross Grand Prix at Sittendorf (Vienna); the Swedish Championship Moto-Cross; the Finnish Moto-Cross Grand Prix at Tikkurila; the Czechoslovakian Moto-Cross Grand Prix; the Soviet Moto-Cross Grand Prix at Kishinev; the East German Moto-Cross Championship; the British Moto-Cross Grand Prix; the German Moto-Cross Grand Prix; the Belgian Moto-Cross Championship; the Moto-Cross Grand Prix of Luxembourg; and the International Moto-Cross at Sint Anthonis (The Netherlands).

Championship moto-cross events

A minor league of moto-cross is restricted to 250-cubic-centimetre (15-cubic-inch) engines, and FIM has 12 such events counting for a secondary championship. The

Contestants manoeuvring over rough terrain in International Moto-Cross Championships of 1969, held at Kishinev, U.S.S.R.
London Daily Express—Pictorial Parade

most important venues are Payerne in Switzerland, Karlovac in Yugoslavia, Bergharen in The Netherlands, and Launsdorf in Austria. FIM and the national clubs also sanction a series of international moto-cross events organized by local clubs. These do not count toward any championship. In addition, local clubs everywhere organize national or regional events throughout the season.

With the approval of the FIM the American Motorcycle Association introduced an eight-race moto-cross series called the Trans AMA in the United States in 1970.

Drag racing. An outgrowth of the speed record trials included in the program of the Bonneville (Utah) Speed Week, motorcycle drag racing is patterned after automobile drag racing and, like it, is an American innovation. It consists of a series of acceleration, or top speed, contests, each between two racers, over one-quarter mile (one-half kilometre) on a straight, smooth-surfaced course. As in the case of auto dragsters, there are no restrictions as to frames, wheels, tires, engines, or most other components; the vehicles are raced in many classes; and there are both gas (gasoline-powered) and fuel events (see also above *Modern automobile sports: Off-road racing*).

Ice racing. Ice racing is the arctic equivalent of speedway racing. The machines resemble each other, except that spiked (studded) tires are used in ice racing. Instead of dirt tracks, ice races are held on ovals from which the snow has been plowed on frozen lakes, or on ice-covered stadium tracks.

The ice racing world championship is a six-stage duel between Sweden and the U.S.S.R. In addition to the championship events, Austrian clubs and the Motorcycling Federation of the Soviet Union stage a number of international ice races.

OTHER EVENTS

Rallies. Motorcycle rallies fall into two categories: first, events where route following and regularity count; and second, touring concentrations with low emphasis on competition. There are 23 important international rallies in Europe (often called reliability trials but not to be confused with off-highway trials). They include the Norddeutsche 2-Tagefahrt, the Austrian Aplenfahrt, the Saale Cup (East Germany), 24-hours of Liège at Zolder, the Cannes–Geneva–Cannes rally, and the Coupe de l'Armistice.

International rallies and touring concentrations

Touring concentrations have become the special preserve of the Italians, the Dutch, and the Belgians, and ten major rallies are held each year in towns such as Modena, Rimini, Alessandria, and Rome, Italy; Gand (Ghent) in Belgium, and Eefde in The Netherlands.

Hill climbs. Motorcycle hill climbs are uphill-road races for Grand Prix and sports-model machines in which each rider races against the clock. Only one machine is in motion at any time. The European Mountain Trophy is decided by points gained in a series of five events: the International Alpland Bergrennen at Mürtzal (Austria), Cernobbio-Bisbino in Italy, Mont Ventoux in France, Monte Generoso in Switzerland, and the Sauerland Bergpreis (Germany). The national clubs also sanction and organize five secondary international hill climbs: Behamberg and Stainz in Austria; Mont Verdun (Lyon) in France; Orsières in Switzerland; and La Rabassada in Spain.

Speed-record trials. Just as there are World Speed Records for automobiles, there are World Speed Records for motorcycles. Only one machine is running at any one time and, as a rule, there are no challengers present. The runs are made in a straight line, one kilometre and one mile in opposite directions, and the average speed for the two runs becomes the official record. In addition to the absolute speed record, there are one-hour and 24-hour world speed records for motorcycles, with and without sidecars. International class records cover distances from one kilometre to 100,000 kilometres with classification according to engine displacement and whether or not a sidecar is used.

Annual maximum speed runs for motorcycles are part of the program of the Bonneville Speed Week on Utah's dry lake. Classification is in accordance with NHRA rules (not FIA). These events led to an outgrowth in the form of the dragster motorcycle and motorcycle drag racing, often using automobile engines.

BIBLIOGRAPHY. GEORGE MONKHOUSE, *Grand Prix Racing: Facts and Figures* (1950), covers the premier events of road racing from the start up to modern times. GRIFFITH BORGESON, *The Golden Age of the American Racing Car* (1966), is a thorough account of U.S. constructors and drivers between 1921 and 1931. EARL HOWE, *Motor Racing* (1951); and S.C.H. DAVIS (ed.), *Motor Racing* (1957), are anthologies covering the state of the art. LAURENCE POMEROY, *The Grand Prix Car, 1906–1939* (1949), is an articulate technical thesis. JACK C. FOX, *The Indianapolis 500* (1967), details the history of the speedway in great statistical detail. ROGER LABRIC, *Les 24 Heures du Mans* (1949), covers all pre-World War II events at the Sarthe circuit (in French). ALFRED NEUBAUER, *Männer, Frauen und Motoren* (1958) and *Herr über 1000 PS* (1959), give an insider's view of the racing world in the period 1919–55 (both in German). WILLIAM BODDY (comp.), *The Story of Brooklands,* 3 vol. (1948–50), is a chronology of events at Britain's first motor speedway from 1907 through 1939. SIR MALCOLM CAMPBELL, *Speed on Wheels* (1949); and S.C.H. DAVIS, *The John Cobb Story* (1953), relate the quest for the World Land Speed Record prior to the arrival of the gas turbine. STIRLING MOSS and LAURENCE POMEROY, *Design and Behaviour of the Racing Car* (1963), is well described by its title yet is written to be fully understandable to the layman. PIERO TARUFFI, *Tecnica e practica della guida automobilista da corsa* (1958; Eng. trans., *The Technique of Motor Racing,* 1959), is an illustrated manual for would-be racing drivers. ENZO FERRARI, *Le Mie gioie terribili* (1962) and *Le Briglie del successo* (1970), are autobiographies of the famous constructor that throw considerable light on his drivers, engineers, friends, and enemies. NORMAN SMITH, *Case History* (1958), covers the career of all leading Grand Prix cars from 1927 through 1957. COUNT GIOVANNI LURANI, *Nuvolari* (1959), is a biography of one of the world's greatest drivers. RONALD HANSEN and FEDERICO B. KIRBUS, *The Life Story of Juan Manuel Fangio* (1956), is a smaller work of the same genre. STIRLING MOSS, *In the Track of Speed* (1957), reviews the early years of his career. GRAHAM GAULD, *Jim Clark: Portrait of a Great Driver* (1968), pays tribute to the Scottish champion. RUDOLF CARACCIOLA, *A Racing Driver's World* (1961), is a personal report on his life and activity. JOHN FITCH, *Adventure on Wheels* (1959), is the story of Fitch's racing career and also an account of the revival of road racing in America after World War II. GERALD ROSE, *A Record of Motor Racing, 1894–1908* (1949), is generally regarded as the most authentic description of motor racing for this period. The *British Racing Drivers' Silver Jubilee Book* (1952), is an anthology covering the history and technology of racing, with contributions by Sir Algernon Guinness, J.D. Benjafield, John Eason Gibson, Lord Brabazon of Tara, Rodney L. de Burgh Walkerley, Georges Roesch, R.R. Jackson, and Desmond Scannell. CHRISTIAN LACOMBE, *La Moto* (1970), is the most complete record of motorcycle racing in all its forms and the evolution of the racing motorcycle published so far (in French). MICHAEL COSTIN and DAVID PHIPPS, *Racing and Sports Car Chassis Design* (1961),

is a handbook for racing car engineers and constructors. MICHAEL HENDERSON, *Motor Racing in Safety* (1968), outlines the need for safety improvements and proposes areas and directions of action.

(J.P.N.)

Mo-tzu

Mo Ti (Mo-tzu; 470?–391 BC?) was an important philosopher and founder of a school of thought bearing his name that, with Confucianism, was, for 200 years, one of the dominant schools in ancient China. Although his name was Mo Ti, he is popularly referred to as Mo-tzu (Motze, Motse, Micius), meaning Master Mo or Teacher Mo. His school was Moism.

Born a few years after Confucius' death, Mo-tzu was a contemporary of Socrates. Mo-tzu attacked the most pressing problem confronting all thinkers in the 5th century BC, when the feudal hierarchy instituted at the beginning of the Chou dynasty (1122–221 BC) was swiftly disintegrating and the country was divided into small, constantly warring, feudal states: how to bring political and social order out of chaos.

According to tradition, Mo-tzu was originally a follower of the teachings of Confucius, until he became convinced that Confucianism laid too much emphasis on a burdensome code of rituals and too little on religious teaching, at which time Mo-tzu decided to go his own way. Confucius, from all accounts, was aristocratic by temperament and orientation and dreamed of a return of the calm and peaceful days of pomp and splendour at the beginning of the Chou dynasty. Mo-tzu, on the other hand, was drawn to the common people and looked much farther back to a life of primitive simplicity and straightforward human relations.

Mo-tzu's life, however, resembled that of Confucius in many important respects. He was widely read and well versed in the tradition of the Chinese classics. Except for a brief period when he held public office, Mo-tzu, in contrast to most Chinese thinkers, lived very simply. He spent most of his life travelling from one feudal state to another in the hope of meeting a prince who would allow him to put his teachings into practice. In the absence of such a prince, he had to be content with maintaining a school and recommending his disciples for administrative positions. He commanded respect partly because he was a teacher who took his own teachings seriously. He not only condemned offensive war but journeyed to distant states to prevent the outbreak of a war whenever he heard of such a possibility.

The *Mo-tzu*

The *Mo-tzu*, the principal work left by Mo-tzu and his followers, contains the essence of his political, ethical, and religious teachings. The gist of it is found in the three sets of chapters of its second section, which give an overview of the ten major tenets: exaltation of the virtuous, identification with the superior, universal love, condemnation of offensive war, economy of expenditures, simplicity in funerals, will of heaven, on ghosts, denunciation of music as a wasteful activity, and antifatalism. Since, following Mo-tzu's death, Moism split into three schools, the three sets of chapters may well represent the three sets of texts preserved by the three schools. The other sections might be listed as follows: I, summaries and abstracts of Mo-tzu's teachings; III, discussions on logic and physical sciences; IV, records of Mo-tzu's doings and sayings; V, a manual of military defense.

As a thinker Mo-tzu is distinctive in his insistence on methodology. He insisted that standards of judgment be established, and his criteria may be summarized as the threefold test and the fourfold standard. The threefold test reminded thinkers that the basis, verifiability, and applicability of any proposition must be analyzed; the fourfold standard reminded thinkers that one should always assess the benefits any proposition could bring to the country and the people. Benefits were defined as enrichment of the poor, increase of the population, removal of danger, and regulation of disorder. To Mo-tzu the tests and standards were indispensable. Generalizing further, Mo-tzu declared that before anything could be said to be good, it was necessary first to demonstrate what it was good for.

The cornerstone of Mo-tzu's system was universal love. If the world is in chaos, he said, it is due to man's selfishness and partiality, and the prescribed cure—in striking parallel with Christianity—is that "Partiality should be replaced by universality," for "when everyone regards the states and cities of others as he regards his own, no one will attack the others' state or seize the others' cities." The same principle was to be applied to the welfare of the family and of the individual. The peace of the world and the happiness of man lie in the practice of universal love. Many objections—its impracticability, its neglect of the special claims of one's parents—were raised against this new doctrine, but Mo-tzu demonstrated that the principle of universal love had in it both utilitarian justification and divine sanction. He spoke of "universal love and mutual profit" in one breath, and he was convinced that this principle was both the way of man and the way of God.

Mo-tzu's stand on religion makes him exceptional among Chinese philosophers. His call to the people is for them to return to the faith of their fathers. He might be said to be a revivalist, a champion of religious orthodoxy with a personal god. To Mo-tzu, there is heaven, heaven has a will, and this will of Heaven is to be obeyed by man and accepted as the unifying standard of human thought and action: "What is the will of Heaven that is to be obeyed? It is to love all the people in the world universally." Heaven not only "desires righteousness and abominates unrighteousness" but also metes out reward and punishment accordingly. The system of Mo-tzu, with its gospel of universal love and the ascetic discipline as exemplified by his own life, soon after the master's death, was embodied in an organized church with a succession of Elder Masters and a considerable body of devotees. The religion prospered for several generations before completely disappearing.

Influence

The teachings of Mo-tzu, however, continued to be held in high respect for several centuries. Down to the beginning of the 2nd century BC, writers referred to Confucianism and Moism in one breath as the two leading schools of thought. But from that time, Moism suddenly disappeared from the intellectual scene. Critics have generally agreed in admiring the high-minded character of Mo-tzu himself but considered his teachings overdemanding and contrary to human nature. It was not until the 20th century that Mo-tzu was rediscovered and his teachings reappraised.

BIBLIOGRAPHY. *The Mo-tzu*, ed. by SUN YI-JANG (1908; in Chinese), is generally recognized as the earliest and also the definitive edition of the text in the modern age. This is an indispensable work to all students of this subject. Other works include: Y.P. MEI, *Motse: The Neglected Rival of Confucius* (1934), a general study of the man and his age, his works, and his teachings, and (trans.), *The Ethical and Political Works of Motse* (1929), the only English translation of the text available; BURTON WATSON (trans.), *Basic Writings of Mo-tzu* (1967), a handy paperback from a series on ancient Chinese philosophers.

(Y.P.M.)

Mountain-Building Processes

A mountain is an area of land that is relatively much higher than the land surrounding it; thus, the so-called hills associated with great ranges such as the Himalayas would be mountains in a less formidable setting. In geology, the technical name for mountain building, orogenesis, is applied not only to the way in which existing mountains attained their high relief and elevation but to the processes that deformed the rocks of ancient mountain regions. Individual mountain elevations evolve in belts or ranges, with a succession of ranges forming a mountain system. Thus, the Himalayas constitute a mountain system within which are many ranges—such as the Karakoram Range—each of which has discrete mountain peaks —such as Mountain Godwin Austen, or K2 (28,380 feet [8,600 metres]).

Kinds of mountains

Structurally, mountains are of several kinds. They include isolated volcanic mountains constructed from accumulated lava flows and extruded igneous rocks. These may rise in only a few years (as Paricutín began to

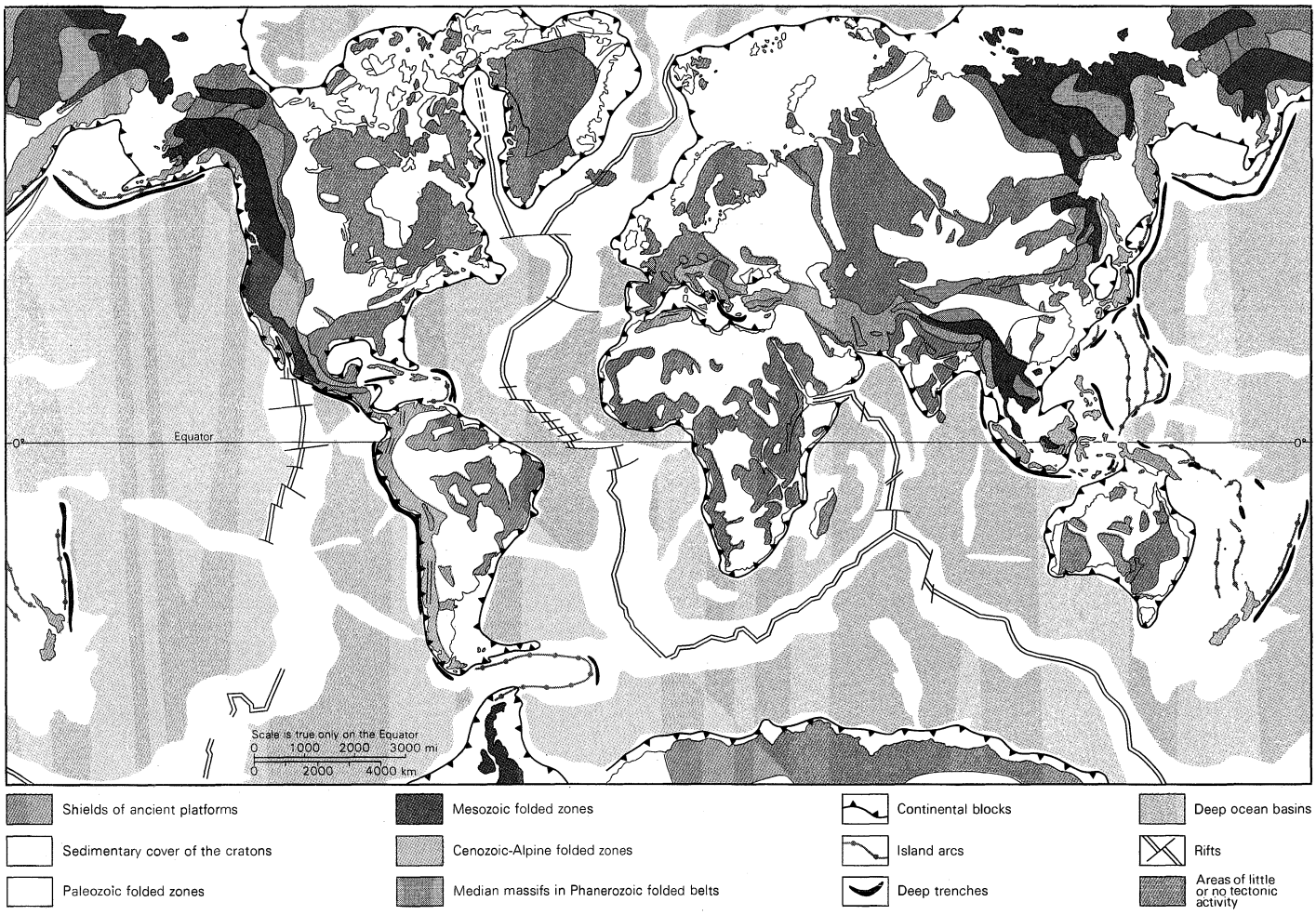


Figure 1: Structural features of the Earth's crust, particularly the regions of folding and faulting.
Adapted from V.E. Khain and M.V. Muratov, "Crustal Movements and Tectonic Structure," In P.J. Hart (ed.), *The Earth's Crust and Upper Mantle;* American Geophysical Union (Washington, 1969)

emerge on a plain in Mexico in 1943) or may grow over thousands or even millions of years. The broad Mauna Loa in Hawaii is built principally of lava flows; more conical mountains, such as Vesuvius in Italy and Fujiyama in Japan, contain a greater proportion of fragmental rocks that represent the explosive ejection of ash and other debris. Volcanoes (*q.v.*) are most abundant in narrow bands, particularly in the island arc areas that rim the Pacific Ocean. Recognizing the origin of volcanic peaks is not difficult; however, understanding the reasons for their distribution in particular areas and for the volcanic activity that caused them creates a project that requires much scientific research and study.

Another kind of mountain is formed by the tilt of blocks along such areas of separation as the rifts and fractures of the earth's crust. The structure of such a range can be simple, as in the tilted volcanic rocks bounded by fractures of the earth's crust along which movement has occurred (faults), exposed in the basin ranges of Oregon or those along the rift valleys (*q.v.*) of East Africa. In other ranges, though the topographic elevations appear to be similar, ancient mountains had to be reduced by erosion to subdued topography before faulting and elevation could produce the present ranges. The Sierra Nevada in California is of this sort.

The great mountain systems of the world comprise broad belts of varied rock types and structural patterns. Erosion through time tends to transform deformed and altered sedimentary and volcanic rocks into curved, elongated, continuous ranges, whereas plutonic igneous rocks (those formed by crystallization deep within the earth's crust) evolve as irregular elevations. Such rocks have distinctive compositions and structures that are recognizable even when only the roots of the ranges have avoided erosion. Processes that lead to rock deformation (*q.v.*) occur deep in the earth because the high temperatures and pressures required occur only at great depth. This fact encourages speculation on the nature of the earth's interior. Though volcanoes and block mountains contrast markedly with mountains produced by other tectonic processes of uplift and deformation, there are significant relationships among them; all the mountains are thought to represent aspects or phases of larger crustal movements.

Knowledge of the origin of mountain systems is derived from analysis of geologically young mountain ranges of high elevation, such as the Himalayas and the Alpine ranges of Asia and southeastern Europe, as well as from older and more deeply eroded ranges, such as the Appalachians of eastern North America and the Caledonian Belt of northwestern Europe. This article deals with mountains in terms of their basic origins, their reason for being. Additional details on the kinds and distribution of mountains are covered in MOUNTAIN RANGES AND MOUNTAIN BELTS. Related information on the processes involved in mountain building and their causes and effects is contained in EARTH, STRUCTURE AND COMPOSITION OF; ISLAND ARCS; RIFT VALLEYS; OCEANIC RIDGES; ROCK DEFORMATION; and SEA-FLOOR SPREADING. EARTH, GEOLOGICAL HISTORY OF contains a summary of mountain building through time.

DISTRIBUTION OF MOUNTAIN BELTS

Relation to global tectonics. The younger mountain belts of the earth appear to be related to global tectonics —that is, to the movements of large plates of the earth's crust on the underlying mantle (Figure 1). The conti-

nents and the oceans exhibit thicknesses of crust that are different from each other. The density of rocks in the continents is less than that of those in the crust below the oceans, and the crust above the Mohorovičić Discontinuity (a zone separating the crust and the denser underlying mantle within which the velocity of seismic waves changes) is relatively thicker in continental areas (40 kilometres or less) than it is below ocean basins (*q.v.*; 10 kilometres or less). It has been a basic tenet of earth scientists that elevations of the earth's surface are related to isostasy (compensatory balance of low regions of dense crust and lighter crustal segments). The gravity, and presumably the density, below oceanic trenches is unusually low. Although geophysical measurements are somewhat variable, the thickness of crust beneath the North American continent is generally greater beneath the mountains than beneath the plains. The thickest continental crust, nearly 70 kilometres (40 miles), is reported to lie beneath the Alps.

Crustal plates, rift systems, and subduction zones

Continental and oceanic crust is separated into plates that are bounded by fracture zones. The plates separate along zones of spreading; it is these zones that form the world rift system. Rift zones classically are exemplified by the Mid-Atlantic Ridge, but some, such as those of the East African Rift Belt, pass into continents. The East Pacific Rise is also of this type; it enters North America at the Gulf of California and passes beneath continental rocks in the western United States. These rift zones are the sites of relatively shallow-focus earthquakes (*q.v.*)—that is, those tens of kilometres deep or less. The plates descend into the mantle and disappear along margins termed subduction zones. These zones are represented by the distribution and concentration of deep-focus earthquakes, which increase in depth along planes that emerge in ocean trenches and are tilted toward and beneath continents or island arcs (*q.v.*). Such seismic bands are called the Benioff zones. On the western margin of South America, these earthquake focuses descend from the Andean Trench under the Andes mountain chain to a depth of 700 kilometres (430 miles).

Most of the world's volcanoes are in the islands or continents bordering the trenches of the Pacific Ocean and northeastern Indian Ocean, where, it is believed, the earth's crust is shortened by consumption of descending plates. These are the same belts in which deep-focus earthquakes descend landward along the Benioff zones. The volcanoes scattered across southern Asia and north of the Mediterranean Sea, such as the biblical Ararat (16,700 feet [5,100 metres]) and Etna (10,800 feet [3,300 metres]) and Vesuvius (4,260 feet [1,300 metres]), are along geologically more complex belts of compression.

A second site of volcanoes is along rift zones in the spreading crust. Those of Iceland and of the Canary Islands lie along the North Atlantic midocean rift; Iceland also has fault-bounded blocks. In East Africa, Kilimanjaro (19,350 feet [5,900 metres]) in Tanzania is along such a rift zone with volcanic block mountains. Volcanoes like those in the mid-Pacific Hawaiian Islands and in the widely distributed oceanic seamounts are not related to the main plate margins, which coincide with the zones mentioned above.

Older mountain belts. Although some of the younger mountain chains of the earth are along continental plate margins, older ones have little apparent correlation with present global plates (Figure 1). It would appear that some, such as the Appalachian-Caledonian Belt, were continuous at some former time and were severed by the opening and spreading of the North Atlantic Ocean Basin through continental drift (*q.v.*). They may have had relationships to ancient global plates in a manner analogous to the relationship of younger mountains and modern plates. The contemporary view is that the crust beneath all present oceans is quite young, not as old as 200,000,000 years.

TECTONIC MOUNTAIN BELTS

The tectonic mountain chains of the earth have such varying compositions and structures that only a few generalizations can be applied to many of the belts. Some of their distinguishing characteristics are used to recognize and locate the roots of ancient mountain chains, now eroded to low level, for the belts are of many ages (Figure 1). At sufficient depth, all parts of continents contain rocks that have compositions and structures resulting from orogenesis (mountain building).

Fold and fault structures. The discrete mountain ranges on the periphery of the tectonic mountain system are formed by the erosion of folded sedimentary rocks. The folds generally face away from the centre of the system; that is, a plane bisecting the folds dips (inclines) toward the interior of the system. In the Appalachians of eastern United States and Quebec and in the Rocky Mountains in Colorado and Alberta, for example, they face toward the continental interior. The folds are broken by faults where strata have become steeply inclined. Low-angle faults frequently separate the folded rocks from an underlying, relatively undeformed basement (literally, bottom rocks) and its rigid cover. The Jura Mountains in the Alpine region contain folds that have moved over salt-bearing strata, and the Cumberland Mountains of eastern Kentucky and Tennessee have folded strata that moved westward on easily deformed and sheared shales.

In many mountain systems, areas of exotic (not common, in terms of local rock types) strata lie in thrust sheets (planes of movement along which some strata are thrust up and over other beds) on the margins of the chains. They are called allochthons (from without or elsewhere) because they were transported from other areas and their lithology (rock type) contrasts with that of rocks of the same age in the autochthon (the local mass or region) below the underlying thrust faults. In New York and western New England and in western and northern Newfoundland, the basement and its cover of Lower Paleozoic strata was buried under sheets that moved westward with a sliding motion from rising welts to the east. On the northern flank of the Alpine System in Austria and Switzerland and on the south flank of the Himalayas in India and Nepal, the folds are most complex. Some have been forced into nearly horizontal attitudes and then thrust along fault planes far from the core of the range. Masses that have been moved in this manner and have then become isolated by erosion are called klippen. Many of the thrust sheets are believed to be gravity slides (the movement was, in part, downslope) generated from rising mountains in the internal parts of the mountain systems.

Age of structures and multiple orogenies

The folds may be of one or of several ages. Thus, the deformation of the Appalachian Mountains from Alabama to Pennsylvania was an event of the Late Paleozoic Variscan orogeny. Similar folds in western New England and Quebec were formed in the Acadian (Shickshockian) orogeny of Middle Devonian time, as well as in an earlier mid-Ordovician phase of the Taconian orogeny. Often folds representing stages of deformation of divergent trends may exist within the strata in a single range.

Structures formed through the erosion of folded and faulted sedimentary rocks on the borders of certain mountain systems have been called the externides, in contrast to the internides, the central part of such mountain belts. The latter region is characterized by the presence of abundant volcanic rocks in the stratigraphic successions (sedimentary layers), severe deformation and metamorphism, migmatization—the alteration of rocks to crystalline texture—and invasion by plutonic rocks. Although these generalizations apply to some mountain systems, they are inadequately descriptive of others.

Rock types of the mountain belts. The oldest rocks within the internides generally are deformed sedimentary and volcanic rocks that delineate the former geosynclines. When they are characterized by volcanic rocks, such geosynclinal belts are called eugeosynclinal. The volcanic rocks within the belts had many sources. There are ophiolites and other submarine pillow lavas that formed on deep sea floors, as well as extrusive lavas and agglomerates (composed of fragments and particles of volcanic rock) resembling those in volcanoes that form

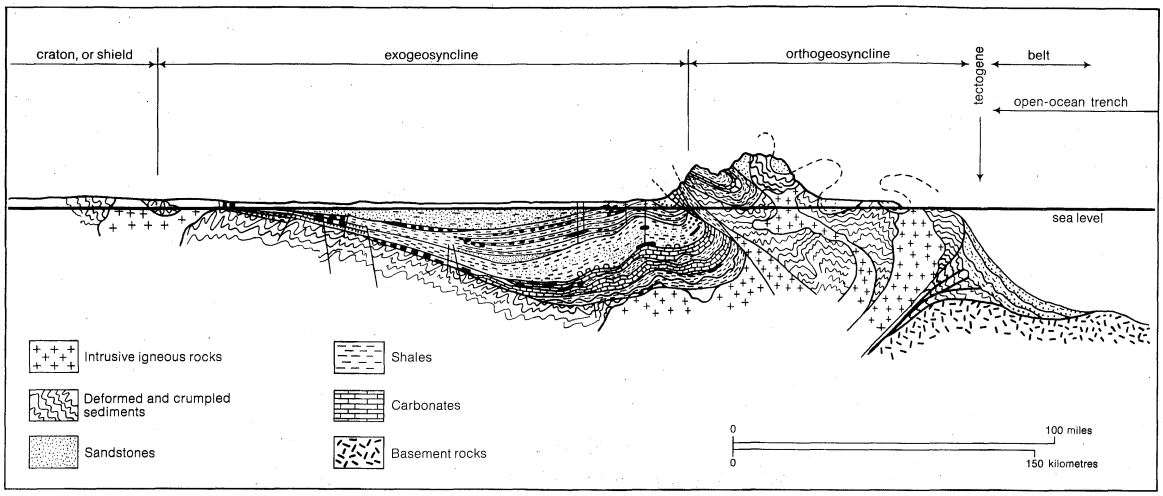


Figure 2: Idealized section across Indonesian–Japanese arc.
Printed with the permission of *The American Association of Petroleum Geologists Bulletin* and the author

modern island arcs. Elsewhere, the belts contain gneisses (a metamorphic rock of banded structure) that show great alteration and deformation, and intrusive igneous rocks. Some belts are stratigraphically overlain by rock successions that have been less deformed. Such tracts represent multiple phases of orogeny within the long and complex history of the belt. They were land areas produced by deformation among the deeply subsided troughs, volcanic islands that comprise the eugeosynclinal belts. A few eugeosynclinal belts contain melanges, a chaotic assemblage of rocks, Shales, graywackes, and volcanic rocks, with interbedded mudstone and boulders of sedimentary and volcanic rocks, occur in lenses (layers that tend to converge) and bands.

Plutonic intrusions

The central parts of the mountain systems are the principal sites of the intrusion of plutonic igneous rocks. Some of these have been intruded within the belt in early phases of orogeny. Not only are they overlain by sedimentary and volcanic rocks that were deformed in later phases but they formed the surface of the uplifted land areas that were eroded to yield igneous rock debris in the succeeding sedimentary deposits. The plutonic rocks were partly emplaced while deformation was progressing and some of them thus bear the structural effect of the deformational stresses. Some so thoroughly assimilate the host, or country, rock as to form migmatites (rocks formed by the mixing of magmatic fluids with pre-existing rock). The later intrusions are relatively more massive and discordant. In many mountain systems, the batholiths (large intrusive bodies of rock) and smaller stocks are exposed in tracts of thousands of square kilometres. Batholiths are generally thought to be intrusions of very great volume that spread downward in the earth's crust. An alternative interpretation is that they are the products of solidification of magma (the liquid melt from which igneous rocks crystallize) that rose rather near to the surface because its density was less than that of the surrounding crust.

THEORIES OF OROGENESIS

Geosynclinal hypothesis. In 1858 it was observed that the thickness of the folded Paleozoic rocks in the Appalachian Mountains of the eastern United States is much greater than that of equivalent undisturbed rocks in the Mississippi Valley. Such belts of relatively thick sediments deposited as the site of deposition subsided were first called geosynclinal in 1873. From these observations came the doctrine that geosynclinal belts become mountain ranges. The American concept was based on the Appalachians; the geosynclines of North America were thought to be separated from permanent ocean basins by borderlands of ancient crystalline rocks. The European analogue was thought to be in the Alpine System, which had quite different characters.

Early in this century, there appeared an interpretation of mountain ranges in which deformation was considered to be bilaterally symmetrical, with overfolds and thrusts spreading out on both sides of an inner metamorphic and intruded zone as the original geosyncline was compressed. More elaborate models with multiple troughs and rises were later developed by structural geologists. Though each of these may have served to explain the principal features of a specific region, the specific analogies are doubtfully applicable to other ranges and systems. Many mountain systems have only a single direction of overfold and are not explainable in terms of these models.

Eugeosynclines and the tectogene

In the late 1930s the main geosynclinal belt of the Appalachians was interpreted as lying in the position of the supposed borderland belt, to the east of the present Appalachian Mountains, where the Appalachian Geosyncline had been first postulated. Rock sequences in the North American continental border region suggested the presence of deeply subsiding troughs with pillow lava flows, associated with volcanic rocks that might have formed in ancient volcanic island chains. Other sections suggested the former presence of tectonic lands, namely land areas that had been raised within the belt. Such belts in which volcanism had been prevalent were called eugeosynclinal and were said to be truly geosynclinal in the sense that they were the principal mobile belts in crustal tectonics.

The consistent presence of ultramafic (rocks consisting almost entirely of iron-magnesium silicate minerals) serpentinized bands was observed within such eugeosynclinal belts as well as in island arcs. At the time, they were associated with the concept of a tectogene, posited as a downward buckling or invagination of relatively light, thin crust descending deeply into the more dense earth mantle, as below oceanic trenches. The hypothetical tectogene was used to account for the surprisingly low gravity near such troughs; the doctrine of isostasy would require belts at very great depths to contain dense rocks of high gravity.

It was believed that this emplacement occurred at the beginning of deformation in each tectogene trough system. Serpentinized ultrabasic rocks (rich in iron and magnesium) were emplaced very early in the orogenic history and they evidently intruded rocks at shallow depths, because they are found within rocks interpreted as being parts of thrust sheets that originated from gravity slides along the borders of orogenic belts. A classic case is that of the ultrabasic rocks of the Bay of Islands, western Newfoundland, which are in masses transported in Middle Ordovician time.

Miogeosynclines and the craton

Eugeosynclinal volcanic belts contrast with linear, nonvolcanic belts that are termed miogeosynclinal; the latter are nearer the stable shield, the craton. The typical miogeosynclinal belt was thought to have thousands of feet of sediments, principally sandstones and carbonate rocks,

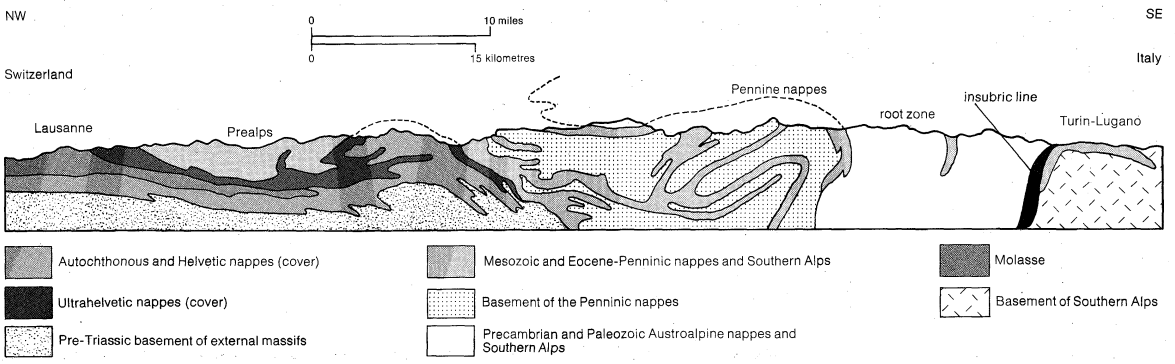


Figure 3: Generalized structure section through the Alps.
From J. Cadisch, *Geologie der Schweizer Alpen*

laid in a belt of differentially warped continental crust 400 kilometres (250 miles) wide. The folded Appalachians in the geosyncline as originally defined, were formed in a miogeosynclinal belt, extending into an ancient depression that subsided within the continent or craton as it received sediment eroded from the rising mountains within the eugeosynclinal belt. At the time that the eugeosynclinal belts were first recognized, it was assumed that ocean basins such as the Atlantic were earth features that had persisted from the very early history of the earth, and that the eugeosynclinal belts developed along the oceanic margins.

Thus, the conventional view in the early years of the 20th century in America, at least, was the geosynclines were independent of ocean basins (*q.v.*) and were intracontinental belts separated from oceans by borderlands. It was next recognized that instead of continent-like borderlands, volcanic-bearing eugeosynclinal belts had existed. The severe metamorphism and intrusion of these belts had led to the erroneous assumption that they had been long-lived continents of ancient rocks. In other continents, particularly in Europe where the Alps had been well studied, there had been suggestions that such volcanic geosynclinal belts might be oceanic, but there was no available way of substantiating such views. The so-called Steinmann Trinity of ophiolites, radiolarian cherts (siliceous rocks formed from the remains of marine organisms), and serpentine was clearly suggestive of oceanic depths, however.

Around 1950 the doctrine of the permanence of ocean basins prevailed. The accretion theory of continental growth was developed (see CONTINENTS, DEVELOPMENT OF) after knowledge of ages of intrusion had been gained by absolute age dating. The continents were thought to expand as orogenesis consolidated the rocks at the margins of ocean basins. This convention was challenged by the discovery that the oceans are geologically quite young. The possibility that other oceans of the past have been severed and destroyed also has been suggested as an alternative doctrine.

Sea-floor spreading and plate tectonics

About 1960 it became apparent that the earth is divisible into large blocks that separate at rifts within oceanic ridges and their offshoots, with which, together, they form the world rift system. Oceans broaden as ocean floors and contiguous continents spread from the midocean trenches. Reversals of magnetism in oceanic cores, when dated by paleontologic and radiometric methods, have been found to replicate quite closely the magnetic anomaly bands that roughly parallel the oceanic ridges. The greater age of ocean floor at distances farther away from the midocean ridges was subsequently proved by the dating of deep sea cores. Basically, the ocean floor at the sites of midocean ridges is relatively recent material, a few thousands of years old, whereas rocks as old as Cretaceous (65,000,000 to 136,000,000 years) in age occur farther away, toward the oceanic margins. Seismic observations show that zones marginal to oceanic trenches descend beneath adjoining continents where deep-focus earthquakes occur to depths of 700 kilometres. By 1968 the concept of global tectonics postulated oceanic crustal plates spreading apart at midocean ridges as in the Atlantic, and descending below continental margins as on the west of South America or below other oceanic plates of midocean volcanic arcs, as in the Pacific arcs. Although such ideas had been expressed many years earlier, their demonstration had depended on advances in geophysical and oceanographic techniques. These developments bear on the understanding of the origin of mountain systems.

The doctrine of the permanence of ocean basins was superseded by the knowledge that oceans are dynamic and ephemeral. The concepts of the relative youth of ocean basins and of the mobility of crustal plates have provided means by which the varied characters of mountain systems can be explained. Mountain building is related to the processes that both produce and destroy crustal plates. Recent volcanic island arcs and mountain belts, for example, exhibit strong correlation with belts of high seismicity as if to suggest that both are dependent on larger earth movements. Deep-focus earthquakes occur along the subduction zones where crustal plates descend. The relations of some recent mountain systems to present crustal structures is thus apparent. Analogies with rocks and structures in other belts lead to the conclusion that they formed along and within similar crustal plates in patterns that have been superseded, for oceans have been formed by spreading and have been destroyed by loss of margins as well as by compression. Mountain ranges have such great diversity that generalizations must be made with great caution. The causes of many of the changes through time are now speculative, but observations confirm the ephemeral nature of many of the larger earth features.

DEVELOPMENT OF MOUNTAIN SYSTEMS

Geosynclinal belts. Mountain systems generally have relatively greater thicknesses of stratified rocks than do the adjoining plains. The sediments are principally of marine origin and represent deposition in a belt that subsided deeply as deposition progressed. Such belts are geosynclinal. These principal mobile belts of the earth, the orthogeosynclinal belts, are separated into eugeosynclinal (truly) and miogeosynclinal (lesser), the former distinguished by its volcanism. Eugeosynclinal belts are made up of several elements in varying proportions: (1) basic lavas such as pillowed basalt, ophiolite, radiolarian chert, and graywackes (*q.v.*) deposited on the sea floor; (2) great thicknesses of intermediate lava (not as rich in iron and magnesium as basic or ultrabasic lavas) and fragmental volcanic rocks; and (3) sedimentary, volcanic, and plutonic rocks that are metamorphosed to varying degrees. Eugeosynclinal belts were subjected to plutonic intrusions and severe metamorphism and alteration of the rocks (forming migmatites) in recurring phases or pulses of orogenesis. The Paleozoic and Mesozoic rock successions of the Pacific coast of North America are prototypes of this sequence. The several elements of the eugeosynclinal belts, as previously stated, may have contrasting significance in terms of global tectonics. Though eugeosynclinal belts characterize the central cores of mountain systems, they are notably narrow or even absent in some of the major systems.

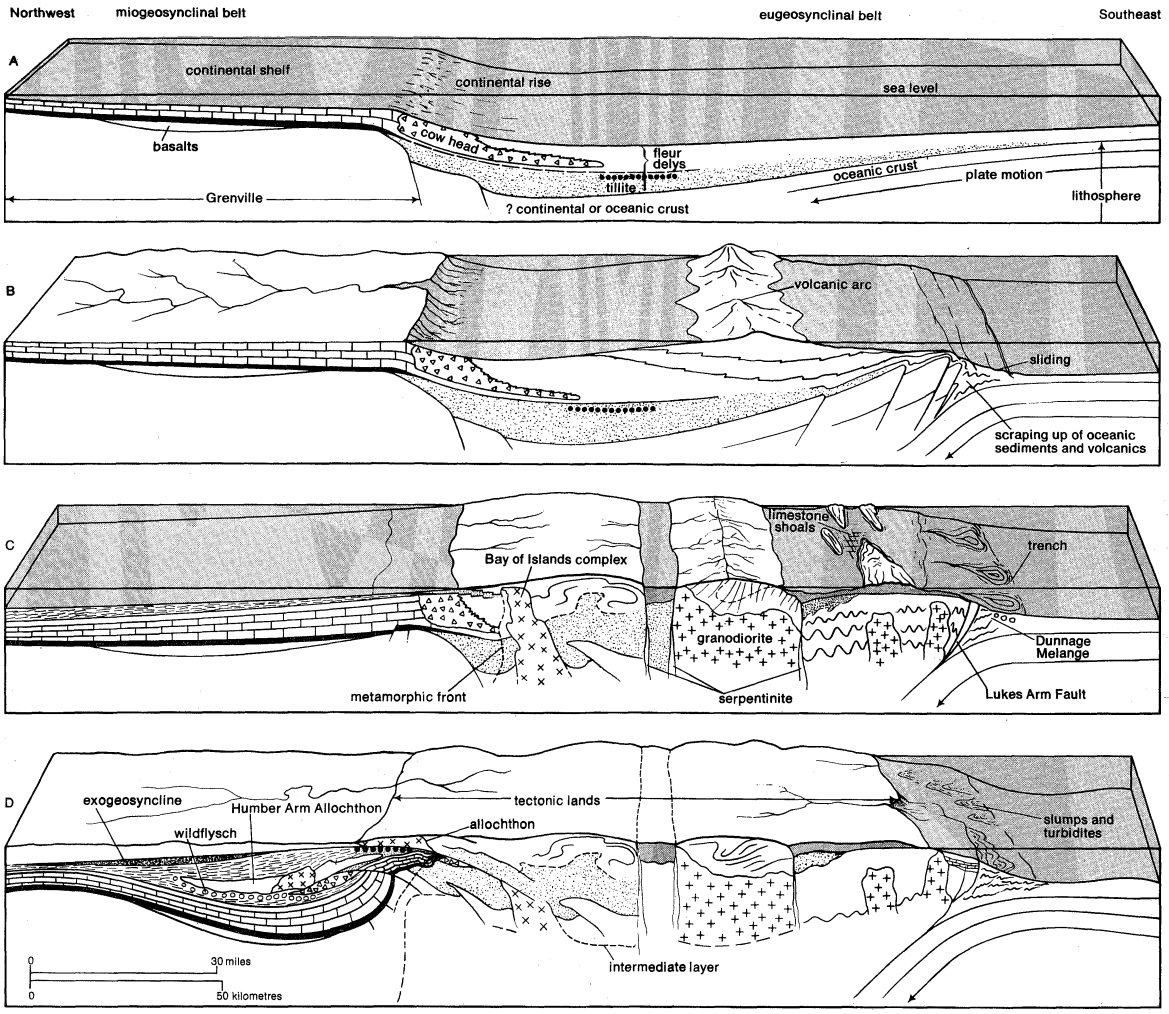


Figure 4: Four stages in the evolution of the continental margin of North America in Newfoundland during Ordovician time (initial stage is at top).
From John Bird and John Dewey, "Lithosphere Plate-Continental Margin Tectonics and the Evolution of the Appalachian Orogen," *Geological Society of America Bulletin*, vol. 81 (1970)

Flysch and other eugeosynclinal deposits

Sedimentary rocks (*q.v.*) associated with the lavas and fragmental volcanic rocks in eugeosynclinal belts are quite varied, reflecting not only the differing sites of deposition but the variety of source rocks accessible to erosion. Some eugeosynclinal belts have alternating coarse- and fine-grained poorly sorted sediments; these are termed flysch deposits to connote transportation by turbidity currents (water–sediment mixtures that are a variety of density currents [*q.v.*]) and rapid deposition. They formed from the erosion of rocks of tectonic islands within the belt or from volcanoes and islands of volcanic rocks and were transported into deeper water. Some eugeosynclinal belts have areas within them that are quite lacking in volcanic rocks, and are not restricted to the eugeosynclinal belts. Wherever deposition was very slow because of distance from source areas and subsidence was great, deep-water-laid "condensed" sections of consolidated clays and siliceous rocks (*q.v.*) such as radiolarian cherts are found. When the source was well-sorted sands on stable adjoining continental platforms and the sediments were carried by submarine currents and sand flow into deep water, the dominant rock is orthoquartzite (a nearly pure quartz sandstone), despite the fact that this rock is typical of miogeosynclinal deposits. Although flysch is a common constituent of eugeosynclinal belts, it is neither invariably present nor limited to them. The type of sediment depends on the exigencies of source and transport.

The lavas, often pillowed, are in associations with thick sequences of varied sedimentary rocks; this suggests that they have flowed into seas in great volcanic troughs. In other places, thousands of metres of lavas, agglomerates, and tuffs are interbedded with marine fossils of shallow water habitat. Such piles of volcanic rocks are laterally limited; they grade into nonvolcanic marine sediments. As they are distributed in belts and are partially subaerial, they may represent island archipelagoes, or arcs, in the ancient seas.

Melanges and chaotic deposits

Occasionally there are melanges, mudstones with some very large boulders, and chaotic mixtures of mud and rocks. Such deposits generally are produced by blocks and masses of sediment that slump from unstable depositional sites into a deep trough. In northern Italy and western Newfoundland, melanges are within gravity slides that have been carried from the margins of the central mountain core regions out onto the externides. Other melanges—*e.g.*, in the Franciscan terrane of late Mesozoic age in California and Oregon, the Lower Paleozoic Dunnage melange of northeastern Newfoundland, and the exotic bouldery mudstones of the Tertiary "root zones" of the northern Alps and the Indus Suture of the Himalayas—seem indigenous to the central regions. Among the rocks in the Franciscan Melange are glaucophane schists (metamorphic rocks deformed under the high-pressure, low-temperature conditions that prevail deep in the crust). There are bands of such schists in two metamorphic belts on Honshu Island, Japan. These frequently have ultramafic intrusions and often are serpentinized.

Subsidence and deformation of deposits

The rocks have been so deformed and altered by later disturbances that it is rarely possible to reconstruct details of the basin of deposition. Abrupt changes in thickness have led to the conclusion that some were not simple troughs but subsided differentially across submarine rifts.

The occasional melanges and the presence of ultramafic rocks and glaucophane schists suggest that these may have been in belts where oceanic floor slid below other plates along subduction zones, like ocean trenches. Perhaps the volcanic piles were in adjacent island arcs where the ocean's floor is consumed as it descends. Or they may have risen and flowed out along rifts in the ocean floor. Eugeosynclinal belts had a complex initial submarine topography.

Bands in the subsiding belt were deformed and raised while remaining areas lay at great depths. As a result areas of older rocks with structures that were formed at crustal depth were raised, eroded, and covered by younger sediments that indicate the occurrence of a smaller number of deformational events. In places the older metamorphosed sedimentary and volcanic rocks were intruded by plutonic rocks, raised, and bounded by faults (rift blocks) yielding coarse sediment that spread into adjoining subsiding troughs. The raising of mountainous lands within the margin of the eugeosynclinal belts had several effects. The poorly consolidated sediments, that had been laid in the geosyncline before it was compressed and rose, slid off the mountainous elevations into the adjoining depressions as thrust sheets or gravity slides, carrying with them the ultramafic rocks that had invaded the trough to shallow crustal depth shortly before it rose. These mobilized rock masses form the allochthons that are widespread in the externides of mountain systems, overlying miogeosynclinal prisms or with their fronts lying buried within.

The miogeosynclinal belts are typically very broad bands on the margins of the continental shields or cratons (stable continental interiors). Quartz sandstones and carbonate rocks are deposited on basement that subsided more rapidly than the craton, thus accommodating greater sediment thickness. They are exemplified by the Paleozoic sequences in the Nevada–Utah region of the western United States. This belt was deposited on a crust that, as the variable thickness demonstrates, was differentially deformed into swells and swales, domes and basins. These miogeosynclinal marginal deposits are quite distinctive and should not be confused with the nonvolcanic, oceanward-subsiding wedges of marine sediment along continental margins. Some of the principal mountain systems of the world, such as the highest ranges in the Himalayas, are of rocks deformed and raised from these nonvolcanic belts.

At times when mountains rose within the eugeosynclinal and miogeosynclinal belts, thick wedges or lenses of clastic (sandstones and shales, principally) sedimentary rocks were spread into the miogeosynclinal belts and cratonal margins as exogeosynclinal (outside or beyond the principal geosynclines) belts or foredeeps. The sediments in the Appalachian Mountains and Allegheny Plateau of Pennsylvania are in such a belt and were the site of the original geosynclinal concept. In some places linear troughs formed within the craton, and associated cratonal uplifts produced coarse sediment that was carried by streams into these zones of deposition, termed zeugogeosynclines. The Rocky Mountains of Colorado and New Mexico were formed by the folding and elevation of the rock within such a region; the typical zeugogeosyncline and a succeeding exogeosynclinal belt are both present and are intracratonic.

Another type of geosyncline is the rift trough or taphrogeosyncline, which is of more limited extent and is essentially postorogenic. Deposits formed in subsiding, tilted blocks received additional sediment eroded from adjacent rising blocks. The taphrogeosyncline is exemplified by the Triassic (190,000,000 to 225,000,000 years ago) basins of the Atlantic coast of North America. The rising margins became block mountains, examples of which are the Sierras and other Basin Ranges. Thus, there is no one simple pattern of structure or mode of origin applicable to all the mountains of the earth.

That the internal parts of the geosynclinal belts underwent a succession of orogenic phases is shown by the multiple ages derived by radiometric analyses of their intrusions and the unconformities that interrupt the stratigraphic successions. In time, the belt undergoes a final paroxysm and metamorphism, and plutonic intrusions serve to consolidate the belt into continental crust.

Through their life, eugeosynclinal belts are subject to lateral as well as vertical stresses. The two flanks are at least partially brought into closer proximity by folding and faulting, as well as by less obvious internal shearing and recrystallization. Workers have recently examined the possibility that the belt was once oceanic and that some of its loss of breadth was caused by the descent of crustal plates within it, perhaps along the melange zones that are present in many of the mountain systems. Estimates of crustal shortening are few and are based on varied assumptions and methods of judgment. The British Caledonides lie between the Moine Thrust of Northwest Scotland and a structural line passing northeastward from South Wales. The Caledonides are separated by the Highland Boundary Fault into a northern band (the "orthotectonic belt"), 125 miles (200 kilometres) wide, which exhibits very great deformation and metamorphism of several phases; and a remaining 240-mile-broad (384-kilometre-broad) belt (the "paratectonic belt") of less intense deformation. It has been estimated that internal deformation of the rocks in this latter zone alone has reduced its breadth from an original width of 600 miles (960 kilometres). The fact that the northern zone is the more severely deformed suggests that the original breadth of the Caledonian eugeosynclinal belt may well have exceeded 1,000 miles (1,600 kilometres), some three times the present width.

Some of the greatest ranges contain but little record of volcanic geosynclinal belts. The highest Himalayas are made of thick sequences of sedimentary rocks, folded and faulted, well to the south of a melange (south of Tibet) that contains the rock types associated with descending oceanic plates. The Indian Peninsula, the northern margin of ancient Gondwana, seems to have pressed in a northerly direction against Eurasia so as to consume any ocean floor that separated them. Similarly, in the Alps, where the Steinmann Trinity of representative oceanic rock was defined, the area of such rocks is limited along the Alpine root zone, and is interpreted as a subduction zone passing below the continental plate to the north. The basement below the associations of ophiolite, chert, and serpentine in the Alps does not seem to be oceanic, however, because it consists of gneissic rocks and granitic intrusions.

Subsequent diastrophism. With the consolidation of the geosynclinal belt by mountain building, it becomes part of the continental crust. Further uplifting and deformation then occur. The Caledonian and Acadian belts in northwestern Europe and northeastern North America were not only deformed by folding and intrusion but they were raised to mountainous elevation in the Devonian Period (*q.v.*), 345,000,000 to 395,000,000 years ago. The attendant erosion produced thousands of feet of a stream-laid sandstone: sediments were deposited in exogeosynclines on their northwestern margin and in intermontane basins. Fracture systems trending along the mountain systems developed. Subsequently, normal, high-angle faults to the north bounded Carboniferous sediments in the Atlantic Provinces of Canada; the faulting was partially contemporaneous with deposition.

Devonian and Triassic events

Similar linear fractures continued to develop. The Newark Group of Triassic rocks extending from Nova Scotia to Georgia contains three or four miles of stream- and lake-laid sediments, as well as hundreds of feet of lava flows. Down-faulted troughs, or taphrogeosynclines, received the detritus from raised fault–scarp mountains. The fault systems have been attributed to tensional fractures formed with epeirogenic (broad, regional upwarping) movements, unwarping the mountain belt after its consolidation. The same region has several sets of transcurrent faults, a fault type in which the rocks on one side moved laterally relative to the other. The fault systems are severed at the Newfoundland continental shelf as a result of sea-floor spreading (*q.v.*) of the Labrador Strait and Atlantic Ocean in later Mesozoic and Cenozoic time.

Basin and Range and Pacific coast events

The faulted block mountains of the Basin and Range Province in the western United States have a very complex and quite different origin. They lie within a region that underwent a succession of mountain-making events from the late Devonian (Antlerian orogeny), through the middle and Late Mesozoic Nevadan and Laramide orogenies, into the Tertiary, climaxed by Late Mesozoic plutonism that consolidated the belt to continental crust. The block mountain structures may not relate to any one of these events but to the spreading of the subcrust along the trend of the East Pacific Rise, a midocean ridge that passes beneath the continent in the Gulf of California and off western Oregon. The transcurrent San Andreas Fault in California has moved right laterally—that is, the west side moved northward relative to the east—from the Mesozoic Era (*q.v.*) onward, for as much as 400 kilometres (250 miles) according to some authorities. Though this fault also lies within an orogenic region, the movement, at least that movement of later date, may be at the junction of two crustal plates. There are similar faults within the Pacific Border northward to Alaska, and so the distribution may be related to the ocean margin. In any case, the block mountains are elevations within a larger system that has rocks that were deformed and intruded long before the events that produced the block ranges.

The elevation of the Sierra Nevada, a block mountain range to the west of the Basin and Range Province, was a Late Tertiary event, following the reduction of the region to low elevation in Early Tertiary time, but the internal structures are much older. The Alps and Himalayas, on the other hand, are younger mountains, formed in belts that were still troughs receiving marine sediments in Early Tertiary time. Indeed, the present elevation of many of the earth's principal mountains is a response to events in Tertiary and even Pleistocene times.

Causes of folding. The conventional explanation of this foreshortening of parts of the earth's crust was that the continents and oceans were essentially permanent, and therefore if the earth's diameter decreased through time, perhaps because of cooling or compaction, then the crust would shorten and contract, causing the folding of crustal rocks.

Another possible explanation is that the core of a mountain range rises vertically with steel faults along the margin of the rising area that spreads toward the surface. The sedimentary rocks of the crust would then move laterally on low-angle or nearly horizontal fault planes, the extended thrust sheets becoming folded in the process. Thus, the gravity slides would acquire their structures during their descent from the rising source of their origin. Such a process does not require crustal contraction, but it would permit permanence of continents and oceans.

A third possible method by which the production of folded mountains has been explained is related to early theories of continental drift (*q.v.*). In this view, the rise and deformation of mountain ranges are attributed to the pressing of the front of a continental block against the ocean. The physical problems of the process deterred consideration of drift as a plausible theory. It is not clear why the stronger continental block should be deformed by the opposition of an oceanic crust weak enough to permit drift into it. And even though the approach might explain mountains on the front of the moving block, it does not explain similar mountain systems, such as the Appalachians of North America, on the other side of the continents. Yet, studies of remanent magnetism of rocks on different continents do seem to show relative movement among them.

Finally, a fourth explanation posits oceans formed by spreading away from a linear belt—*i.e.*, the world rift system. This produces drift of a different kind. The movement is attributed to convection below the crust which moves mantle rocks away from the rift zones and the consumption of oceanic crust by subduction along other bands in ocean trenches. Though plate tectonics seem to account for some of the features of the principal orogenic belts, other ranges seem remote from plate margins, so that the folding of their rocks appears to be due to some other compressive force.

BIBLIOGRAPHY. The following books are suggested as sources of supplementary information: P.B. KING, *The Evolution of North America* (1959), a fine summary of the history of the North American mountain regions; R.A. PHINNEY (ed.), *The History of the Earth's Crust* (1968), papers on recent advances in plate tectonics, though very little directly on mountain building; P.C. BADGLEY, *Structural and Tectonic Principles* (1965), chapters on mountain-building processes, but written before sea-floor spreading and subduction were understood; M. KAY (ed.), *North Atlantic: Geology and Continental Drift* (1969), many papers on the nature and history of the Appalachian-Caledonian belt; A. GANSSER, *Geology of the Himalayas* (1964), an excellent description of this mountain range; J. RODGERS, *The Tectonics of the Appalachians* (1970) and M. KAY, *North American Geosynclines* (1951), discussions of the nature and distribution of geosynclines; P. CHADWICK, "Mountain Building Hypotheses," in S.K. RUNCORN (ed.), *Continental Drift* (1962). Classic works on the Alps are: E.B. BAILEY, *Tectonic Essays, Mainly Alpine* (1935); L.W. COLLET, *The Structure of the Alps* (1927); and J. CADISCH, *Geologie der Schweizer Alpen*, 2nd ed. (1953). F. FUCHS, *Zum Bau des Himalaya* (1967), a monograph comparable to that of Gansser complements and adds to the discussion of this range. Excellent papers concerning plate tectonics and subduction zones are: W. HAMILTON, "Mesozoic California and the Underflow of Pacific Mantle," *Bull. Geol. Soc. Am.*, 80:2409–2430 (1969); J.F. DEWEY and J.M. BIRD, "Mountain Belts and the New Global Tectonics," *J. Geophys. Res.*, 75:2625–2647 (1970); and W.R. DICKINSON, "Plate Tectonic Models of Geosynclines," *Earth and Planetary Sci. Letters*, 10:165–174 (1970).

(M.K.)

Mountaineering

Mountaineering has not always been a sport. Throughout history, as man attempted to use and control the physical elements of the earth, mountains became symbols of his aspirations and ambitions. Because their peaks seemed close to the heavens, certain mountains—Olympus, Etna, Sinai, and Ararat, for example—had early significance for much of mankind as the mythological homes or sacred precincts of the gods. In Japan, Fujiyama is sacred to thousands of pilgrims who climb its slopes every summer, and in the Himalayas certain summits are still considered holy. Thus in May 1955 two two-man English parties stopped just short of the summit of Kānchenjunga, leaving the actual summit untrodden at the request of the Sikkimese government because it was considered sacred to the gods of Sikkim.

Early attempts to ascend peaks were inspired by other than sporting motives: to build altars or to see if spirits or dragons actually haunted once forbidden heights; to get an overview of one's own or a neighbouring country or of unexplored land; or to make meteorological, geological, or other scientific observations. Mountains presented barriers to progress, but over many centuries, the passes among them were gradually opened as routes for military expansion, exploration, communications, and trade. It is a far reach in time, however, from Moses' ascent of Mt. Sinai to the breaching of the Alps with a herd of elephants as a road to the conquest of northern Europe in the 3rd century BC, to the first recorded climbing of Mexico's Mt. Popocatepetl in 1595 by soldiers of Cortés, and a long way again to man's conquest of Mt. Everest. Centuries apart, these way stations symbolize aspects of man's approach to the mountains.

Mountaineering as a sport emerged fewer than 150 years ago, when English, French, and German climbers, together with their Swiss guides, began seriously to penetrate and ascend the mid-European Alps in the mid-19th century.

Mountain climbers must rely upon their own powers of physical endurance and judgment. Nature provides the only competition, demanding of the participant a combination of skill and self-reliance and giving him freedom from man-made rules and regulations. Mountaineering is perhaps the most intensely and purely participant of sports. Climbing mountains embodies the thrills produced by testing one's courage, resourcefulness, cunning, strength, ability, and stamina to the utmost in a situation

Selected List of the World's Highest Mountains

name	location	height		year first climbed
		feet	metres	
Mt. Everest*	Nepal/Tibet	29,028	8,848	1953
K2 (Godwin Austen)	Jammu and Kashmir	28,251	8,611	1954
Kānchenjunga I	Nepal/Sikkim	28,208	8,598	1955
Lhotse I	Nepal	27,923	8,511	1956
Makālu I	Nepal/Tibet	27,824	8,481	1955
Dhaulāgiri†	Nepal	26,810	8,172	1960
Manāslu I	Nepal	26,760	8,156	1956
Cho Oyu	Nepal	26,750	8,153	1954
Nānga Parbat I	Nepal	26,660	8,126	1953
Annapūrna I	Nepal	26,504	8,078	1950
Gasherbrum I	Jammu and Kashmir	26,470	8,068	1958
Broad Peak	Jammu and Kashmir	26,400	8,047	1957
Gasherbrum II	Jammu and Kashmir	26,360	8,035	1956
Gosainthan (Kao-seng-tsan Feng)	Tibet	26,291	8,013	1964
Annapūrna II	Nepal	26,041	7,937	1960
Himālchūli	Nepal	25,892	7,892	1960
Disteghil Sar	Jammu and Kashmir	25,868	7,885	1960
Nuptse	Nepal	25,850	7,879	1961
Masherbrum	Jammu and Kashmir	25,659	7,821	1960
Nanda Devi	India (Uttar Pradesh)	25,646	7,817	1936
Rakaposhi	Jammu and Kashmir	25,550	7,788	1958
Kāmet	India/Tibet	25,446	7,756	1931
Saltoro Kangri	Jammu and Kashmir	25,400	7,742	1962
Janu (Jannu) Peak	Nepal	25,294	7,710	1962
Tirich Mir	Pakistan	25,230	7,690	1962
Makālu II	Nepal/Tibet	25,120	7,657	1954
Minya Konka	China (Szechwan)	24,902	7,590	1932
Annapūrna III‡	Nepal	24,858	7,577	1970
Muztagh Ata (Mu-ssu-t'a-ko-a-t'e Shan)	China (Sinkiang)	24,757	7,546	1956
Annapūrna IV	Nepal	24,688	7,525	1955
Communism Peak§	U.S.S.R. (Tadzhik)	24,590	7,495	1933
Ganesh Himal	Nepal/Tibet	24,298	7,406	1955
Chamlang	Nepal	24,012	7,319	1962
Muztagh Tower	China (Sinkiang)	23,860	7,273	1956
Pumori	Nepal/Tibet	23,442	7,145	1962
Nun Kūn	Jammu and Kashmir	23,410	7,135	1953
Lenin Peak	U.S.S.R. (Tadzhik)	23,405	7,134	1928
Trisūl	India (Uttar Pradesh)	23,360	7,120	1907
Satopanth	India (Uttar Pradesh)	23,213	7,075	1947
Nupcho	Nepal	23,100	7,041	1962
Jugal Himal	India	22,870	6,971	1962
Aconcagua‖	Argentina	22,834	6,960	1897
Siniolchu	Nepal/Sikkim	22,600	6,888	1936
Huascarán¶	Peru	22,204	6,768	1908
Yerupajá	Peru	21,765	6,634	1950
Illimani	Bolivia	21,201	6,462	1898
Chimborazo	Ecuador	20,561	6,267	1880
Mt. McKinley♀	U.S. (Alaska)	20,320	6,194	1913
Mt. Logan	Canada (Yukon)	19,850	6,050	1925
Cotopaxi♂	Ecuador	19,347	5,897	1872
Kilimanjaro□	Tanzania	19,340	5,895	1889
Citlaltépetl (Pico de Orizaba)	Mexico	18,700	5,700	1848
Mt. Elbrus	U.S.S.R. (Georgia)	18,510	5,642	1874
Mt. St. Elias	Alaska/Canada	18,008	5,489	1897
Salcantay	Peru	17,123	5,219	1952
Dykh-Tau	U.S.S.R. (Georgia)	17,070	5,203	1888
Mt. Kenya	Kenya	17,058	5,199	1899
Ararat (Büyük Ağri Dağı)◇	Turkey	16,945	5,165	1829
Margherita	Uganda	16,763	5,109	1906
Mont Blanc▲	France/Italy	15,771	4,807	1786
Ushba	U.S.S.R. (Georgia)	15,400	4,700	1888
Monte Rosa	Switzerland/Italy	15,203	4,634	1855
Dom	Switzerland	14,911	4,545	1858
Weisshorn	Switzerland	14,780	4,505	1861
Matterhorn	Switzerland/Italy	14,692	4,478	1865
Mt. Whitney	U.S. (California)	14,494	4,418	1875
Mt. Rainier	U.S. (Washington)	14,410	4,392	1870
Longs Peak	U.S. (Colorado)	14,256	4,345	1868
Mt. Shasta	U.S. (California)	14,162	4,317	1854
Pikes Peak	U.S. (Colorado)	14,110	4,301	1819
Mauna Kea+	U.S. (Hawaii)	13,796	4,205	?
Grand Teton	U.S. (Wyoming)	13,747	4,190	1898
Jungfrau	Switzerland	13,642	4,158	1811
Eiger (North Face)	Switzerland	13,025	3,970	1938
Mt. Robson	Canada (B.C.)	12,972	3,954	1913
Mt. Cook⊕	New Zealand	12,349	3,764	1894
Ben Nevis	U.K. (Scotland)	4,406	1,343	?

*Highest in the world. †For some 30 years before 1832 thought to be the highest. ‡First major peak first ascended by an all-woman team (Japanese). §Highest peak in the U.S.S.R. ‖Highest peak in Western Hemisphere. ¶First major peak first ascended by a woman. ♀Highest peak in North America. ♂Probably the world's highest active volcano. □Highest peak in Africa. ◇First ascent in modern times. ▲Highest peak in western Europe. +Highest island mountain in the world. ⊕Highest peak in New Zealand.

of inherent risk. Mountaineering, however, to a greater degree than other sports, is a group activity, with each member both supporting and supported by the group's achievement at every stage.

Although mountaineering has a growing number of adherents, it does not attract anywhere near the millions that engage in such other outdoor participant sports as hunting, fishing, and skiing, for example; 500,000 active mountaineers in the world at one time is a generous estimate. But, while other sports are more popular, mountaineering is equally international in flavour—in part because of the distribution of the world's great peaks—if not in numbers of participants.

A selected list of the world's highest mountains giving their locations, heights, and years first climbed appears in the table above. See also MOUNTAIN RANGES AND MOUNTAIN BELTS.

HISTORY

During the 18th century more and more natural philosophers—the scientists of their day—made field trips into the Alps of Europe. They studied the complex structure of those mountains, conducted botanical investigations, and made meteorological observations; the area around Chamonix, France, became a special attraction to these investigators because of the great glaciers on the Mont Blanc chain. There mountaineering in a contemporary sporting sense was born when a young Genevese scientist, Horace-Bénédict de Saussure, on a first visit to Chamonix in 1760, viewed Mont Blanc (at 15,771 feet [4,807 metres] the tallest peak in Europe outside the Caucasus) and determined he would climb it or be responsible for its being climbed. He offered prize money for the first ascent of Mont Blanc, but it was not till 1786, more than 25 years later, that his money was claimed by a Chamonix doctor, Michel-Gabriel Paccard, and his porter, Jacques Balmat. A year later de Saussure himself climbed Mont Blanc, carrying the customary scientific instruments that presumably justified the enterprise. After 1850, groups of British climbers with Swiss, Italian, or French guides, scaled one after another of the high peaks of Switzerland. A landmark climb in the growth of the sport was the spectacular first ascent of the Matterhorn (14,692 ft [4,478 m]) on July 14, 1865. An English artist, Edward Whymper, who had gone to the Alps only to sketch, became one of the outstanding mountaineers of the 19th century and was the guiding spirit of this climb during which a French guide and three others of the party lost their lives on the descent.

First ascent of Mont Blanc

In the mid-19th century the Swiss developed a coterie of guides whose leadership helped make mountaineering a distinguished sport as they led the way to peak after peak throughout central Europe. That one guide lost his life on the ill-fated Matterhorn climb was symbolic of both the successes and risks of the burgeoning sport. An Italian guide, who was Whymper's chief competitor for the first Matterhorn ascent and who climbed it successfully from Italy three days later, in afteryears lost his life descending the mountain in a storm, first, however, bringing his party to safety.

Development of Swiss guide system

Since mountaineering consists of three elements, hiking in lower regions, the technical ability to climb steep rock, and comparable ability on snow and ice, elements of the sport have been frequently pursued for their own sake. A.F. Mummery, a brilliant British mountaineer of the late 19th century, who disappeared on Nānga Parbat (26,660 ft [8,126 m]) on one of the early Himalayan expeditions, was the true forerunner of rock climbing as a sport in its own true right.

By 1870 all of the principal Alpine summits had been scaled and climbers began to seek new and more difficult routes on peaks that had already been ascended; and, as

the few remaining minor peaks of the Alps were overcome, by the end of the 19th century climbers turned their attention to the Andes of South America, the North American Rockies, the Caucasus, Africa's peaks, and finally the Himalayan vastness. Aconcagua (22,834 ft [6,960 m]), the highest peak of the Andes, was first climbed in 1897, and the Grand Teton (13,747 ft [4,190 m]) in North America's Rocky Mountains was ascended in 1898. The Italian duke of the Abruzzi in 1897 made the first ascent of Mt. St. Elias (18,008 ft [5,489 m]), which stands athwart the international boundary of Alaska and Canada, and in 1906 successfully climbed Margherita in the Ruwenzori Group (16,763 ft [5,109 m]) in East Africa. In 1913 an American, Hudson Stuck, ascended Mt. McKinley (20,320 ft [6,194 m]) in Alaska, the highest peak in North America. The way was opening for greater conquests, but it would be midcentury before the final bastion, Mt. Everest, was ascended.

20th-century achievements. As the 20th century wore on, the truly international character of mountaineering began to reveal itself. Increasingly Austrians, Chinese, English, French, Germans, Indians, Italians, Japanese, and Russians turned their attention to opportunities inherent in the largest mountaineering landmass of the planet, the Himalayas. After World War I the British made Everest their particular goal. Meanwhile, climbers from other countries were making spectacularly successful climbs of other great Himalayan peaks. A Russian team climbed Stalin Peak (24,590 ft [7,495 m]), now renamed Communism Peak, in 1933; a German party succeeded on Siniolchu (22,600 ft [6,888 m]) in 1936; and the English climbed Nanda Devi (25,646 ft [7,817 m]) the same year. *The Alpine Journal* of London, a reliable chronicler of ascents, lists no peaks ascended for the first time between 1940–47, a reflection, of course, of the imperatives of World War II. The first "first" to be achieved after the war was that of Satopanth (23,213 ft [7,075 m]) in 1947 by René Dittert and his French climbing party.

At mid-20th century came the successful ascent by the French of Annapūrna I (26,504 ft [8,078 m]) in June 1950; Nānga Parbat (26,660 ft [8,126 m]) by the Germans and Austrians in 1953; K2 (28,251 ft [8,611 m]) by the Italians in 1954; Kānchenjunga (28,208 ft [8,598 m]) by the British in 1955; and Lhotse (27,923 ft [8,511 m]) by the Swiss in 1956. Beyond all of these, however, the success of the British on Mt. Everest (29,028 ft [8,848 m]) when a New Zealand beekeeper, Edmund Hillary, and the Sherpa guide, Tenzing Norgay, stood on the top of the world on May 29, 1953, was a culminating moment. This expedition, led by Col. John Hunt, was the eighth team in 30 years to attempt Everest, and there had also been three reconnaissance expeditions.

Everest, K2, and others

K2 was climbed by two members of an Italian party in July 1954. An Austrian party reached the summit of Cho Oyu (26,750 ft [8,153 m]), which was only slightly lower than Dhaulāgiri, in 1954. The summit party included Pasang Dawa Lama, who had proceeded in three days from an altitude of only 13,000 ft (3,962 m) to the summit (26,750 ft [8,153 m]).

In May 1955, a French party succeeded in getting all of its members plus a Sherpa guide to the summit of Makālu (27,824 ft [8,481 m]), a neighbour of Everest. Also in May 1955, as mentioned above, Kānchenjunga, at 28,208 ft (8,598 m) often considered one of the world's most difficult mountaineering problems, was climbed by an English expedition led by Charles Evans, who had been deputy leader of the successful first climb of Everest, the party stopping just short of the sacred summit at the request of the Sikkimese government. In May 1956 two separate Japanese parties climbed Manāslu (26,760 ft [8,156 m]) in Nepal, the world's ninth highest mountain. Mt. Everest's history did not end with the British ascent in 1953. In 1956 the Swiss placed two parties on its summit, while another Swiss team climbed neighbouring Lhotse (27,923 ft [8,511 m]), the world's fourth highest peak. In 1963 a U.S. team, led by Norman Dyhrenfurth, not only sent two parties from the South Col (high pass) of Everest to the summit but also sent a two-man team to the summit by the formidable West Ridge. Dhaulāgiri (26,810 ft [8,172 m]) was the last of the world's ten highest mountains to be climbed, in 1960, by a six-man Swiss team. The 1953 Nānga Parbat climb was unique in that the last attempt from high camp was a solo climb by Hermann Buhl, a member of a friction-ridden German–Austrian expedition. Because he reached the top late in the day, he was forced to bivouac in the open all night a short distance below the summit and next day was fortunate to regain high camp. The international aspects of mountaineering revealed by the various nationalities that had overcome the world's ten highest mountains was further reflected by the requests filed to climb in Nepal in the early 1970s. Permits requested for parties, seven of them for Everest, included 18 Japanese (one all-woman), 3 British, 2-each Australian and Italian, and 1-each Dutch, Korean, Argentinian, West German, and international.

Seeking new challenges. In the 1960s mountaineering underwent several transformations. As in the golden age of the Alps, once peaks were climbed, the emphasis moved to a search for more and more difficult routes. At the end of the decade, for instance, an international expedition began training for the assault in 1971 of the southwest face of Mt. Everest formerly regarded as unassailable. Climbers from Austria, Canada, France, Germany, Great Britain, India, Italy, Nepal, Norway, the Soviet Union, Switzerland, and the United States, including two women, participated in this enormous enterprise under the leadership of Dyhrenfurth and Col. James O.M. Roberts.

Women mountaineers

The inclusion of women in major expeditions such as this one, and all-women climbing parties, had become increasingly common. An all-woman Japanese climbing team scaled Annapūrna III (24,858 ft [7,577 m]) in 1970 and an all-woman U.S. group climbed Mt. McKinley in the same year.

Just as the most difficult routes in high mountains were increasingly undertaken, so were so-called impossible rock faces scaled through the use of artificial aids and advanced climbing techniques. Smooth faces of granite were overcome in climbs lasting days or even weeks at a time—for example, the 27-day conquest in 1970 of the sheer 3,000-ft (914 m) southeast face of the granite monolith El Capitan in the North American Sierra Nevadas by U.S. climbers.

The risk factor

With these greater skills in mountaineering and the ability to achieve seemingly impossible ascents, the risk factor in mountaineering, despite improved equipment and safety precautions, seemed to lead to more casualties. In the later 1960s the loss of leading climbers assumed depressing proportions. While dangers are many and always present for the inexperienced in the mountains, those dangers associated with bad weather, falling rock, avalanches, ice falls, crevasses, projecting shelves of snow (cornices), took a tragic toll of experienced mountaineers. Although the good climber knows how to identify dangerous conditions, those moments in the mountains for which he cannot plan or prepare sometimes prove to be fatal.

THE SPORT

For most climbers the pleasures of mountaineering are not inherent in the abused word conquest, but rather in the physical and spiritual satisfactions brought about through personal effort and contact with natural grandeur, success and failure, and ever-increasing proficiency. While all mountaineers might like to be a Hillary or Tenzing, and while there is a great difference between a massive party assembled to assault a Himalayan fortress peak and the lesser challenges of weekend cliff, summer trail, or local summit, the joys and challenges of the sport can be shared in any location.

Essential requirements

The nature of mountaineering is such that the elements of danger must always be respected. Nonetheless, the sport is not risky to those who have learned it properly and climb under appropriate conditions. The essential requirements are technical skill, judgment, and leadership ability. More men have died on low hills than on high climbs. The ability and willingness to recognize conditions that are adverse and to make the difficult decision to turn back from the climb when the odds creep up to that

point at which they are against the climbing party must be part of the leader's ability.

Hiking. Hiking has become a widely accepted and respected sport in itself and, as the first and essential element of mountaineering, must be regarded as fundamental. The ability to walk an uphill or downhill trail, rugged or smooth, kilometre after kilometre, is excellent training for the mountaineer and the best indication of his willingness to persist. Hiking is, indeed, the essential element of all climbing, for in the end mountains are climbed by placing one foot before another over and over again. The most arduous hours are those on the lower trails climbing slowly, steadily, but hour after hour, or before dawn on the tedious glacier trek. A climber who in his training has not recognized this disciplined necessity will deprive himself of many successes and pleasures on the heights.

Organization. The basic organization of the sport is the mountaineering, rock-climbing, or hiking club. Every nation with mountaineers has its own clubs, among which the Alpine Club in Great Britain, founded 1857, is perhaps the most venerable. The largest numbers of clubs are in the Alpine countries, in the British Isles including Scotland and Ireland, and in North America. Representative clubs in other areas include those in Spain and Poland, in Argentina, in India and Japan, in New Zealand, and in South Africa. Mountaineering and related clubs serve the purpose of fellowship but have the further intention of promoting responsible leadership, inculcating good practice on the mountains and encouraging members toward increased enjoyment of the hills. Most of the clubs publish annual or periodical reports, journals, or bulletins. Frequently, major mountaineering clubs participate financially in the sponsorship of major expeditions. While most clubs tend to have national or regional names, there is little chauvinism in the sport. This characteristic has been best reflected in the increasing number of international expeditions that has been noted.

Principles. While it is necessary for the complete mountaineer to be competent in all three phases of the sport—hiking, rock climbing, and snow and ice technique—each is quite different. There are wide variations within these categories, and even the most accomplished mountaineers will have varying degrees of competence in each. The good climber will strike that balance that is consonant with his own physical and mental capabilities and approach.

Rock climbing. Rock climbing, like hiking, is a widely practiced sport in its own right. The essentials of this aspect of mountaineering are often learned on local cliffs, where the teamwork of mountaineering, the use of the rope, and the coordinated prerequisites of control and rhythm are mastered. The rope, piton (a metal spike, with an eye or ring in one end, that can be driven into a crack in rock or ice), and carabiner (or snap link, a metal loop or ring that can be snapped into a piton and through which the rope may be passed) are used primarily as safety factors. An exception occurs in tension climbing in which the leader is supported by a judiciously placed series of pitons and carabiners through which the rope is passed. He is then supported on the rope by his fellow climbers as he slowly moves upward to place another piton and repeat the process. Pitons are used with discretion rather than in abundance. When pounded firmly into a crack, they will normally be secure. Various shapes of pitons have been designed to deal with various types of cracks or fissures in the rocks. Pitons are rarely used as handholds or footholds.

Aids and tension climbing

For the majority of rock climbers hands and feet alone are the essential, with the feet doing most of the labour. The layman's notion that one must be extraordinarily strong in arms and shoulders is true only for such situations as negotiation of serious overhangs. By and large, hands are used for balance, feet for support. Hands and arms are not used for dragging the climber up the cliff.

Balance is essential and the body weight is kept as directly over the feet as possible, the climber remaining as upright as the rock will permit. An erect stance enables the climber to use that fifth element of climbing, his eyes. Careful observation as one moves up a cliff will save many vain scrambles for foothold. Three points of contact with the rock are kept, either two hands and a foot or two feet and a hand. Jumping for holds is extremely dangerous because it allows no safety factor. Rhythmic climbing may be slow or fast according to the difficulty of the pitch. Rhythm is not easily mastered and when achieved becomes the mark of the truly fine climber.

The harder the climb, the more the hands are used for support. They are used differently in different situations. In a chimney, a pipelike, nearly cylindrical vertical shaft, they press on opposite sides in opposition to each other. In a crack, they can be used by the balling of a fist to jam in. On slabs, the pressure of the palms of the hand on smooth rock may provide the necessary friction for the hold.

Rope handling. It is always essential that a line of retreat be provided. To climb up beyond the ability to climb down is dangerous. Climbing down steep rock is usually harder than going up because of the difficulty in seeing holds from above and the normal reluctance of a climber to reach down and work his hands low enough as he descends. The quick way down is via the doubled rope in the technique called rappelling. The rope, one end being firmly held or secured, is wrapped around the body in such a way that it can be fed out by one hand slowly or quickly as desired to lower the body gradually down the face of the rock.

Climbing down and rappelling

Rope handling is a fine art, equally essential on snow, ice, and rock. Sufficient rope for the pitch to be climbed and of sufficient length for rappelling is needed. As a lifeline the rope receives the greatest care and respect. A good rope handler is a valued person on the climb. The art of placing pitons and carabiners and stringing the rope to best advantage for safety and economy of effort in ascending or descending, and of keeping it free of cracks where it might jam and places where it might get caught on rock outcrops or shrubbery, or be stepped on by an unwary climber is not learned easily and comes primarily through experience.

Snow and ice. Not least of the hazards of climbing a mountain are the constantly changing conditions of the high reaches where snow and ice are found. A good mountaineer will always want to increase his knowledge about snow conditions. He will wish to know what the sun will do to the snow in a given situation by the time the party is descending the mountain on its way back; he will wish to appraise when crampons (sets of spikes that can be strapped on boot soles for walking on ice) will make the ascent safer or quicker; he will become acquainted with the danger of a possible avalanche. In couloirs (steep-sided gullies) in which there is avalanche potential from above, it is necessary to move as quickly and expeditiously as possible. This condition may be the only one in mountaineering in which speed takes preference over almost every other criterion.

Crevasses on a glacier are found primarily where the glacier bends over and starts down a steeper slope or where the glacier turns a corner. Because glaciers more often than not are snow covered, it is necessary to learn how to detect crevasses simply by observing the surface of the snow. The gaping crack that occurs where the snow, instead of adhering to the lower reaches of the mountain, separates from it and fills part of the glacier field below is called a bergschrund. Bergschrunds increase the difficulty of mountaineering, particularly if snow bridges across them cannot be found.

Crevasses, icefalls, snow cornices

Icefalls on a glacier are a particular danger and have led to some of the mountains' greatest tragedies. Also hazardous are cornices of snow that resemble crests of waves and are formed by the wind on ridges. From the windward side it is hard to determine that they are there. They must be avoided because they can break off precipitously under the weight of a climber. It is necessary to cut across the top of a steep slope rather than to walk within the fracture zone of cornices. It is equally unwise to walk under a cornice if the route can be avoided.

The use of the ice-ax is extremely important as an adjunct to high mountaineering. Consisting of a pick and an adze opposed at one end of a shaft and a spike at the

Snow and ice aids

other, it is used for cutting steps in ice, probing crevasses, direct aid on steep slopes, balance as necessary, and for securing the rope (belaying). Crampons to preclude slipping are useful on steep slopes of snow and ice and in steps that have been cut. By biting into the surface they make progress possible where boots alone would not do. On many slopes crampons also render unnecessary the cutting of steps. On extremely difficult snow and ice, ice pitons and carabiners are used. The pitons when driven in are allowed to freeze in place.

In climbing long snow slopes, a tedious task, it is necessary to strike a slow and rhythmic pace that can be sustained for a long time. It is desirable, too, to make a start on the mountain early in the day when the snow is in hard condition. As in all phases of mountaineering, judgment is important when engaging in snow and ice climbing. The length of the climb, nature of the weather, heat of the sun, and potential avalanche danger must all be considered, as well as the condition of the party in experience, technical skill, acclimatization, and fatigue.

Winter climbing

Snow and ice climbing are often practiced during the winter. Conditions on small mountains can be as severe then as on high mountains in the summer. Low temperatures always present formidable liabilities. While winter mountaineering is limited as a sport, it is growing both as a training procedure and for its own particular pleasure. Such difficult north faces as those of the Eiger and the Matterhorn are now climbed in winter, but only by highly skilled parties after intensive preparations and with proper equipment.

The dangers of mountaineering are sometimes exaggerated in the public mind. Climbers who are well trained, and have good judgment, are usually able to climb for many years. The dangers of the mountain are pitfalls principally for those who are somewhere they should not be, who are improperly equipped, or who do not comprehend the dangers of the mountains or the physical condition required to cope with them. As more and more people enter the sport, however, and more and more difficult and challenging ascents are attempted, the potential of hazards increases—whether they be due to avalanche or rockfall, inexperience or carelessness, timidity or foolhardiness. As when the sport began, mountains still must be approached with understanding and respect.

BIBLIOGRAPHY. Additional information on the various aspects of mountaineering may be found in the following references: T.F. HORNBEIN, *Everest, the West Ridge* (1965); A.J. HUXLEY (ed.), *Standard Encyclopedia of the World's Mountains* (1962); W. NOYCE and I. MCMORRIN, *The World Atlas of Mountaineering* (1969); G.A. SMITH, *Introduction to Mountaineering*, rev. ed. (1967); L.J. THOMAS, *Book of the High Mountains* (1964); J.R. ULLMAN, *The Age of Mountaineering* (1954).

See also *Appalachia* (monthly), Appalachian Mountain Club, Boston; *American Alpine Journal* (annual), American Alpine Club, New York; *Alpine Journal* (semi-annual), The Alpine Club, London; and *The Mountain World* (annual), Swiss Foundation for Alpine Research.

(G.A.Sm./C.D.S.)

Mountain Ranges and Mountain Belts

A mountain is a tract of land rising considerably above the surrounding surface. Mountains are usually found connected in chains or ranges. A relatively flat area elevated above the level of its surroundings is called a plateau. Mountain belts are elongated groups of mountain ranges and plateaus, broken by valleys and basins. The larger mountain belts of the world are aligned in two great systems, one circling the Pacific Ocean (Circum-Pacific Mountain System), the other spanning an east-trending zone from the Mediterranean Sea to Indonesia (Tethyan or Mediterranean–Himalayan Mountain System). Subsidiary mountain belts are those unrelated to both systems; usually they are less elevated and of smaller surface extent; they occur on all continents.

Mountain formation

Mountains may form by the accumulation of volcanic rocks on the earth's surface or by uplift of part of the earth's surface by tectonic (deformational) forces. Most of the great mountain belts have been formed by the compression and uplift of mobile (unstable) zones of marine sediments after contraction to one-half of their original width or less. During contraction the rocks were deformed, warped, folded, and broken into blocks (faulted). Mountain building therefore comprises two partly related processes: compression and contraction of a mobile zone of the earth's crust is called orogeny, whereas broader uplift that is not accompanied by considerable shortening is called epeirogeny. Both are caused by tectonic or deformational forces originating in the interior of the earth. Many of the major mountain belts are the sites of ongoing tectonic activity, as evidenced by earthquakes and volcanism. Most seismic and volcanic zones of the world are therefore related to mountain belts.

Mountain building is an important process in the evolution of continents. The continents formed from intersecting and overlapping mobile zones that were transformed to mountains during several periods of the geological past. Their deeply eroded roots constitute the foundation, or basement, of the continental platforms. Stratified sedimentary rocks generally form only a thin cover on top of this basement.

The existing mountains formed during the Alpine orogeny, which lasted from the Cretaceous Period (65,000,000 to 136,000,000 years ago) to the Holocene Epoch (the last 10,000 years). Their actual topography is determined by uplift that occurred in most mountains during the Pliocene and Pleistocene epochs—that is, during the last 7,000,000 years. Earlier periods of mountain building occurred during the Paleozoic Era, earlier than 225,000,000 years ago (Hercynian and Caledonian orogenies) and in Precambrian time, earlier than 570,000,000 years ago (Pan-African, Grenville, Hudsonian, and Kenorian orogenies).

Total uplift during a particular orogeny can be as great as 30 kilometres (19 miles) in parts of the major mountain belts; in general, it is between one and five kilometres. Rates of uplift are on the order of a few millimetres per year, but mountain building can be continuous for long periods of time—hence the attainment of overall great heights.

This article treats the several types of mountains that occur in nature and the major mountain systems. For further information on the origin and development of mountains, see the companion article MOUNTAIN-BUILDING PROCESSES, and see also ROCK DEFORMATION. An overview of mountain building during geological time is presented in EARTH, GEOLOGICAL HISTORY OF.

TYPES OF MOUNTAINS

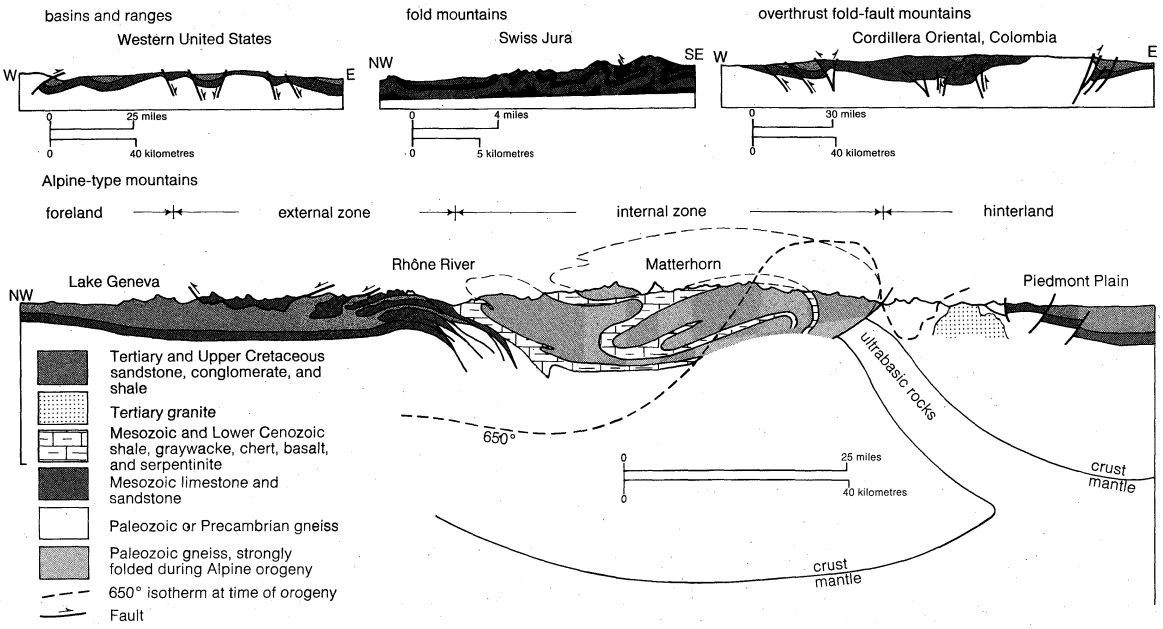
Dome mountains. Dome mountains form by fractureless upwarping or updoming of the land. Their topography is characterized by a relatively flat dissected surface sloping gradually toward the surrounding lowlands. The Black Hills of South Dakota are a typical example.

Block mountains. Block mountains are parts of the crust that have been raised along linear fracture zones in the form of huge blocks. Mountains produced in this way have rectilinear borders and rise abruptly from the adjoining lowlands. The Laurentian Uplands provide a good example.

Effects of block faulting

In many areas block faulting has affected granitic basement (foundation) overlain by a thin cover of stratified sedimentary rocks. Basement and cover have very different mechanical properties and respond differently to tectonic forces; the former will break into blocks that are uplifted, downthrust, or tilted, whereas the weaker sedimentary rocks will become crumpled and folded. This particular type of block mountain is called Germanotype, or Saxonic, after the classical example in central Europe.

Basins and ranges form when irregularly shaped fault blocks are uplifted as the surrounding terrain subsides. The uplifted blocks are eroded by running water and the debris derived from them is transported into the surrounding basins, where it accumulates. Continuing uplift of the ranges leads to upwarping and folding of the sediments of the basin fill at the margin of the ranges. The basins and ranges of Wyoming and Colorado formed in this way.



Representative geologic cross sections of basic mountain types (see text).

Mountains bordering rifts. Rifts are long, linear depressions, 35–350 kilometres (20–200 miles) wide and 300 to several thousand kilometres long, set in a broadly upwarped terrain; they are bounded by faults at both sides. Mountain ranges border the rift valleys and exhibit steep, commonly dissected fault scarps. They slope gradually away from the rifts. The upper Rhine Valley (Rhine Graben) and the Red Sea are rifts, and both are bordered by mountains to the east and west (see further RIFT VALLEYS).

Fold mountains. Fold mountains form by lateral compression and contemporaneous or subsequent uplift in a particular zone. Generally, they are underlain by a relatively thin cover of stratified sedimentary rocks that rests on top of a basement of granite or gneiss. The relatively weak rocks of the cover and the massive rocks of the basement react in different ways during the compression.

Simple and complex folding

Simple-fold mountains form where the rocks of the sedimentary cover have been folded by sliding laterally over the basement, like a carpet lying on a floor. Such mountains exhibit straight, parallel valleys and ridges; the valleys are carved out of soft rocks, whereas hard rocks remain as ridges. The Swiss Jura is a classical fold mountain of this type.

Fold-fault mountains of the overthrust type are more complex and form where compressive forces act upon basement and cover. The rigid basement rocks break into fault wedges along fractures, and the relatively mobile cover is folded into narrow systems of parallel folds. The Cordillera Oriental (Eastern Cordillera) of Colombia is a typical fold-fault belt.

Volcanic mountains. Volcanic mountains are built of overlapping lava flows and layers of tuff (consolidated volcanic dust). Steep volcanic cones (as in Japan) are built by outpouring from a central vent. Their conical shape may be modified by explosion or by collapse of a circular kettle, called caldera (for example, Krakatoa, Indonesia). Broad-shield volcanoes (Mauna Loa, Hawaii) form when less viscous lava flows from a central vent, and volcanic plateaus develop where large volumes of lava erupt from fissures (see further VOLCANOES).

Island arcs. Chains of volcanic islands that are arcuate in plan view are the most conspicuous physiographic element of the western Pacific Ocean. Their convex sides face the open ocean, from which they are separated by a deep oceanic trench. A platform or outer chain of islands rises along an abrupt fault scarp from the trench. The main chain of volcanic cones is located at the concave inner side of the arc. A small ocean basin separates the concave side of the arc from the continent behind.

Island arcs (*q.v.*) are the site of very intense tectonic activity. They constitute the earth's most important earthquake zones, and they are the sites of more than one-half of all active volcanoes. The sources of the seismic activity, or deep-focus earthquakes, are aligned in a plane that dips 30° to 50° from the oceanic trench beneath the arc. Overthrust faulting occurs at the inner wall of the trench, and the crust of the ocean is carried downward along the seismic plane, which is called a subduction zone.

Alpine-type mountains. Alpine-type mountains, such as the Alps or the North American Cordillera, are complex asymmetric belts of folded and faulted rocks. The belts rise abruptly from an adjacent foreland and are subdivided into two zones: an external zone adjoins the foreland, and an internal zone lies away from it.

Zones of Alpine mountains

The external zone is underlain by unmetamorphosed stratified rocks that have been folded or overthrust to form huge nappes (overturned folds). The internal zone is composed mainly of metamorphosed rocks, such as schists and gneisses, that have been deformed; complex overthrusts and great overturned folds are characteristic. These rock structures originated in an intensely deformed zone that is called a root zone. Material has been carried into deeper levels of the earth in the root zone, and it is thought to be analogous to the subduction zone of island arcs.

The metamorphism of rocks, granite intrusions, and volcanic activity that are indicated by Alpine belts also suggest analogies to island arcs: high-pressure metamorphism at low temperatures can occur only at places where cold material is carried into deep crustal levels—that is, in subduction zones.

Alpine-type mountains arose from mobile belts of oceanic or marine sediments called geosynclines. These are subdivided in two zones: a miogeosyncline, adjoining a continental platform, and a eugeosyncline, farther removed from a continent. Well-sorted sandstones and reef limestones, representative of deposits in shallow seas, accumulate in a miogeosyncline, whereas rocks of the eugeosyncline represent more variable conditions. Thick piles of shales, siltstones, and impure sandstones (graywackes) typically are present, and deepwater cherts and thick sequences of volcanic rocks also characterize the eugeosyncline. These volcanic rocks form large submarine-shield volcanoes and, in part, volcanic islands; they represent a rock association similar to the rocks that presently accumulate around island arcs. Generally speaking, the external zone of Alpine belts is composed of miogeosynclinal rocks, and the internal zone is composed of eugeosynclinal rocks. The belts immediately surrounding the Pacific Ocean are an exception because their external zone formed from eugeosynclinal rocks.

Alpine-type belts evolved in two settings: one at the boundary between a continent and an ocean (Circum-Pacific System) and the other between two continental masses (platforms), as in the Tethyan System. In the first case, folding occurs where the oceanic crust is pushed below the former continental margin. In the latter, folding is due to contraction of a former ocean and to "collision" between its bordering continents.

Alpine physiography

The physiography of Alpine belts is essentially determined by the most recent tectonic activity and structures. The main valleys generally follow late orogenic fault zones and warps, and differential erosion (resulting from differential rock resistance) shapes the mountain lands of the external zone. Systems of parallel valleys and ridges form from parallel anticlines and synclines or from parallel nappe zones, as in the Rocky Mountains of Alberta and Montana. Complex nappe structures give rise to complex drainage patterns that reflect the rock structures. The metamorphic rocks of the internal zone show little difference in resistivity to erosion; their physiography bears little relation to fold structure. Drainage patterns are rectilinear, following small faults, warps, or joint systems. Sharp ridges rise from the valleys and are crowned by steep peaks where several ridges meet.

Residual mountains. Residual mountains are remnants of any mountain types that have been greatly worn down by erosion. Landscapes of residual mountains are either produced by differential erosion or are exhumed ancient landscapes.

Differential erosion carves valleys in areas underlain by soft rocks, whereas hard rocks remain as ridges. Landscape forms consequently depend on the tectonic structures. In the southern Appalachians, for example, differential erosion has produced a valley-and-ridge province that mirrors an older fold structure. Rectilinear valleys of the Laurentian Uplands follow a fault-and-joint pattern of Precambrian age.

In other cases, old landscapes have been buried beneath soft sedimentary rocks. Uplift and subsequent removal of the soft cover exhumes the ancient land surface. This happened in the St. Francois Mountains of the Missouri Ozarks, where Precambrian hills were stripped of their sedimentary cover in Tertiary time (7,000,000 to 65,000,000 years ago).

Overview of mountain types

Alpine-type mountains and fold mountains are produced by compression and contemporaneous or subsequent uplift. Dome mountains and block mountains form by vertical movement, eventually accompanied by slight extension or separation of the crust.

Alpine-type mountains and rifts are at the opposite poles of this scale. Alpine-type mountains originated at continental margins during an orogeny, and island arcs are the oceanic counterpart of Alpine belts. Rifts, on the other hand, are zones within the continent that suffered slight extension. Very strong extension of a continent is not possible because the continental crust will split and its fragments will drift apart. This happened in the Red Sea Rift, where strong crustal extension has led to the separation of the African and Arabian platforms. The rift systems, therefore, are the continental counterparts of the midocean ridges, where extension is very strong.

Fold mountains and block-faulted terrains are transitional in type. Both form on continental platforms under the influence of either Alpine-type mountain building or rifting in adjacent areas. Fold mountains develop where segments of the continent have yielded to compressive forces transmitted from adjoining Alpine belts. Block-faulted terrains are either related to Alpine orogeny or to rifting in their vicinity.

WORLD MOUNTAIN SYSTEMS

Circum-Pacific System. The Circum-Pacific Mountain System consists of three segments that exhibit dissimilar geology and physiography. These are (1) the North American Cordillera and the mountain belts of eastern Siberia, (2) the South American Cordillera, and (3) the west Pacific island arcs.

The North American Cordillera is approximately 600–900 kilometres (370 to 560 miles) wide and narrows to 300 kilometres (190 miles) in southern California and southern Mexico. Its present form is dominated by two severely deformed Alpine belts of late Mesozoic and Cenozoic age (the Mesozoic Era ended and the Cenozoic Era began about 65,000,000 years ago) that make up the higher ranges in the east and west; they are separated by a less severely deformed intermontane zone of plateaus, basins, and lower ranges. The eastern belt is composed of several arcuate mountain chains facing the North American Platform (Brooks Range, Mackenzie and Franklin mountains, Rocky Mountains of Alberta and Montana, cordilleras of Mexico), separated by deep recesses. The western belt is curvilinear in plan and comprises the coastal ranges from Alaska to California. This subdivision of the cordillera is poorly defined in the southwestern United States, and the Pacific belt is absent from southern Mexico and Guatemala. Parts of the cordilleran system in the southwestern United States have been destroyed by Tertiary faulting in Nevada and Utah.

North American Cordillera

Intense folding took place in Middle and Late Paleozoic time (about 225,000,000 to 400,000,000 years ago), particularly in the central part of the cordillera. In the Late Paleozoic and Early Mesozoic, the cordillera consisted of several discontinuous ridges, some of which were crowned by volcanoes. Volcanic rocks and sandstones accumulated in the intervening troughs. In the Late Triassic and Early Jurassic (about 190,000,000 years ago), the rocks of the intermontane zone were slightly warped and faulted, whereas intense folding and metamorphism began in the core zones of both Alpine belts. Granites intruded the core zones and the intermontane zone. At the same time, a marginal trough formed east of the cordillera and received the clastic (fragmental sediments) debris produced by the erosion of the rising mountains. Folding was essentially concluded by Middle Tertiary time (about 30,000,000 years ago).

Fault basins broke up the mountain land in Late Cretaceous and Tertiary time and basins were filled with gravels and sands. Volcanic rocks erupted in the core zone of the Pacific belt and in the intermontane zone; they now cover parts of Oregon, Washington, and Mexico. They are surmounted by Pliocene to Recent volcanic cones (formed during the last 7,000,000 years, approximately) in the Coast Ranges of Washington and Oregon and in central Mexico. Intense earthquake activity is related to strike-slip faulting (essentially a lateral movement along a fault) in California and Alaska, where the coastal zone has drifted several hundreds of kilometres northward since Tertiary time. The elevations of the cordilleran system are largely determined by Late Cenozoic faulting, uplift, and upwarping, and subsequent erosion.

Mountain belts of eastern Siberia and Central America

The East Siberian belts are comparable to the North American Cordillera. They also consist of two Alpine belts separated by an intermontane zone. The continental belt comprises the mountains of Verkhoyansk, trending from the Arctic Ocean to the Sea of Okhotsk, whereas the Pacific or more oceanward belt comprises the ranges from Anadyr to Kamchatka. Both are separated by the Kolyma Massif.

The mountains of Central America and the Antillean Island Arc connect the cordilleras of North and South America. Both are underlain by Mesozoic and Cenozoic graywackes, cherts, and voluminous volcanic rocks that were folded during Tertiary time (2,500,000 to 65,000,000 years ago). Cones of active volcanoes are set upon the mountains of Guatemala, El Salvador, Honduras, and Costa Rica and form many of the islands in the eastern Antilles.

The mountain belts of the Andes bound South America on the west. The system consists of one mountain chain, the High Cordillera, in southern and central Chile; it widens in Peru and Bolivia, where it is built of two chains—the Eastern and Western Cordilleras (known as Cordillera Oriental and Occidental in South America)—separated by a high plateau (Altiplano). It narrows again in Ecuador. In Colombia and Venezuela it branches to form the Western, Central, and Eastern Cordilleras. A narrow coastal plain or a chain of lower mountains (coastal cordillera) accompanies the Andes to the west. A narrow

shelf and the Peru–Chile trench form the boundary between the Andes and the Pacific Ocean Basin.

Only the Western Cordillera of Colombia, part of the Eastern Cordillera in Venezuela, and the southernmost High Cordillera of Chile can be described as Alpine-type belts. The remainder of the Andes form fold-fault mountains. The Alpine belts consist of thick shales, graywackes, cherts, and volcanic rocks of Triassic, Jurassic, and Early Cretaceous age (130,000,000 to 225,000,000 years ago). These rocks were strongly folded and were metamorphosed in certain areas during the later Cretaceous and Tertiary. The remainder of the cordillera has a basement of Paleozoic and Precambrian metamorphic rocks. The basement has been covered by very thick andesites (volcanic rocks with more silica than basalts) and interbedded sedimentary rocks in a narrow belt following the Western Cordillera of Ecuador, Peru, and northern Chile during Jurassic and Cretaceous time. Elsewhere, Late Paleozoic and Mesozoic rocks were laid down in continental basins. Only for short periods, and in limited areas, did the sea cover the eastern parts of the Andes during the Mesozoic.

The rocks of the Andes were warped, faulted, and overthrust during the Late Cretaceous and Tertiary. Intense folding is limited to parts of the Western and Eastern cordilleras. At the same time, fault basins were filled with thick conglomerates and sandstones derived from the rising mountain chains. Huge granite massifs intruded the western parts of the Andes during the Mesozoic and Cenozoic.

Large parts of the western Andes were covered by thick sheets of rhyolitic volcanic rocks during the Late Tertiary. Innumerable andesite cones, many of which are still active, crown the Western Cordillera. Destructive earthquake activity is related to underthrusting of the crust of the Pacific Ocean below the Andes and to normal faulting in the mountain belt.

Island arcs of the western Pacific Ocean

Two sets of island arcs frame the western Pacific Ocean. The outer arc system, located at the boundary of the main oceanic basin, comprises the Aleutian, Kuril-Kamchatka, and Bonin-Mariana arcs, the Bismarck Archipelago and Solomon Islands, the Tonga-Kermadec arc, and New Zealand. The inner system consists of the Sakhalin-Japan-Ryukyu arc and the Philippine arc, and it joins the Tethyan Mountain System through the island arc of Indonesia. Both systems meet in triple junctions located in central Kamchatka, central Japan, and western New Guinea.

The smaller islands of the outer arcs are composed of Late Tertiary volcanic rocks. The part of the outer arcs from New Guinea to New Zealand and the inner-island arcs are Alpine-type belts underlain by Paleozoic and Mesozoic graywackes, cherts, shales, and very voluminous volcanic rocks that were folded during several orogenies between Triassic and Recent times. They were intruded by huge granite massifs during the Mesozoic and Cenozoic. The rocks have been metamorphosed in large zones of the arcs.

Andesitic cones of latest Tertiary and Recent age are set upon the inner parts of the island arcs. Many of them are still active. Intense seismic activity is caused by underthrusting of the crust of the ocean below the arcs, by normal faulting in the arcs (particularly in Japan), and by strike-slip faulting (particularly in New Zealand).

The Tethyan System. The Tethyan Mountain System separates the continental platforms of Eurasia from Africa and India. In the west it consists of two continuous Alpine-type belts separated by an intermontane zone. Both belts are composed of arcuate mountain ranges 1,000–2,000 kilometres (600–1,200 miles) long and separated by deep recesses; they face away from the intermontane zone. The northern belt is represented by the Betic Cordillera, the Alps, Carpathians, Balkan Mountains, Caucasus, and the Elburz ranges; the southern belt comprises the Tell Atlas, the Apennines, the mountain lands of Greece and Yugoslavia, the Taurus and Zagros, and the mountains of Oman and Baluchistan. The Hungarian basin, Rhodope Mountains (eastern Greece, southern Bulgaria), northern Turkey, and the plateaus of Iran and Afghanistan belong to the intermontane zone. From Afghanistan to Indonesia only one south-facing chain of Alpine belts is known (the Himalayas, the mountains of western Burma, and the Indonesian island arc); somewhat older ranges, whose relations to the Tethyan System are poorly known, exist in southeast China and Indochina, however. Several belts of fold-fault mountains are set in the foreland of the Tethyan System: the Pyrenees, Cantabrian, Iberian, and Catalonian mountains in Spain, the High Atlas in northwest Africa, and the Jura Mountains of Switzerland and France.

Early Mesozoic limestones resting on a metamorphosed basement of Paleozoic age form the framework of the external zones of the Alpine belts (that is, the northern half of the northern belt and the southern half of the southern belt) and the foreland fold belts. The internal zones of the belts are mainly made of Mesozoic shales and graywackes; an ophiolite association—radiolarian chert, submarine basalt, and serpentinite (composed of magnesium silicate minerals)—is a characteristic component. The intermontane zone consists of Mesozoic limestones, shales, and, in certain areas, volcanic rocks overlying a Paleozoic basement. These rocks were folded and overthrust during late Mesozoic and Cenozoic time.

Formation of the Tethyan System

The Tethyan Mountain System rose from a wide sea named the Tethys. The Tethys System began to form in the Carboniferous and Permian time (around 280,000,000 years ago) and reached its greatest width during the Early Mesozoic (200,000,000 years ago). The limestones of the external zones of the Alpine belts were deposited in shallow water close to the continents farther north and south. The deepwater shales and the ophiolites of the internal zones accumulated in two continuous zones of deep water. The intermontane zone again represents sediments of shallow water. Folding began in the intermontane zone and in the most internal parts of the Alpine belts during the Late Cretaceous (about 70,000,000 years ago).

Cordilleras began to rise in the internal parts of the Alpine belts during the Cretaceous. The clastic debris produced by their erosion accumulated in deep basins along the mountain front. The larger parts of the mountain belts were folded and uplifted in successive steps during the Tertiary Period (2,500,000 to 65,000,000 years ago). Simultaneously, the remaining marine basins became shallower. At the close of orogeny in the Late Tertiary, sands and gravels, derived from the uplifted mountain belts, accumulated in shallow seas and in flood plains bordering the rising mountains to the north and south. The rocks were metamorphosed in the internal zones of the Alpine belts and in parts of the intermontane zone during mountain building. Relatively small granite massifs intruded these zones, and some andesitic volcanic eruptions occurred.

During the Pliocene and Pleistocene (the last 7,000,000 years, approximately) the mountain ranges were uplifted and erosion produced their present landscapes. Earthquake activity is related to the underthrusting of the foreland below the Alpine ranges in the Himalayan-Indonesian segment of the mountain system, whereas it is related to strike-slip faulting (lateral movements) in Iran, Turkey, and Greece. Volcanic activity is strong in Indonesia and moderate in other areas.

Subsidiary mountain ranges. Parts of the platform areas of North and South America, Africa, India, and Australia form broadly upwarped domes. The regional upwarps commonly culminate one or a few hundred kilometres from the coast. They slope gently toward basins in the centre of the large continents (Amazon Basin, Congo Basin, Mississippi Basin). A narrow coastal strip is either sharply downwarped or downfaulted. Impressive mountainous areas with Alpine scenery developed where rivers dissected a greatly upwarped terrain located close to the shore of the ocean (for example, Serra da Matiqueira, Brazil; Cape Mountains, South Africa). The regional domes are parts of Precambrian Shield areas (ancient stable platforms) or parts of Alpine-type belts of Paleozoic age. These regions were eroded during the Early Mesozoic and subsequently were upwarped. The last ma-

jor period of uplift occurred in the Pliocene and Pleistocene (the last 7,000,000 years, approximately).

Basin and range areas of the United States and Asia

The Basin and Range Province of the southwestern United States is characterized by uplifted fault blocks separated by broad, flat basins. The Basin and Range Province cuts across parts of the North American Platform and the Cordillera. In the Platform area (Wyoming and Colorado), the ranges consist of metamorphic rocks of Precambrian age and Paleozoic limestones and quartzites that are undisturbed, except for the faulting and folding that produced the ranges. The ranges of the Cordilleran Province (Utah and Nevada) are underlain by Paleozoic rocks that had been folded and overthrust during Alpine-type mountain building prior to range faulting. Upfaulting of the ranges and subsidence of the basins began in the latest Cretaceous (about 70,000,000 years ago) and continued in successive pulses through much of the Tertiary. The sands and gravels produced by erosion of the ranges were deposited in the basins. After Middle Miocene time (about 15,000,000 years ago) the whole province was uplifted, and faulting continued at some range boundaries.

Basins and ranges and fold-fault belts make up much of central Siberia, Mongolia, Sinkiang, Tibet, and China. Some of the highest mountain ranges of the world are located in this region, among them the Pamir, Kunlun, Tien Shan, and Altai mountains. The ranges are separated by basins, the largest of which are the basins of Sinkiang, the Gobi desert, and the Huang Ho.

In northcentral Siberia (Zubaykalye or Transbaikalia) the higher ranges are underlain by a basement of metamorphic rocks of Precambrian age. In Tibet, Mongolia, Sinkiang, and the adjoining parts of the Soviet Union, the basement consists of Early Paleozoic belts of the Alpine type. The basement was eroded and covered by sedimentary rocks during Late Paleozoic and Early Mesozoic time (around 225,000,000 years ago). Great uplift of the ranges began in the Jurassic, and rocks were intensely folded in some zones. Basins subsided and were filled with sand and gravel derived from the rising ranges. Renewed uplift, gentle bending of sediments, and subsidence of basins continued during the Cretaceous and Tertiary. The main mountain-building period occurred during the later Pliocene and the Pleistocene. Mountain building continues in parts of the region, as indicated by seismic activity, as in the Pamir region, in Tibet, and in Sinkiang.

The block-faulted uplifts of western Europe (Spanish Meseta, Massif Central in France, German *Mittelgebirge*, and the mountains of Bohemia) are underlain by Precambrian and Paleozoic rocks that were folded and partially metamorphosed during Late Paleozoic time (around 200,000,000 years ago). The terrain was eroded during the Permian and then uplifted in successive steps during the Mesozoic and Cenozoic. Pliocene uplift shaped the present ranges.

Rift systems must also be mentioned within the context of subsidiary mountain ranges. The largest rift system in the world consists of the Red Sea Rift and its continuations to the north (Jordan Valley) and the south (East African rifts in Ethiopia, Kenya, Tanzania, and Malawi). Smaller rifts exist in central Europe (Upper Rhine rift) and Central Asia (Lake Baikal rift). Most of the rifts formed during the Early Tertiary (about 60,000,000 years ago). They are still tectonically active and therefore show considerable seismic activity. The Red Sea Rift is continuous with the oceanic-rift system of the Carlsberg Ridge in the Indian Ocean.

BIBLIOGRAPHY. The literature on mountain ranges and belts is both voluminous and diverse. Overviews are presented in the books by L.C. KING, *Morphology of the Earth* (1962); and L.V. DESITTER, *Structural Geology* (1956), among others. Other works, such as that by J. CADISCH, *Geologie der Schweizer Alpen*, 2nd ed. (1953), cover specific mountain belts or parts of mountain systems, but the interested reader should consult articles in a variety of periodicals, notably: *Geologische Rundschau*, the *Bulletin of the Geological Society of America*, the *Journal of Geophysical Research*, and publications of the Geological Survey of Canada, the U.S., the U.S.S.R., and other countries.

(E.Di.)

Mozambique

Mozambique (officially Portugal's Província de Moçambique), also known as Portuguese East Africa, is located on the southeastern coast of Africa between latitudes 10° 27′ and 26°52′ south. Its elongated territory of 303,073 square miles (784,961 square kilometres)—it is the second largest of Portugal's African possessions—has an Indian Ocean coastline of more than 1,700 miles (2,730 kilometres). Mozambique is bounded on the north by Tanzania, on the west by Malawi, Zambia, Rhodesia, and the Transvaal Province of South Africa, and on the south by Swaziland and by the Natal Province of South Africa. To the east lies the Mozambique Channel, which separates the African mainland from the island of Madagascar. (For details on associated physical features see articles on EAST AFRICAN LAKES; EAST AFRICAN MOUNTAINS; MOZAMBIQUE CHANNEL; and ZAMBEZI RIVER. For historical aspects see SOUTHERN AFRICA, HISTORY OF.)

Mozambique's spatial morphology is complex; in the south, it is only 55 miles (88 kilometres) wide; in the middle, about 200 miles (320 kilometres); and in the north, nearly 500 miles (800 kilometres). A wedge of its territory also extends westward along the Zambezi River, between Zambia and Rhodesia. Malawi, in turn, penetrates deeply into northern Mozambique.

Mozambique's pronounced territorial elongation has political, economic, and social significance. Its northern sectors lie adjacent to militantly nationalist black African states, while the south borders white-controlled South Africa.

With over 8,200,000 inhabitants, Mozambique is Portugal's most populous overseas domain. Although officially an overseas province of Portugal, Mozambique remains in effect a colonial dependency of Lisbon. The constitutional amendment of 1952, whereby the territory's colonial status was revised into a provincial one, served to shield it for a time from the scrutiny of United Nations committees concerned with colonial dependencies until, in 1960, the United Nations General Assembly proclaimed it to be a non-self-governing territory. Mozambique has not, however, been exempted from the political and military pressures of African nationalism. Throughout the late 1960s and into the early 1970s, the territory has faced attacks by Tanzania-based nationalist forces that, aided by local allies, have wrested a large part of the north from Portuguese control. While Portugal has maintained substantial armed forces in the field to combat the onslaught in the north, efforts have intensified to accelerate the pace of development elsewhere, to reduce the susceptibility of the territory to nationalist influence, and to link the country more closely to Portugal.

THE LANDSCAPE

Mozambique occupies southern Africa's largest coastal plain; nearly half of the total area lies below 750 feet (229 metres). This coastal lowland is widest in the south, where it covers virtually the entire width of the territory, but northward it becomes narrower. In central and northern Mozambique an interior plateau develops whose elevations range from 500 to 2,000 feet (152 to 610 metres). Toward the western boundaries with Malawi, Zambia, and Rhodesia, this plateau rises still higher, and mountainous terrain occurs. Highest elevations along the Zambia–Malawi border range from 5,741 feet (1,750 metres) in the extreme north, near Lake Malawi in Niassa District, to 7,241 feet (2,207 metres) in the northwest, near the intersection of the Zambia–Malawi boundary in Tete District.

The lengthy coastline is marked by several distinct features. In the extreme south lies the Bay of Lourenço Marques, also known as Delagoa Bay, protected from the Indian Ocean by a sandy spit. Delagoa Bay is the best natural harbour on the East African coast, and on its shore the capital of the territory, Lourenço Marques, has grown. Northward, from Delagoa Bay to the delta of the Zambezi River, the coast is sandy and swampy, with mangroves and offshore bars. North of the Zambezi Delta, however, sandy stretches of coast are interrupted by rocky cliffs and headlands as well as islands, many of

which are of coral formation. In addition to the Zambezi, some 50 other rivers reach the coast of Mozambique, including the Limpopo, the Save, the Búzi, the Lúrio, and the Messalo. In the far north the Rovuma River forms the boundary between Mozambique and Tanzania.

The Zambezi: a physiographic frontier

The pivotal geographical feature of Mozambique, in the interior as well as on the coast, is the Zambezi River (*q.v.*). Almost all the higher and more rugged terrain lies to the north of the river, while to the south the gently undulating coastal lowland prevails. Apart from the higher elevations near the Rhodesian Plateau, there is only one marked feature of positive relief to be located south of the Zambezi: the Serra de Gorongoza in Sofala District. This steep-sided granite massif, an inselberg (prominent residual mountain), reaches 6,112 feet (1,863 metres), but it is only some 12 miles (19 kilometres) wide. Elsewhere, southern Mozambique presents a gently rolling, rather featureless landscape, with an occasional group of low hills, broken by the wide alluvial valleys of the rivers. North of the Zambezi River, the undulating plateau has a landscape of the savanna type, with latosolic soils (red residual soils, leached of silica, and containing iron and aluminum hydroxides). Toward the interior, some spectacular mountains rise above the plateau surface. Lake Malawi (Nyasa) lies in the southern part of the East African Rift Valley, whose fault structures are thought to extend at least to the mouth of the Zambezi.

Mozambique's topography generally declines in elevation from west to east, so that, from the Rovuma River in the north to the Limpopo in the south, streams have an eastward orientation. The low gradients on the coastal plain, especially in the south, nevertheless render some of the stream courses irregular. The higher parts of the territory are sustained by the crystalline rocks of the ancient African basement complex or shield. In the Zambezi Basin and on the coastal plain, dipping toward the Mozambique Channel, lie more recent sedimentary rocks deposited during the Triassic Period (about 260,000,000 to 200,000,000 years ago). Still more recent sediments, of Late Cretaceous (about 100,000,000 to 70,000,000 years old) and Eocene (about 60,000,000 to 40,000,000 years old) age, lie in the eastern parts of the coastal plain. A notable feature is the rhyolite-sustained Lebombo Range, which coincides with the political boundary between Mozambique and Swaziland and also extends northward along the Transvaal border. The Lebombo may be the eastern flank of the Swaziland Rift Valley, an echelon continuation of the East African Rift Valley.

With the exception of northern Mozambique east of Lake Malawi, Mozambique's major rivers rise to the west of the territory, on the African high veld. The headwaters of the Limpopo lie on the margins of the Kalahari Desert: the Save River rises on the Rhodesian Great Dyke, and the Zambezi rises in the Angola–Zambia borderland. Since the interior plateaus receive moderate rainfall at best, with marked seasonal variation, water levels alter considerably. The largest drainage basin is that of the Zambezi River, which, with its Shire tributary (the outlet of Lake Malawi), drains most of the centrally situated districts of Tete, Manica e Sofala, and Zambézia.

Good soils are the exception rather than the rule in Mozambique. The soils of the plateau areas, derived from underlying granites, gneisses, and schists, are ferralitic (leached and iron–aluminum rich) and are generally rather infertile. In the dry interior to the north of the Limpopo Valley, sandy, ferruginous (iron-bearing) soils prevail, whose high permeability increases the area's dryness. Along the coastal plain there are waterlogged and organic soils, with some poorly developed alluvials. In the alluvial lowlands of several of the river basins, and in the *machongas* (humus-rich, moist patches of lowland between the coastal plain's dunes) lie Mozambique's most productive soils.

CLIMATE

As Mozambique lies on the tropical East African coast, apparently in the path of the southeast trade winds, and as the warm Mozambique Current flows southward offshore, one would expect a rainy, moist climate to prevail. This would seem the more likely as the mountains of the interior cause air currents to rise, thus increasing the probability of rain. Much of Mozambique is, however, subhumid, and drought is frequent in many areas. Lourenço Marques, for example, receives a mere 30 inches (7,595 millimetres) of rain in an average year; Inhambane, 36; Beira, 10; António Enes, 42; and Tete, 23. With the prevailing tropical temperatures (Lourenço Marques has an annual average of 73° F [23° C], while Quelimane's warmest month averages 82.5° F [28° C] and its coolest month 69° F [21° C]), much of this rainfall is lost by evaporation or by plant transpiration. Only north central and northwestern Mozambique generally maintain a favourable moisture balance.

Rainfall

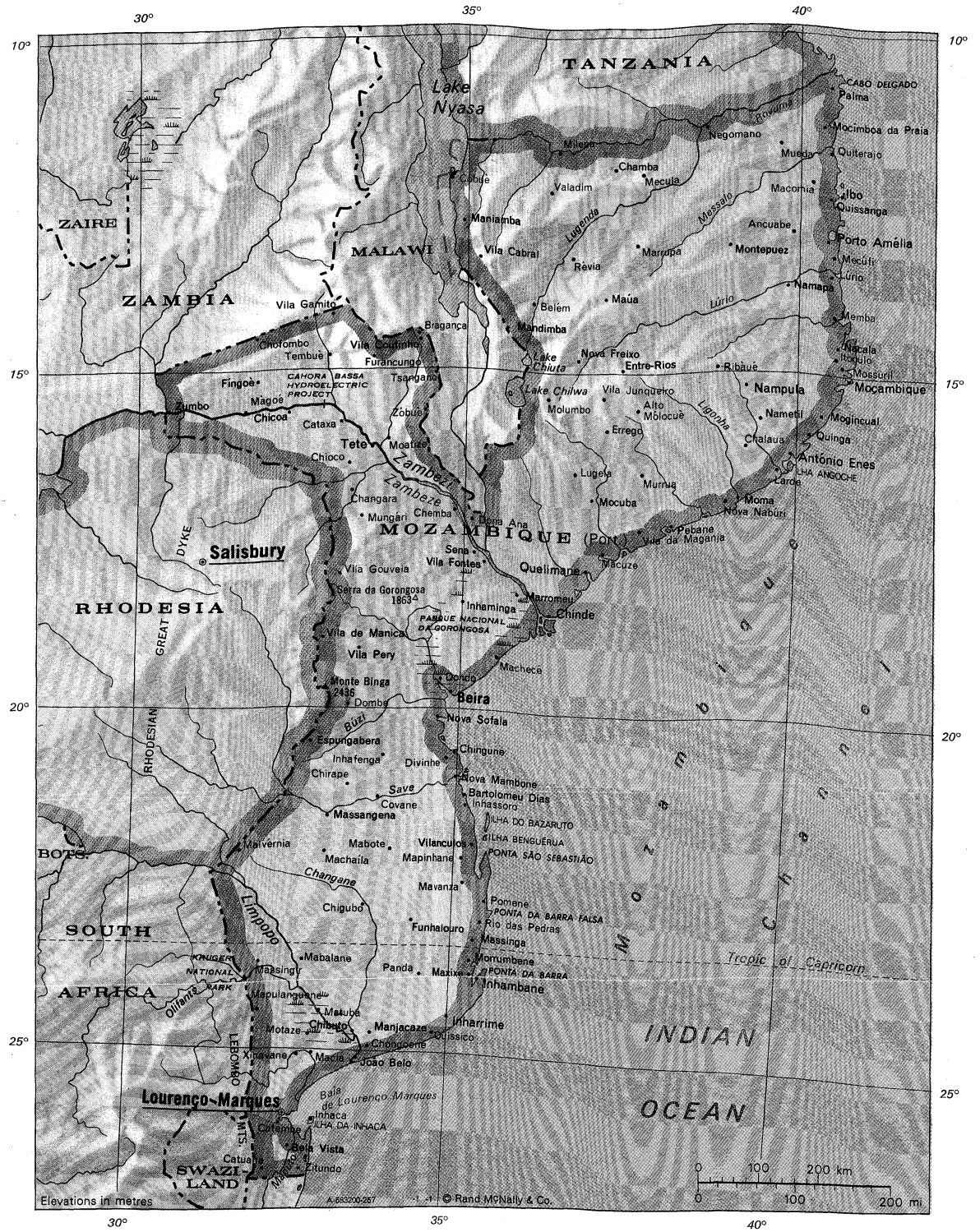
Several reasons for Mozambique's comparatively low rainfall have been suggested. The configuration of the coastline may be such that the moisture-bringing trades blow parallel to most of it, rather than approaching it at an angle. Madagascar, across the Mozambique Channel, may interfere with the westward-moving trades and reduce their moisture content. The air layer that reaches the coast may be thin and therefore capable of generating little rain. Mozambique may also be under the influence of a permanent South African anticyclone, or high pressure area, repelling air masses moving in from the Indian Ocean.

Mozambique is tropical savanna country extending south beyond the Tropic of Capricorn. It is, nevertheless, its altitude rather than its latitude that moderates temperatures. Although Lourenço Marques lies nearly 1,000 miles south of Porto Amélia, its annual average temperature is only slightly lower, and its annual range only slightly greater. Chillier nights and cooler winter months are recorded in the higher zones along the Rhodesian and Zambian borders. There is a distinct rainy season, which is also the high-sun period, from October (or November) until March. The dry season usually lasts from April into September and sometimes longer. Annual precipitation figures show that Mozambique north of the Zambezi is a region of greater moisture and lesser drought incidence (although the Zambezi Basin around Tete is itself moisture deficient) but that the south suffers from drought and low rainfall, except along the littoral and against the Rhodesian Plateau slopes.

VEGETATION AND ANIMAL LIFE

Mozambique's vegetation is mostly tropical forest and savanna, with many barkcloth trees. Along the coast, especially between 17° and 20° south, the coconut palm is common. The coconut plantation at Quelimane, with 50,000 acres (202 square kilometres) and 4,000,000 palms, is often identified as the largest in the world. Several other types of palm, including the date palm, also occur. Along the upper reaches of the major rivers, notably the Zambezi and the Limpopo, the mopane tree prevails on the savanna, where the baobab tree is also a characteristic feature. Areas of mangrove exist along the coast, especially in the vicinity of the Zambezi Delta. Bamboo and spear grass occur abundantly along the river banks and in the marshy areas. Patches of forest, which include stands of ironwood and ebony, are generally found on better drained slopes amid savanna country. In the narrow ravines tree growth is dense, including the large Senegal khaya, used for timber.

Despite excessive hunting and the virtual extermination of game in some areas by professional hunters and poachers, Mozambique still remains rich in animal life. Lions are still numerous, and the leopard is often seen on the granite hills of the plateau slopes and in the denser bush. The cheetah, spotted hyena, jackal, serval (a long-legged wildcat), civet cat, mongoose, and wild dog occur in several areas, notably in the Moçambique and Manica e Sofala districts. Some animals also roam into the country from South Africa's Kruger National Park, which lies to the southwest. The elephant is common, as is the black rhinoceros; the white rhinoceros occurs, although more rarely. Many of the lower courses of Mozambique's rivers are inhabited by the hippopotamus. Among the herd



MOZAMBIQUE

MAP INDEX

Cities and towns

Alto Molócuè...15·38s 37·42e
Ancuabe.....12·58s 39·54e
António Enes...16·14s 39·58e
Bartolomeu Dias.....21·10s 35·09e
Beira.....19·49s 34·52e
Bela Vista.....26·20s 32·40e
Belém.....14·13s 35·58e
Bragança.....14·23s 34·20e
Cataxa.....15·58s 33·12e
Catembe.....26·00s 32·33e
Catuane.....26·48s 32·18e
Chalaua.....16·06s 39·11e
Chamba.....12·07s 36·57e
Changara.....16·54s 33·14e
Chemba.....17·08s 34·52e
Chibuto.....24·44s 33·33e
Chicoa.....15·37s 32·24e
Chigubo.....22·50s 33·34e
Chinde.....18·37s 36·24e
Chingune.....20·38s 34·55e
Chioco.....16·25s 32·50e
Chirape.....21·18s 33·33e
Chofombo.....14·35s 31·50e
Chongoene.....25·00s 33·47e
Cóbuè.....12·04s 34·50e
Covane.....21·22s 33·56e
Divinhe.....20·40s 34·49e
Dombe.....19·59s 33·25e
Dona Ana.....17·25s 35·07e
Dondo.....19·36s 34·44e
Entre Rios.....14·57s 37·20e
Errego.....16·02s 37·14e
Espungabera...20·29s 32·48e
Fíngoè.....15·12s 31·50e
Funhalouro.....23·03s 34·25e
Furancungo....14·55s 33·35e
Ibo.....12·20s 40·35e
Inhaca.....26·01s 32·58e
Inhafenga.....20·35s 33·53e
Inhambane.....23·51s 35·29e
Inhaminga.....18·24s 35·00e
Inharrime.....24·29s 35·01e
Inhassoro.....21·33s 35·11e
Itoculo.....14·42s 40·18e
João Belo.....25·04s 33·39e
Larde.....16·28s 39·43e
Lourenço Marques.....25·58s 32·35e
Lugela.....16·25s 36·43e
Lúrio.....13·35s 40·30e
Mabalane.....23·37s 32·31e
Mabote.....22·03s 34·09e
Machaila.....22·15s 32·55e
Machecé.....19·17s 35·33e
Macia.....25·03s 33·10e
Macomia.....12·15s 40·08e
Macuze.....17·42s 37·11e
Mágoé.....15·48s 31·43e
Malvérnia.....22·06s 31·42e
Mandimba.....14·21s 35·39e
Maniamba.....12·43s 35·00e
Manjacaze.....24·44s 33·53e
Mapinhane.....22·19s 35·03e
Mapulanguene .24·29s 32·06e
Marromeu.....18·20s 35·56e
Marrupa.....13·03s 37·30e
Massangena....21·32s 32·57e
Massinga.....23·20s 35·25e
Massingir.....23·51s 32·04e
Matuba.....24·27s 32·55e
Maúa.....13·51s 37·10e
Mavanza.....22·43s 35·08e
Maxixe.....23·51s 35·21e
Mecúfi.....13·17s 40·30e
Mecula.....12·04s 37·40e
Memba.....14·11s 40·30e

MAP INDEX (continued)

Milepa.....11·43s 36·20e
Moatize.....16·08s 33·45e
Moçambique....15·06s 40·50e
Mocímboa da Praia.....11·20s 40·21e
Mocuba.....16·50s 36·59e
Mogincual.....15·35s 40·25e
Molumbo.....15·27s 30·15e
Moma.....16·44s 39·14e
Montepuez.....12·32s 40·27e
Morrumbene ...23·39s 35·20e
Mossuril.....14·58s 40·42e
Motaze.....24·48s 32·52e
Mueda.....11·39s 39·33e
Mungári.....17·12s 33·31e
Murrua.....16·20s 37·56e
Nacala.....14·32s 40·37e
Namapa.....13·43s 39·50e
Nametil.....15·43s 39·21e
Nampula.....15·07s 39·15e
Negomano.....11·27s 38·31e
Nova Freixo.....14·49s 36·33e
Nova Mambone.20·59s 35·01e
Nova Nabúri....16·46s 38·57e
Nova Sofala....20·09s 34·42e
Pafúri.....22·27s 31·21e
Palma.....10·46s 40·29e
Panda.....24·02s 34·45e
Pebane.....17·10s 38·08e
Pomene.....22·53s 35·33e
Porto Amélia....12·58s 40·30e
Quelimane.....17·53s 36·51e
Quinga.....15·49s 40·15e
Quissanga.....12·25s 40·29e
Quissico.....24·42s 34·44e
Quiterajo.....11·48s 40·25e
Révia.....13·23s 36·31e
Ribauè.....14·57s 38·17e
Rio das Pedras..23·12s 35·23e
Sena.....17·27s 35·00e
Tembuè.....14·52s 32·58e
Tete.....16·13s 33·35e
Tsangano.....15·08s 34·32e
Valadim.....12·22s 36·10e
Vila Cabral.....13·18s 35·14e
Vila Coutinho...14·37s 34·19e
Vila da Maganja.....17.18s 37.30e
Vila de Manica..18·56s 32·53e
Vila Fontes.....17·50s 35·21e
Vila Gamito.....14·12s 33·00e
Vila Gouveia....18·03s 33·11e
Vila Junqueiro..15·25s 36·58e
Vilanculos.....22·01s 35·19e
Vila Pery.....19·08s 33·29e
Xinavane.....25·02s 32·47e
Zitundo.....26·45s 32·50e
Zóbuè.....15·38s 34·26e
Zumbo.....15·36s 30·25e

Physical features and points of interest

Barra Falsa, Ponta da, *point*.22·55s 35·37e
Barra, Ponta da, *point*.....23·45s 35·30e
Bazaruto, Ilha do, *island*.....21·40s 35·28e
Benguérua, Ilha, *island*.....21·58s 35·28e
Binga, Monte, *mountain*.....19·47s 33·04e
Búzi, *river*.....19·50s 34·43e
Cahora Bassa Hydroelectric Project.....15·34s 32·38e
Changane, *river*.24·43s 33·32e
Chilwa, Lake....15·12s 35·50e
Chiúta, Lake....14·55s 35·50e
Delgado, Cabo, *cape*.....10·40s 40·35e
Gorongosa, Parque Nacional da, *national park*...18·48s 34·15e
Gorongosa, Serra da, *inselberg*.....18·24s 34·06e
Indian Ocean...25·00s 40·00e
Inhaca, Ilha da, *island*.....26·03s 32·57e
Lebombo Mountains.....26·15s 32·00e
Ligonha, *river*...16·54s 39·09e
Limpopo, river..25·15s 33·30e
Lourenço Marques, Baie de, *bay*.....25·48s 32·51e
Lugenda, *river*..11·25s 38·33e
Lúrio, *river*.....13·35s 40·32e
Maputo, *river*...26·11s 32·42e
Messalo, *river*...11·40s 40·26e
Mozambique Channel.....20·09s 40·00e
Nyasa, Lake.....12·00s 34·30e
Rovuma, *river*...10·29s 40·28e
São Sebastião, Ponta, *point*....22·07s 35·30e
Save, *river*.....21·00s 35·02e
Zambezi (Zambeze), *river*.....18·55s 36·04e

animals, the buffalo is especially numerous, and herds of zebra are often seen. Giraffes and wart hogs are also to be found. There are many antelopes. The scaly anteater is ubiquitous, and there are many species of rodents. Several kinds of monkeys and lemuroids exist, although the anthropoids are absent.

Snakes are common, especially the python, cobra, puff adder, and viper. Crocodiles abound in the rivers, and there are many lizards, chameleons, and several species of tortoises. In addition to scorpions and centipedes, insects are prolific. There are swarms of mosquitoes, locusts, tsetse flies, hippo flies, cockroaches, phylloxera (plant lice), termites, soldier ants, and flying ants.

Mozambique's bird life is varied. Flamingoes inhabit the northeast, especially Moçambique District, and species of cranes, storks, herons, pelicans, and ibises occur. The eagle, buzzard, and crow are omnipresent, as is the vulture. Guinea fowl, partridge, bustard, quail, wild goose, teal, and other kinds of duck are all found in large numbers.

TRADITIONAL REGIONS

As already mentioned, the Zambezi River forms a significant dividing line between northern and southern Mozambique. South of the river are the territory's major cities, Lourenço Marques and Beira. Southern Mozambique is adjacent to the Transvaal Province of South Africa and to Rhodesia, and its ports have long functioned as ports of entry and exit for South African and Rhodesian goods. African labourers by the tens of thousands have gone to work in the mines and on the fields of these neighbouring countries, and the Portuguese administration has benefitted from the resulting income derived both from labour agreements and from individual taxes. In southern Mozambique, also, are the chief sites of the agricultural settlements by means of which the administration seeks to accelerate development and to consolidate control through the encouragement of European immigration. The transport system is also more developed in the south.

Development in the south

Southern Mozambique is, moreover, relatively remote from the African nationalist threat in the north. Whereas southern Mozambique contains the majority of the approximately 100,000 Europeans who live in the territory, northern Mozambique contains some of its most densely peopled African districts. The northern districts of Zambézia and Moçambique together are inhabited by nearly 3,500,000 people—over 40 percent of the total population of the entire territory. With its larger black population, its remoteness from the capital, Lourenço Marques, its rudimentary transport system, and its lengthy boundary with Tanzania, northern Mozambique is vulnerable to attack and has consequently become the main theatre of the struggle Portugal is waging against black nationalism in Africa. While this fact is of immediate import, there are also other factors that distinguish the north from the south—the lack of economic opportunity inland, where Malawi and Zambia generate far less employment or trade than South Africa and Rhodesia, the scarcity of alternatives to subsistence agriculture, the imprint of Islām on some of the African peoples, and the generally stronger resistance to change.

HUMAN SETTLEMENT

The vast majority of Mozambique's African peoples live in dispersed homesteads, small hamlets, or villages. Although various types of dwellings are built, most African houses are cylindrical, with a conical thatched roof and with walls of wattle and daub. In the hamlets and villages, a cluster of these houses surrounds either a cattle corral or, occasionally, a central meeting ground dominated by a large tree. Surrounding each settlement are the fields worked communally by its inhabitants. Maize (corn) is the staple for most of the territory, although yields per acre are low and the annual production highly variable. Famine remains a local threat.

Much of rural Mozambique is thinly populated bush country, but pressure on the land occurs in the coastal areas of Moçambique and Zambézia districts and in southern Inhambane and Gaza as well as in the Lourenço Marques District. Because of a legacy of past slave raiding, endemic sleeping sickness, and low soil fertility, Niassa District in the north remains the emptiest area of the province. In the more fertile areas, Africans have combined some cash cropping, especially of cotton and rice, with their cultivation of staples.

Considerable population agglomerations have occurred in the hinterlands of the coastal towns, both as a result of greater soil fertility and as a result of the drawing power of such towns or cities as Lourenço Marques, João Belo, Inhambane, Beira, Quelimane, António Enes, and Moçambique. In these areas the rural homestead or hamlet pattern begins to vanish as truck farming appears and as the modernizing influence of the town alters the nature of settlements. Vegetable plots, some tobacco fields (mainly in the north), fields of rice and maize, and some sugar and citrus plantations occur. The European plantations and the fields devoted to cash cropping are almost everywhere interspersed with areas of African settlement or of bush country where little has changed.

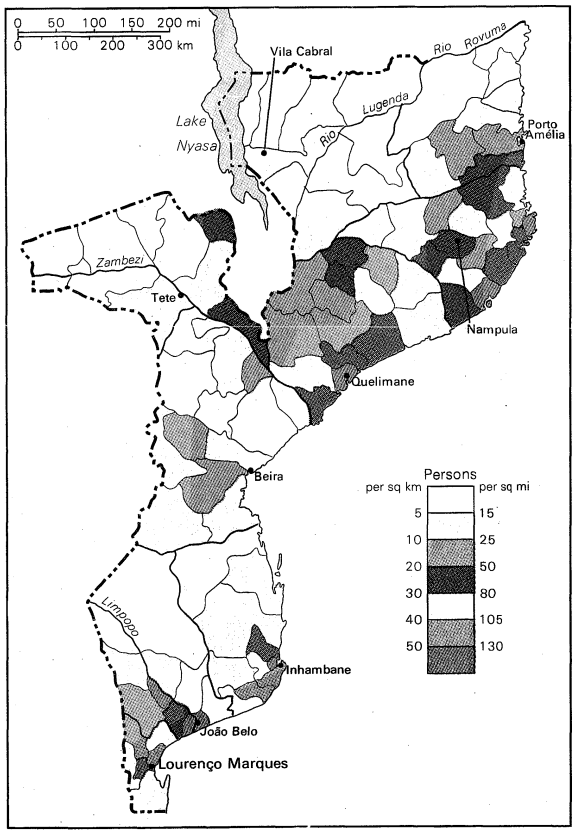
Lourenço Marques and Beira, both with metropolitan area populations exceeding 150,000, are modern, clean, and attractive cities. They are well planned, with wide avenues and ample shade trees. A meticulous attention to detail, as demonstrated by some inlaid sidewalks and by well-kept and terraced public gardens, suggests an administrative intention to provide a cultural focus for the territory. Large apartment houses and elaborate public buildings dominate the urban landscape. In the suburban areas there are villas and smaller homes in the characteristic Portuguese style. Both Lourenço Marques and Beira are somewhat compact; around the periphery of these cities, however, unplanned African shantytowns have grown up.

PEOPLE AND POPULATION

The African majority

The overwhelming majority (97 percent) of Mozambique's population of over 8,200,000 is black African. Most of the Africans belong to the Central Bantu group. Two major historical events that shaped the peopling of Mozambique—the great Bantu migration from the north and west and the northward migratory movement that followed the rise of the Zulu Empire in Natal to the south—created two subdivisions of the Central Bantu group in the territory. Again the Zambezi is a crucial dividing line: north of the river the people trace their lines of descent on the matrilinear principle, while patrilinear peoples prevail in the south.

The two main groupings thus identified can be further subdivided on an ethnolinguistic basis. The African peoples south of the Zambezi are related to those in Rhodesia. Approximately from the latitude of Beira southward are the Thonga, while to the north live the Shona (or Mashona). The Thonga and Shona are primarily farmers, although such groups as the Shangana (a subgroup of the Thonga) own sizable cattle herds in Gaza District. In the Zambezi Valley and northward two clusters are also recognizable. In the east are the Makua (Macua) and Makonde peoples, who both form part of the Yao tribal grouping, and who occupy most of the land between Malawi and the coast. The Makonde, who have a history of opposition to Portuguese rule and are currently involved in the war against the Portuguese in the north, are also settled in Tanzania, on the north bank of the Rovuma River. To the west live the Maravi as well as other Yao peoples, all of whom have tribal affiliations across the frontier. These Yao live in the northern corner adjacent to Lake Nyasa, as well as in Tanzania, while the Maravi occupy the Zambezi corridor north of the river, adjacent to Zambia and Malawi.



Population density of Mozambique

The former status of *assimilados* (*i.e.*, assimilated, or living as a European) has been replaced by universal full citizenship, but the non-African minority still constitutes the territory's elite. An extrapolation from the 1960 census suggests that there are now more than 150,000 Europeans in Mozambique, not counting the 60,000 armed forces that were in the territory in 1969. The European minority that controls most of the country, and which is almost entirely Portuguese, is concentrated in the urban centres and in the most productive agricultural zones. There are also an estimated 40,000 mestizos (persons of mixed ancestry). The largest cities contain Chinese, Goan, and Pakistani quarters, and these minorities play a leading role in trade and commerce. The Chinese community in Beira has long maintained its own primary school; in Lourenço Marques many shopkeepers, especially in the "Old Town" of the city, are Indian or Chinese. In the mid-1960s, membership in the territory's various economic associations numbered 4,000, out of which about 1,000 were non-Europeans, most of whom were Indians and Chinese. There are also some South Africans and Rhodesians, most of whom are officials or engaged in trade.

Mozambique's language pattern is as complex as its population is varied. African peoples speak several distinct Bantu languages, most of which are mutually unintelligible. The closest any African language comes to serving as a lingua franca is Swahili, which is spoken in the towns and villages along the coast, especially north of the Zambezi. Portuguese is the official language, but the literacy rate is estimated at only 6 percent. One result of the labour migration to the South African and Rhodesian mines has been to foster English among the African workers. A substantial number of the Portuguese population also speak English.

A similar variation occurs in religious affiliations. In the north, Islāmic traditions are superimposed on essentially animist African societies. In the south, Christianity has made deeper inroads, primarily in the form of Roman Catholicism. The position of the church has been defined by the Portuguese concordat and missionary agreement of 1940, which created a missionary see at Lourenço Marques, to which two dioceses at Beira and Nampula are suffragan. There is also some Protestant missionary activity. The religious diversity of Mozambique finds expression in the architecture of Lourenço Marques, which includes a Catholic cathedral and several Catholic churches, an Anglican cathedral, a Methodist church, a synagogue, a mosque, and a Chinese temple.

DEMOGRAPHY

The 1960 census reported a population of 6,600,000, of whom 6,460,000 were identified as "tribalized Africans," 148,000 as "civilized" (including the European and other non-African minorities), and between 4,000 and 5,000 as "westernized Africans." Reliable birth and mortality rates have never been available, and growth-rate estimates average 2.0 percent (births, 43 per 1,000; deaths, 23 per 1,000). Officially, infant mortality is reported as being between 90 and 100 per 1,000, but some observers state that there are areas where it may be near 50 percent. Emigration and immigration data are similarly unavailable. Further complexities have arisen as a result of the political conflict. More than 10,000 Makonde have reportedly fled to Tanzania in the past few years, and many Lomwe—a subgroup of the Yao people, mentioned earlier—have migrated into Malawi. Approximately 40 percent of all Thonga males at one time or another work in the mines of South Africa or Rhodesia, and an unknown percentage of them remain there. Some Shona, on the other hand, have crossed the border from either South Africa or Rhodesia to settle in southern Mozambique.

White immigration

White immigration has grown steadily over the past half century, but in the 1950s and 1960s the rate increased greatly. In 1930 there were only some 17,000 whites in Mozambique; in 1940, about 27,000; and in 1950, 48,000. In the 1950s the European population nearly doubled, and the 1970 estimate is about 150,000—excluding the armed forces. White immigration has been stimulated by increasing economic development, and by governmental policy, which has encouraged Portuguese to emigrate to Mozambique or Angola. While the majority of Portuguese emigrants settle in Angola, Mozambique has nevertheless received a share (25,344 in 1967). In the valleys of some of the territory's rivers, such as the Limpopo

and the Revuè, settlement schemes have been laid out, and farming families from Portugal are being encouraged to settle by offers of land and loans. In the basin of the Zambezi River, the Cabora Bassa Hydroelectric Project was under construction in the early 1970s. Although primarily aimed at the generation of cheap power, the Cabora Bassa Project will also serve to expand the irrigated acreage of land in the Zambezi Basin and will thus stimulate farm settlement on a large scale. Flood control and irrigation over a 3,000,000-acre (12,141-square-kilometre) area are foreseen.

Mozambique, Area and Population

	area*		population	
	sq mi	sq km	1960 census	1970 census
Districts (*distritos*)				
Cabo Delgado	30,260	78,374	549,000	567,000
Gaza	31,866	82,534	682,000	753,000
Inhambane	26,436	68,470	583,000	747,000
Lourenço Marques	6,480	16,783	437,000	799,000
Manica e Sofala	50,137	129,854	780,000	1,085,000
Moçambique	30,218	78,265	1,452,000	1,735,000
Niassa	46,384	120,135	281,000	297,000
Tete	38,886	100,714	471,000	492,000
Zambézia	39,722	102,880	1,369,000	1,757,000
Total Mozambique	300,389	778,009	6,604,000	8,234,000†

*Areas for the districts and the total are land areas. The total area for Mozambique is 303,073 sq mi (784,961 sq km). † Figures do not add to total given because of rounding.
Source: Official government figures.

Population distribution is characterized by (1) a concentration along the coast, except along the Zambezi Valley; (2) strong rural agglomeration, especially in Moçambique and Zambézia districts and in the south; and (3) very low densities in the areas of the upper Limpopo-Save, upper Zambezi, and upper Rovuma-Lugenda rivers. African and non-African population distributions should, however, be viewed separately. More than 90 percent of the Africans live in the rural areas, in the traditional manner, but between 75 and 80 percent of the Europeans live in the cities and towns. The Indian and Chinese minorities are almost completely urbanized.

Reliable demographic statistics being scarce, it is difficult to discern trends. The rate of European immigration may nevertheless be expected to continue to increase. Large areas in the north being out of Portuguese control, no accurate population count can be expected in this region for some time. Mozambique is known to have inadequate medical facilities (except in the towns), and the increase of less than 800,000, or 12.5 per thousand, in the African population between 1950 and 1960 may be attributable to a moderately high birth rate and high death rates. Some authorities feel that the birth rate remains moderate because of the absence of many male migrant labourers in South Africa and Rhodesia and because of a high infant mortality rate.

THE NATIONAL ECONOMY

Mozambique is not a rich territory. The great majority of the people are engaged in subsistence agriculture. No known mineral deposits are comparable to those of South Africa, Rhodesia, Zambia, or Zaire. Mozambique's assets are its soils, labour, and location. Its income comes chiefly from agriculture and transport.

The balance of trade normally shows a large deficit, which is partially made up by income from transportation. In 1968, for example, the value of exports totalled $100,000,000, and of imports $151,000,000. A large part of the $51,000,000 deficit was made up by earnings deriving from the transit trade link connecting South Africa and Rhodesia with Lourenço Marques and Beira. The remaining deficit, nevertheless, still totalled $11,200,000.

Portugal and South Africa are Mozambique's chief trading partners. Portugal buys about 40 percent of the territory's exports and supplies more than one-third of its imports, and South Africa's share in both directions averages 12 percent annually. India and the United Kingdom are also major partners; India buys much of Mozambique's cashew harvest. Cotton, sugar, and other agricultural products constitute the main volume of exports, while various kinds of machinery and textiles account for most of the import value.

RESOURCES

Mozambique's mineral resources may not yet be fully known, for recent discoveries reported include gold in Manica e Sofala, iron ore in Moçambique District and near Tete, and manganese near the South African border. Other hypothetical reserves include oil along the margin of the coastal plain. Hitherto, however, mining in Mozambique, despite an interesting history, has produced few results. As recently as the late 1960s the only substantial mining operation was the working of the coal deposit near Tete in the Zambezi Valley, used by the Mozambique and Rhodesian railroads. With the advent of the Cabora Bassa Project and with the recent discovery of iron ore, the possibility of an economic transformation of the area is mooted. Other minerals that have been mined include tantalite (Mozambique reportedly possesses the largest deposit of this rare mineral in the world) at Murrua in the north, gold in Bragança, iron in Namapa (where a Japanese firm is investing in its exploitation), and manganese and asbestos in Catuane in the southwest corner of the territory. In addition, beryl, graphite, ilmenite, mica, and bismuth have been extracted—all in small quantities—as well as natural gas near Maamba, northwest of Lourenço Marques. Despite reports of new discoveries (that of uranium was announced in 1970), the contribution of the territory's mineral resources to the economy nevertheless remains small.

Mineral prospects

Mozambique's biological resources have been put to various uses. Natural products, such as copra and cashew nuts, are collected by the Africans, and the cashew-nut industry has assumed large proportions following the construction of a processing plant at Vila de João Belo. Mozambique exports a small amount of timber, mainly from logging areas in Manica e Sofala District. More than half the exports are in sawn timber, used for railroad ties for South Africa. The varied fauna of the territory draws many hunters and visitors to Parque Nacional da Gorongosa. Thousands of South African tourists visit the coasts of Mozambique each year.

The electric power resources of Mozambique will be much increased by the completion of the Cabora Bassa Project. In the late 1960s the only hydroelectric project operating in Mozambique was that on the Revuè River, which generates power for an area that includes Beira, as well as Umtali in Rhodesia. The installation was recently enlarged, and an additional project was begun on the Maputo River in the extreme south of the territory. Meanwhile, Mozambique is also served by some 688 thermal power plants, most of which are small in scale; those at Lourenço Marques have a 53,000 kilowatt capacity.

SOURCES OF NATIONAL INCOME

The agricultural sector contributes the bulk of the exports by value. Cashew nuts and cotton each now account for about one-fifth of all export income. Sugar and related products account for about 10 percent, and vegetable oils, tea, sisal, and copra for smaller percentages. Timber exports, mainly from Manica e Sofala District, account for about 3 percent of annual export revenues. Mahogany, ebony, and sandalwood are the most important timbers produced. The nascent fishing industry catches fish for the home market.

Although the beginnings of a mining economy exist in Mozambique, the mining sector is still small. Petroleum products in the late 1960s accounted for about 11 percent of annual income, and coal for less than 1 percent. Iron is expected to increase in importance when development plans for the Nampa deposit are completed and exploitation begins.

The processing of primary products is the main industrial activity. Most industrial establishments are engaged in the pre-export treatment of cotton, sugar, tobacco, cashews, and other agricultural products or in the manu-

facture of products for the local market, such as footwear, foodstuffs, cement, fertilizers, beverages, and textiles. There has always been a shortage of capital, and the close proximity of South Africa and Rhodesia discourages industrial development in Mozambique.

As already mentioned, it is from the transit trade that Mozambique reaps major benefits. The junction of Lourenço Marques, which is connected by rail to Johannesburg and the Witwatersrand, as well as to Swaziland, is to serve as an outport for South African trade, and the prospect is for increasing revenues from this source. The Mozambique Convention between Portugal and South Africa stipulates that 40 percent of the southern Transvaal's transit traffic must go through Lourenço Marques. Beira has similarly earned revenue as the chief outlet for Rhodesia and Malawi and, to a lesser extent, for Zambia and even Katanga. Since 1955 the Pafuri Line has linked Lourenço Marques with the Rhodesian rail system.

Despite warfare in the north, efforts are being pursued to connect the port of Nacala in the Moçambique District with the Malawi system, thus establishing another outlet through Mozambique for a landlocked country. In the late 1960s the income derived from the transport operations of the ports and railroads was approximately $50,000,000 to $60,000,000, or over one-third of the value of all exports combined.

The African labour force

The African labour force is another source of revenue. For a century Africans have been volunteering for work in the South African mines; in 1897 the first of a series of formal agreements between the Portuguese administration and a South African government regulating migrant labour from Mozambique was signed. Apart from taxation on the earned income, withheld and paid directly to Lourenço Marques, South Africa pays the Portuguese administration a fee for every labourer who legally enters South Africa for contract work. In addition, passport and registration fees must be paid by the worker. Earlier in this century revenue obtained from migrant workers was larger, as a forced-labour system within Mozambique induced many Africans either to volunteer for work in South Africa or to enter South Africa illegally. Today conditions are different, and income from migrant labour has declined to about 2 percent of the annual total.

Tourism represents an expanding source of revenue for Mozambique, although in the late 1960s the credit derived from the industry was still below $5,000,000 annually.

MANAGEMENT OF THE ECONOMY

The economy tends to be centralized, so that private enterprise is somewhat restricted. With minor exceptions, all surface transport and communications systems are under state ownership, as are the municipal services. Mineral development has been left to private enterprise, except in large operations, where joint government–private sponsorship prevails. Manufacturing, banking, general commerce, and finance are open to private entrepreneurship. In agriculture the situation is complex. The government supports land-settlement projects and also plays a shareholder's role in farming cooperatives. There are also many privately owned farms and plantations.

The tax structure is complicated and cumbersome. Eight taxes are levied: (1) a tax paid by all government servants; (2) a personal income tax; (3) an industrial tax on corporate profits; (4) an urban building tax, drawn from rental incomes; (5) a tax on concessions granted by the government for exploitation of land, livestock, minerals, forests, or fish resources; (6) an excess profits tax; (7) a real estate transfer tax; and (8) an extraordinary tax, instituted in 1963 to raise revenue for defense.

Mozambique has participated in successive Portuguese National Development Plans (Plan de Fomento), of which the first was introduced in 1953, and the second in 1959. The agricultural settlement schemes form part of the third plan. While there has been an expectation that the Cabora Bassa project will signal a period of development, accompanied by a liberalization of economic policy—the 1965 Foreign Investment Law is seen as an augury of this—a shadow has been cast upon these hopes by conflict in the north. A number of the new developments—the extended northern railroad extension, the Namapa iron fields, and the Cabora Bassa Project itself —are near areas where Portuguese control is tenuous.

TRANSPORTATION

Transport networks are insufficiently developed, having been constructed more to serve the transit trade than the territory's own needs. Although it is possible to travel by rail from Lourenço Marques to Johannesburg and from Beira to Salisbury, there is no direct rail link between Mozambique's two major cities nor even a good road connecting them. Roads as well as railroads centre on three distinct foci: on Lourenço Marques in the south, Beira in the central region, and Moçambique in the north.

Roads. In recent development plans, the much-neglected road system of Mozambique received much attention, but in 1970 the territory, over 300,000 square miles (780,000 square kilometres) in extent, still had only about 1,100 miles (1,800 kilometres) of surfaced roads and well under 25,000 miles (40,000 kilometres) of roads of all descriptions, including mere tracks. Among the surfaced or graded gravel roads are the route from Lourenço Marques to Ressano Garcia at the Transvaal border, together with some side roads; the road from Beira to Umtali in Rhodesia, and a connecting route that leaves it at Vila Pery and runs to Tete on the Zambezi; and a stretch of road along the far north coast. The road from Lourenço Marques to Inhambane is being continued to connect with the Beira–Umtali highway. When completed it will provide the first all-weather, north–south link to the south of the Zambezi. Enormous areas of Mozambique remain virtually without roads.

Railways. Mozambique's transit function is emphasized by the disconnected pattern of its railroads. There are about 2,300 miles (3,700 kilometres) of rail in three systems. The Lourenço Marques system, owned by the Mozambique government, serves the Witwatersrand and Johannesburg; it also extends into Swaziland, from where it transports iron ore. In addition, feeder lines serve local areas in the Incomáti, Maputo, and Limpopo valleys. The Pafúri (Malvérnia) Line carries nearly half of the output of the copper fields of central Africa to the port of Lourenço Marques. The Beira system, whose main arterial line is state owned but which has two privately run feeders, serves Rhodesia, Malawi, Zambia, and Zaire. The system is V-shaped, with the western artery serving the Rhodesian interior and the northern lines serving Tete and Malawi. The Mozambique system begins at the port of Nacala and extends to Vila Cabral and the highlands near Lake Nyasa. It is the longest of the domestic lines. A connecting line is being constructed to link it with the Malawi system.

Ports. In a continent in which natural harbours are scarce, Mozambique is well endowed. Delagoa Bay, on the shores of which the port-capital of Lourenço Marques has developed, has a reputation as one of the best natural harbours in the world. Lourenço Marques can berth 16 oceangoing vessels in addition to smaller coastal steamers and oil tankers, for which special facilities exist. Several deep channels cross the bay to the open sea but dredging is necessary. Beira, a river port 15 miles (22 kilometres) from the open sea, is less favoured. The harbour is shallow, sand and silt accumulations necessitate constant dredging, and there is a troublesome bar across the mouth of the Púnguè River. Despite these disadvantages, Beira has a capacity nearly half that of Lourenço Marques. In addition to these two ports, there are several deepwater harbours along the coast. Nacala is the principal port of the north—the old Moçambique harbour is virtually abandoned—and is the eastern terminal for the Mozambique railroad system. Quelimane, located on the Cuacua, a Zambezi Delta distributary, is connected by a short railroad to the Mocuba agricultural area. Other smaller ports include Mocímboa da Praia, Porto Amélia, Memba, António Enes, Moma, Pebane, Macuze, Chinde, Mambone, and Inhambane.

Air transport. Lourenço Marques and Beira both have major international airports and are served by foreign

carriers as well as by the Mozambique airline, DETA (Direcção e Exploração dos Transportes Aéreas [Direction and Exploitation of Air Transport]). In addition, some 16 airports serve the smaller towns, and there are several dozen airfields and landing strips in the more isolated parts of the territory. Regular services are provided to the larger airports by DETA and to the international airports by Portuguese and South African carriers. The airport at Nampula in the north is an international terminal and is also being used for military purposes.

ADMINISTRATION

By the constitutional amendment of 1952, which proclaimed Mozambique to be an overseas province, the territory became an integral part of the Portuguese state. The Central Government in Lisbon, through the Overseas Ministry (which is headed by a Cabinet member), supervises the affairs of the territory. The Minister for Overseas Provinces is advised by an Overseas Council, consisting of persons representing metropolitan and overseas interests. In the Portuguese National Assembly, Mozambique holds seven of 130 seats.

The administrative structure

The province is administered by a governor general, who is appointed through the Overseas Ministry; although he has administrative powers in Mozambique, he may be overruled by Lisbon. His residence is in Lourenço Marques, where he is assisted in his executive work by an advisory group, the 18-member Economic and Social Council. This council is composed of representatives chosen by various interest groups in the province, representing business, labour, cultural organizations, and the civil service, as well as of a number of appointees. In his legislative functions the governor general consults a 29-member Legislative Council, which includes appointed as well as elected members. Some of the elected representatives are chosen by the general vote of eligible voters in the province. Although prospects of a legitimate opposition coming into existence have improved, there is no legal opposition party in Mozambique. The only political party is the National Union (*União Nacional*).

Administratively, Mozambique is divided into nine districts, or *distritos*, each with a local governor and an advisory board. These districts are further subdivided into smaller units; those enclosing urban places are called councils (*concelhos*) and those composed of rural areas are circumscriptions (*circunscrições*). African chiefs are paid by the government.

The political process is thus that prevailing in the corporate state. All executive officials in the territory, down to the level of the presidents of the urban councils, are appointed, not elected. The nominated members for the Social and Legislative councils must be acceptable to the administration. A rejected nominee may nevertheless present himself for election.

With the abolition of an official distinction between "indigenous" and "assimilated" (or "civilized") African residents of Mozambique, the composition of the voters' rolls is changing. Even should black voters come to outnumber white, however, control would remain in the hands of the central administration. Two challenges to the status quo have, nevertheless, come into existence. One is represented by the Associação do Naturais de Moçambique (Association of the Native-Born of Mozambique), an association of Mozambique-born businessmen, civil servants, labour people, and others. This organization has at times opposed all influences of metropolitan Portugal in Mozambican affairs and has taken on the characteristics of a political dissent group. Some of its members favour an independent Mozambique within the framework of a greater Portuguese community. The second challenge is represented by the two major African nationalist parties—both based in Dar Es Salaam, Tanzania—whose partisans are combating Portuguese authority in the north. One is FRELIMO (Frente de Libertação de Moçambique, "Mozambique Liberation Front"), and the other is COREMO (Comité Revolutionário de Moçambique, "Revolutionary Committee of Mozambique").

Justice. Portuguese law is applicable in Mozambique, but with some modifications. The Portuguese criminal code applies in all respects, but the civil code is appropriate to all persons who choose to be governed by it, rather than by the customary laws of traditional African society. Where conflict occurs, written law prevails. Some outlawed practices include witchcraft and sorcery as well as mutilation of the body. There are ordinary, special, municipal, and labour courts. The court of final appeal is the Metropolitan High Court in Lisbon.

Armed forces. All males are eligible for the armed forces, and military service is compulsory. Service, which begins at the age of 21, is normally of two years' duration. Since 1967, however, military service has also been open to women; the age of eligibility begins at 18, and the period of active duty is of four years' duration. For metropolitan Portuguese, two of the four years' service must be in Africa. The 1970 estimates suggest that some 45,000 troops were in the field in Mozambique.

SOCIAL SERVICES

Social services in Mozambique are poor. The latest available official figures in the early 1970s indicated that in a country of 8,000,000 people there were fewer than 1,200 teachers and fewer than 20,000 students at the second level of education (*i.e.*, ages 15 to 19). At the first level, including the initial years of schooling, attendance in 1964 was 358,378, but the attrition rate in the second and third year is about 90 percent. Portuguese reports for 1970 indicated 496,381 students were at the primary level, most of them in the special African schools. There is a university in Lourenço Marques.

Health and welfare

Health and welfare services are similarly rudimentary. The high infant mortality rate has already been mentioned. In the early 1970s there were less than 500 doctors in Mozambique, or about one doctor for every 16,500 persons, with a ratio of just over 1.5 hospital beds per 1,000 people.

The growth of shantytowns on the outskirts of the cities and towns has presented a challenge that has so far not been met, since funds for rehabilitation are not available.

The Portuguese secret police, PIDE (Polícia Internacional e de Defesa do Estado, or International Police for the Defense of the State), became operative in Mozambique for the first time in the late 1950s; since then it has questioned or detained suspected sympathizers of the African insurgency movements.

LIVING CONDITIONS

The cost of living in Mozambique is higher than in Portugal. Until the early 1960s, a dual system of wages existed that favoured the whites; Africans were paid much less for the same work and were not permitted to belong to workers' associations. Wages are low by South African standards, and many skilled and semiskilled white workers depart for South African jobs. Taxes are also high for the wage earner—a tax of 40 percent is levied on a salary of $3,000 (U.S.).

Health conditions favour the white man rather than the African; the inadequate doctor and hospital services are concentrated in the towns, and the rural areas have little medical service. A small number of medical personnel, sustained by the state or by the church, were active in some rural areas. The incidence of bilharzia—a disease communicated by water—has increased because of new irrigation projects; over 1,000,000 people were treated in outpatient clinics in 1967. Smallpox vaccination is compulsory, and it is estimated that 80 percent of the population have been vaccinated. Malaria, sleeping sickness, yaws, and diarrheal infections (the last probably the chief cause of the high infant mortality) all afflict the population, especially the African majority.

The average white resident has a better life than his African counterpart. Until the early 1960s, policies of forced labour, compulsory cropping, corporal punishment, and wage discrimination were in effect, resulting in international criticism of the Portuguese administration. While the decade of the 1960s saw the termination or revision of many of these policies, it was nevertheless too late to stave off some of the consequences. Reform in

Mozambique did away with the *shibalo* labour system; the word *shibalo* is derived from the Swahili word *shiba*, meaning serf. Under this system an African male between the ages of 18 and 55 who could not prove that he was leading a productive existence could be put to work for the state at a negligible wage. It also eliminated the *assimilados* system, whereby a few Africans who could qualify could achieve "assimilated" citizenship status. The effects of such systems, however, take time to disappear. Yet while Mozambique has always maintained a social stratification based on class, and to a considerable extent on religion, discrimination because of skin colour has always been less noticeable than elsewhere in Southern Africa. While some critics have maintained that the degree of integration achieved represents a facade behind which general racial discrimination prevails, patterns of racial mixing are observable in the public facilities of Lourenço Marques and Beira. The Portuguese no longer can afford to maintain a dual system of assimilation and subjugation. They now seek the acquiescence, if not the loyalty, of the African peoples of the territory.

CULTURAL LIFE AND INSTITUTIONS

Portuguese administrators have always taken a fierce pride in Portugal's cultural heritage and have consistently sought to inculcate Portuguese traditions.

To a considerable extent the cultural effort has taken root. Mozambique has developed a tradition that is a mixture of Portuguese and local culture. This blend is expressed in the architecture of Lourenço Marques, in the writings and poetry of black as well as white authors, in painting and sculpture, and in music. There is considerable interest in and encouragement of African traditional arts.

The major daily newspaper is Lourenço Marques' *Notícias*, which is state controlled and subject to censorship. The Radio Clube de Moçambique serves Mozambique, as well as broadcasting in English and Afrikaans to South Africa and Rhodesia.

CONCLUSION

In Mozambique, Portugal is seeking to free itself from the liabilities of an earlier colonial period in which laxity went largely unpunished and oppression provoked neither criticism nor retaliation. In a decade of adjustment and far-reaching reform, the Portuguese have sought both to redress past errors and to maintain their position. Thus, while a full-scale war is being fought in the north, the Portuguese administration is nevertheless planning to promote future development—an intention symbolized by the Cabora Bassa Project, which is due to be completed in the late 1970s. In the meantime the cost of the war in the north is heavy. Together with the wars in Angola and Portuguese Guinea, it accounted for 40 percent of Portugal's 1970 budget. Economists have calculated that Portugal can sustain the conflict indefinitely, but, on the other hand, the African armed opposition is obtaining more manpower and material, as well as more support from abroad. An increase in hostilities is anticipated, for while the African nationalists intend to prevent the construction of Cabora Bassa, the South African government, for its part, has shown a growing interest in the struggle.

Portuguese officials have repeatedly expressed the hope that Mozambique may in time emerge as an autonomous state on the Brazilian model, forming part of a greater *luso*-tropical world. For this hope to be realized, the most evident requirement is time—time to repair past mistakes and time to generate new development. The question to be answered is whether sufficient time will be available.

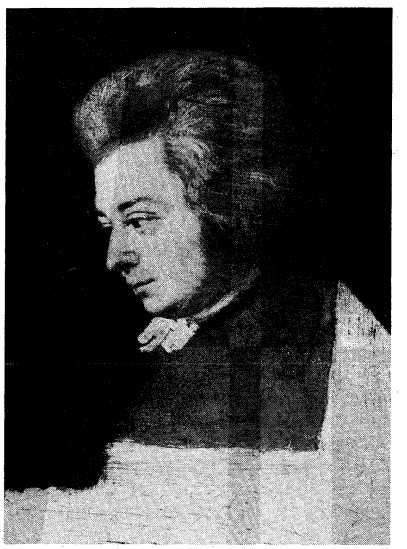
BIBLIOGRAPHY. D.M. ABSHIRE and M.A. SAMUELS (eds.), *Portuguese Africa: A Handbook* (1969), a good source for general information, but variable in quality; J. DUFFY, *Portuguese Africa* (1959), a substantial volume, historical in orientation, that deals also with contemporary problems; *Portugal in Africa* (1962), a more popular volume; A. DE FIGUEIREDO, *Portugal and Its Empire: The Truth* (1961), an analysis of the contemporary situation in Portuguese Africa, written by a Portuguese citizen; M. HARRIS, *Portugal's African Wards* (1958), a description of the forced labour and forced cropping system, written by a well-known anthropologist; A. MOREIRA, *Portugal's Stand in Africa* (Eng. trans. 1962), a reply by a Portuguese cabinet member to the criticisms of Portuguese colonial policy; R.J. HAMMOND, *Portugal's African Problem: Some Economic Facets* (1962), a description of the composition of Portuguese Africa's society, with an analysis of urban–rural contrasts, wage distribution, and class structure; J.H. WELLINGTON, *Southern Africa*, vol. 1, *Physical Geography*, vol. 2, *Human and Economic Geography* (1955), frequent references to Mozambique; G.P. MURDOCK, *Africa: Its Peoples and their Culture History* (1959), discussions of several Mozambique ethnic groups, including the Yao and Thonga peoples; C.F. SPENCE, *Moçambique: East African Province of Portugal* (1963), data on the evolution of settlement, on the exploitation of resources, and on patterns of livelihood in the early 1960s.

(H.J.deB.)

Mozart, Wolfgang Amadeus

Regarded by many as the greatest musical genius of all times, Wolfgang Amadeus Mozart was one of the world's three or four leading operatic composers and one of the central figures of the Viennese classical school.

By courtesy of the Internationale Stiftung Mozarteum, Salzburg, Austria



Mozart, unfinished oil portrait by Joseph Lange, 1789. In the Internationale Stiftung Mozarteum, Salzburg, Austria.

Early life and works. Mozart was born in Salzburg, Austria, on January 27, 1756. His parents were Leopold Mozart, composer and vice Kapellmeister to the Archbishop of Salzburg, and his wife, Anna Maria, *née* Pertl, daughter of an official from St. Gilgen. Of their seven children, only two survived: a daughter, Maria Anna (known as Nannerl), born on July 30, 1751, and Wolfgang, who was christened Joannes Chrysostomus Wolfgangus Theophilus (in German, Gottlieb, in Latin, Amadeus; later, Mozart himself generally signed his letters and compositions either "W.A. Mozart" or "Wolfgang Amadé Mozart"). Both children soon evinced profound musical talent and were taught the science of music by Leopold, who, apart from being a well-known violinist and composer, was the author of a celebrated theoretical treatise, *Versuch einer gründlichen Violinschule*, which was published at Augsburg the year of Wolfgang's birth and which went through numerous editions in various languages until long after its author's death. One of the best, and certainly one of the most authentic, descriptions of Wolfgang's childhood occurs in a letter from Leopold to Wolfgang of February 16, 1778.

> As a child and a boy you were serious rather than childish and when you sat at the clavier or were otherwise intent on music, no one dared to have the slightest jest with you. Why, even your expression was so solemn that, observing the early efflorescence of your talent and your ever grave and thoughtful little face, many discerning people of different countries sadly doubted whether your life would be a long one.

At the age of five, Wolfgang was beginning to compose minuets and other pieces—a sufficient number of them have been preserved to show us his immediate grasp of musical forms—and he soon became proficient both on the harpsichord and on the violin. By the end of the 1770s he was gradually to abandon any public performances on the violin and concentrate on the pianoforte, becoming one of its foremost virtuosos.

Leopold considered that his children's talents might be better displayed and put to more profitable use than in provincial Salzburg, and he therefore began a series of journeys throughout Europe in which Nannerl and Wolfgang played separately and together.

The early tours. The first tours, made in 1762, may be summarized as follows: Munich in January and February (for the Elector); Vienna from September 18, 1762 (where Wolfgang had an unprecedented success at the imperial court); Pressburg; Vienna; Salzburg, where they returned on January 5, 1763.

The first grand European tour lasted from June 9, 1763, to November 30, 1766, and included the principal towns of southern Germany, the Rhineland, Brussels, and Paris (for the first winter, 1763–64, where Wolfgang published his first compositions, four violin sonatas, K. 6–9 in Köchel's catalog of 626 items), followed by a year and a half in London. There, Johann Christian Bach, youngest son of Johann Sebastian, became a friend of the family, and in London Mozart wrote his first symphonies (K. 16, 19) under the influence of J.C. Bach and his colleague Karl Friedrich Abel; the J.C. Bach influence was a vital one, lasting for many years. The Mozarts spent the winter of 1765–66 in Holland and then returned to Austria via Brussels, Paris, Geneva, Berne, and Munich.

First published works

The family left Salzburg for a second trip to Vienna on September 11, 1767, and returned on January 5, 1769. In Vienna, Wolfgang composed his first German operetta, *Bastien und Bastienne*, K. 50, and his first Italian opera, *La finta semplice*, K. 51, on a libretto by Marco Coltellini, who also provided books for Christoph Willibald von Gluck and Franz Joseph Haydn. Although the Italian opera was ordered by the Emperor, intrigues prevented its being produced; this is the beginning of a sorry tale of such resistance, not only to Mozart but also to Haydn, whose opera *La vera constanza* (composed 1776, produced 1779), commissioned by the court opera, was withdrawn by the composer in the face of strong opposition from the Italian party, led (more or less openly) by the composer Antonio Salieri. For the consecration of the Waisenhäus (Orphanage) Church on December 7, 1768, Wolfgang composed a new mass (recently identified as K. 139, in C minor) and other works, which he conducted before the Emperor and the court. Upon his return to Salzburg, the benevolent Archbishop Sigismund von Schrattenbach, always a friend and patron of the Mozart family, had the rejected *La finta semplice* performed in the archiepiscopal palace and made Wolfgang his *Konzertmeister* (without salary, since Wolfgang was hardly ever in Salzburg).

The Italian tours. The first Italian tour, which lasted from December 13, 1769, to March 28, 1771, included Rovereto, Verona, Milan, Parma, Bologna, Florence, Rome, and Naples. On their return, Leopold and Wolfgang spent three months in Bologna, where Wolfgang studied counterpoint with Padre Martini, and more than an equal period at Milan, where he executed a commission for a new opera, *Mitridate, rè di Ponto*, K. 87, which was first produced at the Teatro Reggio Ducal on December 26, 1770, with a large orchestra (14 first and 14 second violins; in all about 60 players) and repeated 20 times with enormous success. In March 1771 they were at Padua, where Wolfgang was asked to write a new oratorio, *La Betulia liberata*, K. 118, "which he can do at his leisure."

On the second Italian journey, which lasted from August 13 to December 15, 1771, Wolfgang went to Milan to compose and produce a dramatic serenata, *Ascanio in Alba*, K. 111, commissioned by the empress Maria Theresa for the marriage of her son, Archduke Ferdinand, to Princess Maria Beatrice d'Este of Modena. He composed the whole score in four weeks, and it was performed with great success on October 17, 1771, to the extent that Leopold had to write that "Wolfgang's serenata has killed Hasse's opera," which had been performed the day before —and that despite the colossal reputation of *il sassone* ("The Saxon"), as Johann Adolf Hasse was called in Italy.

On his third Italian journey, from October 24, 1772, to March 13, 1773, Mozart returned to Milan to execute the new commission for the Teatro Reggio Ducal, *Lucio Silla*, K. 135, an opera seria, first performed on December 26, 1772. Two weeks later Leopold could write, "Thank God, the opera is an extraordinary success, and every day the theatre is surprisingly full"; the opera ran for about 20 performances. *Lucio Silla* is a significant milestone in Mozart's stylistic development; it is full of felicitous details and the orchestration is particularly full and rich. Meanwhile, Archbishop Schrattenbach had died, to be succeeded by Hieronymus, Count Colloredo, bishop of Gurk, with whom the Mozarts never had the same friendly relations that they had enjoyed with his predecessor. For the new archbishop's installation on April 29, 1772, Mozart composed the festival opera *Il sogno di Scipione*, K. 126. While at Milan, he wrote the famous motet *Exsultate, jubilate*, K. 165, for the celebrated castrato Rauzzini, one of the singers in *Lucio Silla*.

Vienna, Salzburg, and Munich. The Mozarts visited Vienna again from July 14 to September 26, 1773. Father and son hoped to secure an appointment for the latter at the imperial court—to no avail. But in Vienna, Mozart came under the profound influence of Haydn's music, whose style—in common with that of many Austrian composers of the early 1770s (*e.g.*, Florian Leopold Gassmann, Carlos Ordoñez, J.B. Vanhal)—had undergone a radical change. From lighthearted rococo music, the new Haydn style now swerved abruptly to the *Sturm und Drang* ("storm and stress"), the title of a literary movement in Germany, derived from a characteristic play of that title by a German playwright, Friedrich Maximilian von Klinger. The Austrian musical "revolution" preceded that in German literature by some years, and the *Sturm und Drang* style began with Haydn as early as *c.* 1767–68.

Early quartets and symphonies

Mozart composed six string quartets (K. 168–173) under the shadow of Haydn's Opus 20 quartets of 1772; and Mozart's stormy *Symphony in G Minor*, K. 183, is a typical and excellently composed product of the restless Viennese symphonic style of this period—even the use of four, rather than two, horns is characteristic (see Vanhal's *Sinfonia in G Minor* and Haydn's *Symphony No. 39 in G Minor*, for two obvious precursors in the field). For the rest of 1773 and almost all of 1774, the family remained at Salzburg, Wolfgang diligently composing. Perhaps his best known compositions of this interlude are the *Bassoon Concerto in B Flat Major*, K. 191, and four symphonies (K. 199–202), of which the A major, K. 201, is the most popular.

Leopold and Wolfgang went to Munich from December 6, 1774, to March 8, 1775, where Wolfgang carried out a commission for an opera buffa, *La finta giardiniera*, K. 196. The day after its performance in January 1775, at the Residenz-Theater, Wolfgang writes: "Thank God! My opera was performed yesterday, the 13th, for the first time, and was such a success that it is impossible for me to describe the applause . . . the whole theatre was . . . packed." If Mozart hoped to enter the Elector's service, nothing came of the plan, and he returned to Salzburg, where he remained for the rest of 1775, the whole of 1776, and three-quarters of 1777. A very large quantity of music was composed, including the opera *Il rè pastore*, K. 208, an attractive and sophisticated piece on a rather over-used libretto by Pietro Metastasio; several masses, in which Wolfgang attempted to create a popular, "folklike" mass—of these the brilliant *Mass in C Major* (*Credo Mass*), K. 257, is particularly successful; the five famous violin concerti (K. 207, 211, 216, 218, 219); various serenades and divertimenti, foremost among them the enchanting *Serenata notturna*, K. 239, for strings and kettledrums, and the grand and lengthy *Serenade in D Major* (*Haffner Serenade*), K. 250, out of which

Mozart later created a symphony (adding kettledrums and making some other alterations), which has, curiously, escaped attention and has not even a Köchel number. Mozart also wrote several works for the pianoforte, an instrument that was gradually assuming great importance in his artistic world (six sonatas for Baron Dürnitz, K. 279–284; the *Triple Concerto*, K. 242; two concerti, K. 238, 246; etc.). Mozart's relations with the archbishop Colloredo became increasingly strained, and sometime in August 1777 (the document is undated) the composer applied for discharge from his duties. On August 28 the Archbishop dictated his reply, that "father and son have permission to seek their fortune elsewhere, according to the Gospel"; however, the final break between Mozart and the Archbishop did not occur until 1781.

Paris, Salzburg, and Munich. In September 1777 Mozart and his mother left Leopold behind and started on a long journey through Germany to Paris, where Wolfgang hoped to find some permanent appointment. They went to Munich, then to Mannheim, where Mozart fell in love with Aloysia, second daughter of the prompter and copyist Fridolin Weber. Aloysia had an exceptionally well trained coloratura soprano voice, and Mozart entertained the idea of going with her to Italy, but his father sternly forbade it. In Paris Mozart did not really gain any foothold, and on July 3, 1778, his mother died. He slowly returned home, arriving at Salzburg in the middle of January 1779. When on his return journey he stayed with the Webers, who had meanwhile moved to Munich, Aloysia scarcely seemed to recognize him. ("Today," writes Mozart on December 29, 1778, "I can only weep . . .").

Works of early maturity

Of the various compositions written on this trip, the best known are the six violin sonatas (K. 301–306), to which should be added a seventh (K. 296), some piano sonatas (including the famous A minor, K. 310), the music to *Les Petits riens* (a ballet by Jean-Georges Noverre, written in Paris), and three symphonic works for Joseph Legros of the Concert Spirituel, of which only one—the *Symphony No. 31 in D Major* (*Paris Symphony*), K. 297—has survived in its authentic and original state. Meanwhile, various court musicians had died at Salzburg, and Mozart applied in January 1779 for the position of court organist, formerly held by Anton Cajetan Adlgasser. His request was granted. The whole of 1779 and most of 1780 were spent in Salzburg, where Mozart's music enters its first period of consistent greatness. This new maturity, with dark-hued undertones that were to assume an ever-increasing importance in his style, is found in everything Mozart composed—the series of church music (particularly the two great C major masses, the *Coronation Mass*, K. 317, and the solemn and grandiose *Missa solemnis* in C major, K. 337); the dramatic and extremely modern music to *Thamos, König in Ägypten*, K. 345, which brought him into contact with Emanuel Schikaneder, whose touring company (the Böhm-Schikaneder Troupe) was then visiting Salzburg; two symphonies, of which the *Symphony No. 34 in C Major*, K. 338, is outstanding among Austrian symphonies at that time; the celebrated *Sinfonia concertante in E Flat Major for Violin, Viola and Orchestra*, K. 364; and his greatest and most profound orchestra serenade, the *Serenade in D Major* (*Post Horn Serenade*), K. 320, which Mozart later shortened and adapted into a symphony like that fashioned from the *Haffner Serenade*.

While in Salzburg Mozart, through friends at Munich, received a commission to compose a new opera seria, *Idomeneo*, K. 366 (and the ballet music K. 367). He went to Munich early in November 1780, and on January 29, 1781, the new opera was most enthusiastically received; the Archbishop being in Vienna at that time, Leopold and Nannerl came to Munich to witness Wolfgang's new triumph. For *Idomeneo* was the first stage work that revealed Mozart at the summit of his powers, and, although it is cast in the antiquated form of an opera seria, the music is so profound, so brilliantly composed, and so astonishingly orchestrated, that it retains its dramatic impact, despite the stereotyped libretto by the Salzburg court chaplain Abbate Varesco. Of other compositions written in Munich, the *Kyrie in D Minor*, K. 341, and the exquisite *Serenade in B Flat Major*, K. 361, for 13 wind instruments are outstanding—the latter showing an entirely new and original treatment of the wind band as a whole and the clarinet family (including basset horn) in particular. Mozart always preserved a special affection and aptitude for the clarinet, whose beauty had been scarcely realized by his precursors (Haydn, for example, had hardly used it since a few wind-band works of *c.* 1760).

Early in March 1781 the Archbishop summoned Mozart to Vienna, where he arrived on March 16. Mozart enjoyed the patronage of many members of the Viennese aristocracy, including Countess Thun and Count Cobenzl, and the Archbishop enjoyed the reflected glory of his young composer; but Colloredo did not allow him to play in houses other than his own and forced him to eat with the servants. The already strained relations between the two soon came to a head, and in a shameful scene the Prince-Archbishop of Salzburg addressed Mozart in the language of the street (using the third person singular reserved for low-grade servants), and Mozart resigned. Shortly afterward, not having received his official dismissal, Mozart brought it in person, only to be subjected to more street language from the Archbishop's high steward, Count Karl Arco, who earned for himself a dubious immortality by actually kicking Mozart out "on his behind" (Mozart's own words).

Later life and works. Mozart was now entirely on his own, the first composer after George Frideric Handel who risked an existence without some official patronage. After years of waiting, the emperor Joseph II finally engaged him on December 7, 1787, as chamber composer at 800 gulden annually (Gluck, his predecessor, had received 2,000 gulden). At the beginning of 1781, however, Mozart's prospects—despite Leopold's dismayed reactions from Salzburg—seemed favourable. Apart from his success at various concerts, at which the Emperor also applauded, Mozart soon received a commission to write a German opera for the National Singspiel, founded by the Emperor in 1778. This work, *Die Entführung aus dem Serail* (*The Abduction from the Seraglio*), on a libretto by Gottlieb Stephanie, the younger, after Christoph Friedrich Bretzner, was produced with great success on July 16, 1782, though a strong cabal had already aligned itself against Mozart. Meanwhile, he had moved in with the Webers, who had come to Vienna from Munich, and he fell in love with the third daughter, Constanze. Despite Leopold's strenuous protests, Mozart married Constanze at St. Stephen's Cathedral on August 4, 1782. There has been a great deal of criticism about Constanze—that she was fickle, a bad manager of finances, and so on—but they were very fond of each other, and Mozart wrote her affectionate letters all his life. The considered opinion of the great 19th-century Austrian scholar C.F. Pohl was that "she was a good and loving wife, accommodated herself in everything to her husband's disposition, and restrained him from many heedless actions."

Marriage to Constanze

Mozart soon came to hold his own subscription concerts, either at the Burgtheater or the Mehlgrube, which consisted of a full orchestra, with one or more vocal soloists, and Mozart himself playing his own concerti for piano and improvising on that instrument; his improvisations were much admired. It was for these concerts that Mozart wrote most of his great piano concerti, from K. 413 to K. 595. Mozart was not only a great pianist but was profoundly interested in the mechanical aspects and construction of the piano, to the development and perfection of which he contributed a great deal. Some of his improvisations may, perhaps, be gathered from such work as the *Fantasy in C Minor*, K. 475. Mozart had a pedal constructed for his piano, which he used when improvising and possibly also in realizing the basso continuo parts of his concerti.

Despite an intense activity as a pianist, Mozart's finances were never very secure. He perhaps reached the apex of his worldly gains shortly before he composed *Figaro*, and was living in comfortable quarters at the Domgasse behind St. Stephen's. Gradually, from 1785 to 1791, his

financial situation became more and more catastrophic, and he was forced to borrow ever-increasing sums from his friends, and particularly his Masonic brothers—Mozart was an ardent Freemason and a member of the Lodge "Zur gekrönten Hoffnung." It was from his fellow Mason, the banker Michael Puchberg, that Mozart borrowed most freely. The ever-present threat of these debts hung over Mozart's existence until his death.

Wolfgang and Constanze visited Leopold and Nannerl in Salzburg in August, September, and October of 1783; the visit was not a success, and Constanze was not much liked by the Salzburg Mozarts. Mozart performed the "Kyrie" and "Gloria" of his great (unfinished) *Mass in C Minor*, K. 427, in St Peter's in August 1783; although a torso, the mass is a noble work, showing Mozart's renewed study of Bach and Handel at the Sunday concerts of Baron Gottfried van Swieten, director of the National Library in Vienna, who was later to be closely connected with Haydn and to become a patron of the young Beethoven.

The great symphonies

Mozart had little time in the last ten years of his life to compose symphonies, but those he did manage to write are all masterpieces: the *Haffner*, K. 385, reduced from another serenade for that Salzburg family; the *Linz*, K. 425, composed on the way back from Salzburg to Vienna in October and November 1783; the *Prague*, K. 504, written for the Bohemian capital; and the last three (E flat major, G minor, and *Jupiter* in C major of 1788, K. 543, 550, 551). The last four, in particular, carry the art of the 18th-century symphony far beyond anything hitherto achieved except in a few of Haydn's works.

Leopold returned the visit to his son and daughter-in-law, from February 11 to April 25, 1785, and was witness to his son's material and artistic successes. Mozart was in the process of completing the six string quartets dedicated to Haydn (K. 387, 421, 428, 458, 464, 465), and when the last three were played to Haydn in Mozart's flat, Haydn said to Leopold: "Before God and as an honest man I tell you that your son is the greatest composer known to me either in person or by name." The string quartets also show Mozart to have mastered the form in a way far superior to anyone except Haydn; the ensuing works, particularly K. 499 in D major, but even the more superficial "Prussian" quartets for Berlin, show his complete grasp of this central form of chamber music. His quartets, indeed, are perhaps only surpassed in quality by the string quintets (of which two, C major and G minor, K. 515 and 516, are the most popular, though a lesser known one in D major, K. 593, is equally fine).

Late operatic works

The great operas that were composed after *Die Entführung* in these last ten years are as follows: *Le nozze di Figaro* (*The Marriage of Figaro*), K. 492, the first of three masterpieces written on librettos by Lorenzo da Ponte, first performed with enormous success on May 1, 1786, at the Burgtheater in Vienna and later repeated, with even more approbation, at Prague; *Don Giovanni*, K. 527, again on a libretto by da Ponte, first played at the National Theatre, Prague, on October 29, 1787, with sensational success (but when repeated at Vienna the next year, it was a failure although a few—among them Haydn—violently defended it); *Così fan tutte*, K. 588, the last da Ponte opera, produced at the Vienna Burgtheater on January 26, 1790, was not entirely successful, partly due to the fact that the emperor Joseph II died on February 20, throwing Vienna into some confusion; but *Così* was perhaps too richly scored, too cynical, too erotic, and too overpoweringly intense for the easygoing Viennese, who were amused by *Figaro* but less enchanted by its political undertones; *La clemenza di Tito*, K. 621, an opera seria on a libretto by Metastasio, was commissioned by the Estates of Bohemia for the coronation of the new emperor Leopold II at Prague in 1791; it was given there on September 6, but was badly sabotaged by the court in general and the Empress in particular, who referred to it as a *porcheria tedesca* ("German muck"). *Die Zauberflöte* (*The Magic Flute;* libretto by Emanuel Schikaneder), K. 620, was written for the little Theater auf der Wieden and made Schikaneder wealthy and famous. The first performance took place on September 30, 1791, and, while the initial reception was rather cool, audiences soon took to its fairy-tale "machine comedy" world, with its strong Masonic elements, the whole made immortal by Mozart's fantastic music. (Beethoven was to admire it particularly because, he said, in it Mozart wrote in every kind of form; it is certainly true that in *Die Zauberflöte* we find everything from north German *Choral-Vorspiel* to fugue, and from coloratura aria to *Volkslied*.)

In another attempt to save his failing finances, Mozart undertook an extensive concert tour, leaving Vienna on April 8, 1789, with his pupil and noble patron, Prince Karl Lichnowsky, who was later to be a firm supporter of Beethoven. Mozart went to Dresden, Leipzig (where he studied J.S. Bach's manuscripts), Berlin, and Prague, returning to Vienna on June 4, 1789; the financial results of the tour were negligible, as were those from his last concert tour to Frankfurt am Main, where the emperor Leopold II was crowned on October 9, 1790. His financial position was now catastrophic, for he was deeply in debt.

In the summer of 1791 a stranger commissioned the *Requiem Mass*, K. 626, from Mozart under rather mysterious circumstances; in fact the man behind it was Count Franz von Walsegg, an amateur musician who ordered works from professional composers and performed them under his own name. Mozart was not able to complete the *Requiem*, but from oral instructions and sketches (some have come to light as late as 1960), his pupil Franz Xavier Süssmayr was able to finish the work and present it to Walsegg.

Mozart's death

Mozart became seriously ill in the late autumn of 1791, and his rapid decline and death soon led people to whisper that he had been poisoned. But in the mid-20th century the medical research of Carl Bär showed that Mozart's heart had been strained by his youthful attacks of rheumatic fever and that his final attack of that same fever in December 1791, plus excessive bloodletting, contributed to the failure of his heart. He died in Vienna on December 5, 1791, and the next day he was buried, unattended by anyone except the gravedigger, at St. Mark's cemetery. Constanze was left, brokenhearted and in a state of nervous collapse, with huge debts. She later married the Danish diplomat Georg Nicolaus von Nissen and died in Salzburg in 1842.

Assessment. Because most of Mozart's major works were not printed until after his death—and some not until the middle and end of the 19th century—Mozart did not create a school. Although his operas were widely circulated all over Germany by 1795, it was many years before his work *in toto* was known to the average European musician; London in 1795, for instance, knew hardly anything of Mozart's music. During the Romantic age Mozart became the object of a veneration such as has hardly ever been accorded to any other musician except—later—Bach and Beethoven; and even today the sheer beauty, perfection, and profundity of his music continue to astonish and delight us. If Mozart wrote with staggering ability in every single form—including even dance music—perhaps his greatest single achievement is in the characterization of his operatic figures. They are not only subtle and very human, they are also capable of changing as the course of the drama unfolds. After Mozart perhaps only Giuseppe Verdi ever succeeded in creating such vital operatic personalities. Altogether it is possible that in Mozart the music of Western civilization reached an apex of perfection to which future ages will look back, just as, for centuries, men regarded with nostalgia the products of ancient Greece at the summit of its achievements.

MAJOR WORKS

The Köchel catalog forms the basic list of the works of Mozart. The compositions of Mozart are almost always identified by the letter K. (or, in German, KV, for *Köchelverzeichnis*, "catalog-register") followed by an Arabic numeral, which indicates the chronological order in which the work was composed.

Ludwig Ritter von Köchel is best remembered for the monumental *Chronologisch-thematisches Verzeichnis*, a thematic catalog that lists in chronological order of composition all the works of Mozart. Mozart did not always date

his compositions; as a result, there was a great deal of confusion as to the order and period of his works. It was to the task of bringing order out of chaos that Köchel set himself. Köchel had, however, a base upon which to build. Mozart himself had compiled a list of his works from February 1784. This list, plus an imperfect catalog by Aloys Fuchs (1837), and a publication based on Mozart's autograph scores by Johann Anton André (1828) were available to Köchel. Köchel classified the hundreds of unnumbered Mozart compositions into 23 categories and, on the basis of stylistic development and Mozart's musical handwriting, assigned a date of composition to each work. Each composition was given a single number, whether it was a single minuet for piano (K. 1, composed in 1762), an entire opera (K. 492, *The Marriage of Figaro*, 1786), or a work of several sections, such as K. 626, the *Requiem Mass* (1791—the final entry in the catalog). Köchel also included as appendices lists of works lost, existing as fragments, or of doubtful authenticity.

The Köchel catalog was published by Breitkopf and Härtel in 1862 and appeared in six editions. The third and definitive revision of the catalog is Alfred Einstein's (1937); he assigned new numbers of more accurate chronology. This revision, in turn, was supplemented by its author, and in 1947 these supplements appeared as an appendix to the 1937 revision.

Theatre music

OPERAS: 16 operas, including *Idomeneo, rè di Creta*, K. 366, first performed 1781; *Die Entführung aus dem Serail*, K. 384, 1782; *Le nozze di Figaro*, K. 492, 1786; *Don Giovanni*, K. 527, 1787; *Così fan tutte*, K. 588, 1790; *Die Zauberflöte*, K. 620, 1791; *La Clemenza di Tito*, K. 621, 1791.

OTHER THEATRE WORKS: Music for three plays; one Latin comedy (intermezzo); two serenatas; one ballet.

Church music

MASSES: *Requiem Mass* in D minor, K. 626, completed 1791 by Süssmayr; eight masses, including K. 317 in C major (*Coronation*), 1779; ten *Missae Breves*.

OTHER CHURCH WORKS: Four litanies; seven offertories; two vespers; two antiphons; three *Regina coeli* settings; two German motets; two Latin motets, K. 165 (*Exsultate jubilate*), 1773, and K. 618 (*Ave, verum corpus*), 1791; psalms and kyries; Passion cantata; oratorio, K. 118 (*La Betulia liberata*), 1771.

Secular vocal music

57 solo arias with orchestra; 9 vocal trios and 5 duets with instrumental accompaniment; 36 German songs for solo voice and piano; 35 unaccompanied canons and one sacred madrigal.

Instrumental music

SYMPHONIES: 41 symphonies, including K. 297 in D major (*Paris*), 1778; K. 385 in D major (*Haffner*), 1782; K. 425 in C major (*Linz*), 1783; K. 504 in D (*Prague*), 1786; K. 551 in C major (*Jupiter*), 1788; fragments of 7 other symphonies and 9 *sinfonie* related to other works.

SINFONIE CONCERTANTI: K. 364 in E flat major (for violin, viola, and orchestra), 1779; K. App. 9 (297b) in E flat major (for oboe, clarinet, horn, and bassoon), 1778.

CONCERTI: (PIANO): 21 concerti for piano and orchestra, including K. 537 in D major (*Coronation*), 1788; arrangements, separate movements, and cadenzas; one concerto for three pianos; one for two pianos. (VIOLIN): five concerti for violin and orchestra, all 1775; K. 190 (*Concertone* for two violins), 1773; other movements for violin and orchestra. (OTHER INSTRUMENTS): one concerto for bassoon, K. 191, 1774; two for flute, 1778; four for horn, 1782–86; one for clarinet, K. 622, 1791.

OTHER INSTRUMENTAL WORKS: 23 divertimenti; 13 serenades, including K. 250 in D major (*Haffner*), 1776; K 375 in E flat major (for two oboes, two clarinets, two basset horns, four horns, two bassoons, and double bass or double bassoon), 1781; and K. 525 in G major (*Eine kleine Nachtmusik*), 1787; one notturno and 2 cassations; marches and dances. (FOR PIANO): 20 sonatas; 16 sets of variations; 14 minuets; *Forty-three Pieces;* 6 fantasies, 2 with fugues; 2 fugues; 3 rondos; a number of other separate movements; 2 sonatas and one fugue for two pianos. (FOR STRINGS): 6 quintets, including K. 515 in C major and K. 516 in G minor, both 1787; 25 quartets, 6 dedicated to Haydn (including K. 458 in B flat, *The Hunt*), 1784, and 3 dedicated to the King of Prussia, 1789–90; 2 string duets and 3 trios. (FOR STRINGS WITH ACCOMPANIMENT): 35 sonatas for violin and piano; 2 piano quartets; 8 piano trios, including K. 498 in E flat major (for clarinet, viola, and piano), 1786; 2 sets of variations for violin and piano. (FOR OTHER INSTRUMENTS): 3 flute quartets; one oboe quartet; one clarinet quintet in A major, K. 581, 1789; 2 other quintets; 5 divertimenti for two clarinets and bassoon; 12 duets for wind instruments; other works for wind; 3 pieces for mechanical organ and 2 for glass harmonica and wind instruments.

BIBLIOGRAPHY. The Mozart literature is so vast that a mere listing of the titles would require several hundred pages. The following, therefore, is a rigorous selection of what might be called the basic Mozart literature. The most complete Mozart bibliography in recent years will be found in the Mozart article contained in *Die Musik in Geschichte und Gegenwart*, vol. 9, col. 812–839 (1961). FRANK X. NIEMTSCHEK, *Leben des K.K. Kapellmeisters Wolfgang Gottlieb Mozart* (1798; Eng. trans., *The Life of Mozart*, 1956), is the first book-length biography using authentic documents and containing information given to the author by Mozart's widow, Constanze. It is not scientific and is fragmentary but contains much invaluable material. G.N. VON NISSEN, *Biographie W.A. Mozarts* (1828), a large-scale biography written by Constanze Mozart's second husband, is essential to all Mozart scholars. O. JAHN, *W.A. Mozart*, 4 vol. (1856–59; Eng. trans., *Life of Mozart*, 3 vol., 1960), is the first scholarly attempt at a full-scale Mozart biography; it is in many respects superseded but contains important material. G. NOTTEBOHM, *Mozartiana* (1880), contains much material then new and reproduces a number of documents of which the original sources have since disappeared. H. ABERT, *W.A. Mozart* (1923; new ed. rev. by Abert's daughter, A.A. ABERT, 2 vol., 1956), is the great standard work of the 20th century that no Mozart lover or scholar can be without; apart from dealing with Mozart's music brilliantly, Abert also provides the first serious examination of the Italian operas by Mozart's precursors and contemporaries. A. EINSTEIN, *Mozart* (1945; Eng. trans., 2nd ed., 1956), is the sophisticated and intelligent result of many years' research for the 3rd edition of the Köchel catalog. R.M. HAAS, *Wolfgang Amadeus Mozart*, 2nd ed. (1950), is a brilliant study of particular perception concerning the music itself. The *Mozart-Jahrbuch* (1950–) is the yearly journal of Mozart studies in which the world's leading Mozart scholars offer their new discoveries. *Acta Mozartiana* (1954–) is a smaller publication but contains valuable material. O.E. DEUTSCH, *Mozart: Die Dokumente seines Lebens* (1961; Eng. trans., *Mozart: A Documented Biography*, 2nd ed., 1966), contains all the known contemporary documents about Mozart and is essential for any serious study of the composer. The same author's *Mozart und seine Welt in zeitgenössischen Bildern* (1961) is the finest iconographical study of Mozart and the world in which he lived. H.C. ROBBINS LANDON and D. MITCHELL (eds.), *The Mozart Companion*, 2nd ed. (1965), is a collection of essays on Mozart's music by a group of specialists, C. BAR, *Mozart—Krankheit, Tod, Bergräbnis* (1966), is a doctor's serious and penetrating analysis of Mozart's death, in which the theory that he was poisoned is disproved.

(H.C.R.L.)

Muʿāwiyah I

Muʿāwiyah ibn Abī Sufyān, a leader of early Islām, founded the Umayyad dynasty, which ruled the empire of the Caliphate from AD 661 to 750; he stands out as one of the few caliphs who is depicted both in Muslim historiography and in modern scholarship as a decisive force in Islāmic history. Undoubtedly one reason for the prominence assigned to him is that he was a controversial figure. Pious scholars of the dominant Sunnī sect of Islām together with writers of the minority, dissenting Shīʿah, have always heaped opprobrium on Muʿāwiyah: the Sunnīs because of his deviations from the pattern of leadership set by the Prophet Muḥammad and the "rightly guided" caliphs; the Shīʿah because he had usurped the caliphate from ʿAlī ibn Abī Ṭālib. Although Muʿāwiyah has been and still is condemned for his sins from these two quarters, he has also been the subject of lavish praise in Arabic literature as the ideal ruler. In other words, unlike most of the other caliphs, Muʿāwiyah looms large in Islāmic history because he has consistently aroused partisanship at different extremes. But beneath the biassed portraits given in traditional Muslim historiography, there is a person whose actual accomplishments were of great magnitude quite apart from partisan value judgments and interpretations. These accomplishments lay primarily in political and military administration, through which Muʿāwiyah was able to rebuild a Muslim state that had fallen into anarchy and to renew the Arab–Muslim military offensive against unbelievers.

Early years. It is ironic that a man who was to become political–religious head of Islām was born (*c.* 602) into a clan (ʿAbd Shams) that rejected Muḥammad, the Prophet, in his home city, Mecca, and continued to oppose him on the battlefield after he had emigrated to Medina. Muʿāwiyah did not become a Muslim until Muḥammad had conquered Mecca and had reconciled his former enemies by gifts. Possibly as a part of Muḥammad's policy of conciliation, Muʿāwiyah was made a scribe in his service. But Muʿāwiyah's contributions to Islāmic history are wholly associated with his career in Syria, which began shortly after the death of the Prophet, when he, along with his brother Yazīd, served in the tribal armies sent from Arabia against the Byzantine forces in Syria. Upon the death of Yazīd in 640, Muʿāwiyah was appointed governor of Damascus by the caliph ʿUmar and gradually gained mastery over other areas of Syria. By 647 Muʿāwiyah had built a Syrian tribal army strong enough to repel a Byzantine attack and in subsequent years to take the offensive against the Byzantines that resulted in the capture of Cyprus (649) and Rhodes (654) and a devastating defeat of the Byzantine navy off the coast of Lycia in Anatolia (655). At the same time, Muʿāwiyah periodically dispatched land expeditions into Anatolia. All these campaigns, however, came to a halt with the accession of ʿAlī to the caliphate, when a new and decisive phase of Muʿāwiyah's career began.

As kinsman of the slain caliph ʿUthmān, Muʿāwiyah bore the duty of revenge. Because ʿAlī neglected to apprehend and punish the murderers, Muʿāwiyah regarded him as an accomplice to the murder and refused to acknowledge his caliphate. Thereupon ʿAlī marched to the Euphrates border of Syria and engaged Muʿāwiyah's troops at the famous battle of Ṣiffīn (657). Muʿāwiyah's guile turned near defeat into a truce. Resorting to a trick that played upon the religious sensibilities of ʿAlī's forces, he persuaded the enemy to enter into negotiations that ultimately cast doubt on the legitimacy of ʿAlī's caliphate and alienated a sizable number of his supporters. When these former supporters—the Khārijites—rose in rebellion against ʿAlī, Muʿāwiyah took advantage of ʿAlī's difficulties in Iraq to send a force to seize control of Egypt. Thus, when ʿAlī was assassinated in 661, Muʿāwiyah held both Syria and Egypt and, as commander of the largest force in the Muslim Empire, had the strongest claim to the caliphate. ʿAlī's son Ḥasan was persuaded to remove himself from public life in exchange for a subsidy, which Muʿāwiyah provided.

The battle of Ṣiffīn

Caliphate (661–680). During his twenty-year governorship of Syria and during the war against ʿAlī, Muʿāwiyah had succeeded in recruiting and training a large Arab tribal army, which was remarkably loyal to him. It was therefore natural that he should base his caliphate in Syria, with Damascus as the new capital of Islām. But if Muʿāwiyah's chief support came from the tribes of Syria, the tribes of other areas posed the chief threat to his reign. It is not surprising then that early Umayyad government followed certain tribal principles as a means of retaining and winning the loyalty of the Arabs. The clearest examples of such a policy are provided by Muʿāwiyah's adoption of two tribal institutions: the council of notables—the *shūrā*—convoked by the caliph for consultation and the delegations—the *wufūd*—sent by tribes to keep the caliph informed of their interest. Within this context, Muʿāwiyah ruled as a traditional Arab chieftain. Although he may not have consciously encouraged renewed warfare against non-Muslim territory as a means of directing Bedouin aggressive tendencies into channels that would aggrandize Islām and stabilize his own power, there is no doubt that warfare served these purposes during his reign; and in this respect it is significant that Muʿāwiyah used the Syrian army only for domestic defense and for campaigns against the Byzantines, who threatened the borders of Syria. During the civil war, Muʿāwiyah had purchased a truce with the Byzantines in order to free his army for the struggle against ʿAlī. Soon after his accession to the caliphate, however, he curtailed the payment of tribute and sent expeditions against the Byzantines on an almost yearly basis. These campaigns served both to fulfill Muʿāwiyah's obligation to conduct holy war (*jihād*) against unbelievers and to keep his Syrian troops in fighting trim. Otherwise, the war against Byzantium was inconclusive. Even though two expeditions reached the vicinity of Constantinople, the Arabs never succeeded in permanently occupying territory in Asia Minor beyond the Taurus Mountains. Troops stationed in other parts of Muʿāwiyah's empire were sent on campaigns into remote areas. In North Africa, raids were conducted as far west as Tlemcen in present-day Algeria. More permanent, however, was the conquest of Tripolitana and Ifrīqīyah, which was consolidated by the foundation in 670 of the garrison city of Qairawan, soon to become the base for further expansion later in the Umayyad period. At the same time, a vigorous campaign was being conducted in the east by which Muslim borders were extended to the Oxus River, and Khorāsān was established as an Umayyad province.

Administrative developments

It had become apparent during the reigns of the first caliphs that tribal tradition and the practice of Muḥammad in Medina were inadequate resources for administering a vast empire. To solve this problem, Muʿāwiyah resorted to a solution that lay at hand in Syria—that is, the imitation of administrative procedures that had evolved during centuries of Roman and Byzantine rule. Although the process by which the borrowing took place is not clear, it is clear that Muʿāwiyah initiated certain practices that were apparently inspired by the previous tradition. Basically, he aimed at increased organization and centralization of the caliphal government in order to exert control over steadily expanding territories. This he achieved by the establishment of bureaus—*dīwāns*—in Damascus to conduct the affairs of government efficiently. Early Arabic sources credit two *dīwāns* in particular to Muʿāwiyah: the *dīwān al-khatam*, or chancellery, and the *barīd*, or postal service, both of which were obviously intended to improve communications within the empire. Prominent positions within the nascent bureaucracy were held by Christians, some of whom belonged to families that had served in Byzantine governments. The employment of Christians was a part of a broader policy of religious tolerance, necessitated by the presence of large Christian populations in the conquered provinces, especially in Syria itself.

Such administrative innovations coupled with the observance of tribal traditions caused historians of a later period to deny Muʿāwiyah the religious title of caliph and to characterize him as a king (*malik*) instead. As a symbol of the increasingly secular nature of the caliphate, derived in part from a non-Islāmic tradition, the title is apt for Muʿāwiyah and for most of the Umayyads. It is particularly appropriate for the most startling of all of Muʿāwiyah's innovations, the one by which he secured the allegiance of the tribes for the caliphate of his son Yazīd and thereby established the practice of hereditary rule in Islām. As an alternative to the various unreliable precedents for selecting a caliph, this measure was certainly consonant with Muʿāwiyah's policy and achievement as caliph, which, in summary, consisted of invigorating the theocratic origins of Islāmic governance with borrowings from other traditions better adapted to the demands of tribesmen and the needs of an empire. Muʿāwiyah died in Damascus in 680.

BIBLIOGRAPHY. The only detailed study of Muʿāwiyah in a Western language is HENRI LAMMENS, "Études sur le règne du Calife Omaiyade Moʿāwia I," *Mélanges de la Faculté Orientale de l'Université St. Joseph de Beyrouth*, 1:3–108 (1906), and 2:1–172 (1907). This work is summarized in Lammens' article, "Muʿāwiya," in the *Encyclopaedia of Islam*, 1st ed., vol. 3 (1936).

(D.P.L.)

Muḥammad

The founder of the religion of Islām and of the Arab Empire, Muḥammad initiated religious, social, and cultural developments of monumental significance in the history of mankind. The full form of his name is Abū al-Qāsim Muḥammad ibn ʿAbd Allāh ibn ʿAbd al-Muṭṭalib ibn Hāshim.

LIFE AND DEEDS

Early life. Born in Mecca *c.* 570 after the death of his father, ʿAbd Allāh, Muḥammad was at first under the care of his paternal grandfather, ʿAbd al-Muṭṭalib. Because the climate of Mecca was considered to be unhealthful, he was given as an infant to a wet nurse from a nomadic tribe and spent some time in the desert. At six he lost his mother, Āminah of the clan of Zuhra, and at eight his grandfather. Though his grandfather had been head of the prestigious Hāshem (Hāshim) clan and was prominent in Mecca politics, he was probably not the leading man in Mecca, as some sources suggest. Muḥammad came under the care of the new head of the clan, his uncle Abū Ṭālib, and is reputed to have accompanied him on trading journeys to Syria. About 595, on such a journey, he was in charge of the merchandise of a rich woman, Khadījah of the clan of Asad, and so impressed her that she offered marriage. She is said to have been about 40, but she bore Muḥammad at least two sons, who died young, and four daughters, of whom the best known was Fāṭimah, the wife of Muḥammad's cousin ʿAlī, who is regarded as Muḥammad's divinely ordained successor by the Shīʿah branch of Islām. Until after Khadījah's death in 619, Muḥammad took no other wife. The marriage was a turning point in Muḥammad's life. By Arab custom, minors did not inherit, and therefore Muḥammad had no share in the property of his father or grandfather; but by his marriage he obtained sufficient capital to engage in mercantile activity on a scale commensurate with his abilities.

Prophetic call and early religious activity. Muḥammad appears to have been of a reflective turn of mind and is said to have adopted the habit of occasionally spending nights in a hill cave near Mecca. The poverty and misfortunes of his early life doubtless made him aware of tensions in Meccan society. Mecca, inhabited by the tribe of Quraysh (Koreish), to which the Hāshim clan belonged, was a mercantile centre formed around a sanctuary, the Kaʿbah (Kaaba), which assured the safety of those who came to buy and sell at the fairs. In the later 6th century there was extensive trade by camel caravan between Yemen and Syria (Gaza and Damascus), bringing goods from India and Ethiopia to the Mediterranean; and the great merchants of Mecca had obtained monopoly control of this trade. Mecca was thus prosperous, but most of the wealth was in a few hands. Tribal solidarity was breaking up; merchants pursued individual interests and disregarded their traditional duties to the unfortunate. About 610, as he reflected on such matters, Muḥammad had a vision of a majestic being (later identified with the angel Gabriel) and heard a voice saying to him, "You are the Messenger of God." This marked the beginning of his career as messenger (or apostle) of God (*rasūl Allāh*), or Prophet (*nabī*). From this time, at frequent intervals until his death, he received "revelations"—that is, verbal messages that he believed came directly from God. Sometimes these were kept in memory by Muḥammad and his followers, and sometimes they were written down. About 650 they were collected and written in the Qur'ān (or Koran, the sacred scriptures of Islām), in the form that has endured till today. Muslims believe the Qur'ān is written in the very words of God himself, and sound scholarship requires that Muḥammad be adjudged sincere in his view that the Qur'ān was not the product of his conscious mind.

Muḥammad's vision

Muḥammad is said to have been perturbed after the vision and first revelation but to have been reassured by his wife, Khadījah. In his later experiences of receiving messages there was normally no vision. Occasionally there were physical concomitants, such as perspiring on a cold day, and these gave rise to the suggestion, now agreed to be unwarranted, that he was an epileptic. Sometimes he heard a noise like a bell but apparently never a voice. The essence of such an experience was that he found a verbal message in his heart—that is, in his conscious mind. With the help of Khadījah's Christian cousin Waraqah, he came to interpret these messages as in general identical with those sent by God through other prophets or messengers to Jews, Christians, and others and to believe that by the first great vision and by the receipt of the messages he was commissioned to communicate them to his fellow citizens and other Arabs. The proclamation of the messages must have been accompanied by explanation and exposition in his own words.

Soon he gathered some sympathetic friends who accepted his claim to be a prophet and joined him in common worship or prayers. These culminated in an act of prostration in which they touched the ground with their foreheads in acknowledgment of God's majesty—still a cardinal act in Islāmic worship. In about 613 he began preaching publicly, and he and his followers spent their days together in the house of a young man named al-Arqam. It is probable that they sometimes worshipped together in the Kaʿbah.

Pagan religious milieu

The people of Mecca at the time nominally worshipped many gods, but few believed that man was dependent on supernatural powers. The merchants thought most things could be accomplished by wealth and by human planning. Some men regarded Allāh as a "high god" who stood above lesser deities. (Allāh, the Arabic word for God, is used by Christian Arabs as well as Muslims.) The earliest passages of the Qur'ān revealed to Muḥammad emphasize the goodness and power of God as seen in nature and in the prosperity of the Meccans and call on the latter to be grateful and to worship "the Lord of the Kaʿbah," who is thus identified with God. Gratitude is to be expressed in generosity with one's wealth and avoidance of niggardliness. As a sanction, men are warned that they will appear before God on the Last Day to be judged according to their deeds and assigned to heaven or hell.

By proclaiming this message publicly, Muḥammad gained followers—39, it is said—before he entered the house of al-Arqam. The names of 70 followers are known prior to the appearance of opposition to the new religion, and there were probably more. Most were young men under 30 when they joined Muḥammad. They included sons and brothers of the richest men in Mecca and thus were by no means plebeians, though they might be described as persons excluded from the most lucrative forms of commerce. A handful of Muḥammad's early followers are spoken of as "weak," which merely means that they were not of the tribe of Quraysh and so not effectively protected by any clan. The new religion was eventually called Islām—*i.e.*, "'submission [to God]"—and its adherents were called Muslims—*i.e.*, "those submitting [to God]"—though the Qur'ān speaks of them primarily as "the believers."

Opposition at Mecca. Although Muḥammad's preaching was basically religious, there was implicit in it a critique of the conduct and attitudes of the rich merchants of Mecca. Attempts were made to get him to soften his criticism by offering him a fuller share in trade and a marriage alliance with one of the wealthiest families, but he decisively rejected such offers. About 615 more active opposition appeared. Points in the message of the Qur'ān were questioned, such as the assertion that men would be resurrected before the Judgment. Commercial pressure was brought to bear on Muḥammad's supporters, and in some families there was mild persecution of junior members who followed him. It is sometimes suggested that the main reason for opposition was economic—namely, the merchants' fear that the new religion would destroy the recognition of the Kaʿbah as a sanctuary—but this is unlikely. Certainly attacks on idols appeared in the Qur'ān, and Islām began to be characterized by the insistence that "there is no deity but God" (Allāh), but no attack was made on the Kaʿbah, and the idols mentioned had their chief shrines elsewhere. A leader of the opposition arose in the person of Abū Jahl, a contemporary of Muḥammad, who probably felt that the latter, despite his claim to be "only a warner" (of Judgment to come), was building up for himself a position of authority that might one day make him politically supreme in Mecca, because Arabs deeply respected the kind of wisdom or knowledge that Muḥammad clearly had. In about 616 Abū Jahl organized a boycott of the clan of Hāshim by the chief

Commercial pressure, boycott, and persecution

clans of Mecca, allegedly because the clan continued to protect Muḥammad and did not curb his preaching; but, since few of the clan were Muslims, other questions may have been involved. After three years the boycott petered out, perhaps because some of the participants found they were harming their own economic interests.

Both Muḥammad's wife Khadījah and his uncle Abū Ṭālib died in about 619, and another uncle, Abū Lahab, succeeded as head of the clan of Hāshim. He was closer to the richest merchants, and at their instigation he withdrew the protection of the clan from Muḥammad. This meant that Muḥammad could easily be attacked and therefore could no longer propagate his religion in Mecca. He left for the neighbouring town of aṭ-Ṭā'if, but the inhabitants were insufficiently prepared to receive his message, and he failed to find the support he hoped for. Before returning to Mecca, he had to beg for the protection of the head of another clan; it was granted, but conditions were presumably attached. In 620 Muḥammad began negotiations with clans in Medina, leading to his emigration, or *hijrah*, there in 622.

It is difficult to assess the nature and extent of the persecution of the Muslims in Mecca. There was little physical violence, and that almost always within the family. Muḥammad suffered from minor annoyances, such as having filth deposited outside his door. The persecution is said to have led to the emigration of some of the Muslims to Ethiopia about 615, but they may have been seeking opportunities for trade or military support for Muḥammad. Some remained until 628, long after Muḥammad was established in Medina. Whether or not the persecution was unbearable, the Muslims were very bitter about it.

The emigration from Mecca to Medina. In the summer of 621, 12 men from Medina, visiting Mecca for the annual pilgrimage to the Ka'bah (then a pagan shrine), secretly professed themselves Muslims to Muḥammad and went back to make propaganda for him at Medina. At the pilgrimage in June 622 a representative party of 75 persons from Medina, including two women, not merely professed Islām but also took an oath to defend Muḥammad as they would their own kin. These are known as the two Pledges of al-'Aqaba. Muḥammad now encouraged his faithful Meccan followers to make their way to Medina in small groups, and about 70 emigrated thus. The Meccans are said to have plotted to kill Muḥammad before he could leave. With his chief lieutenant he slipped away unperceived, used unfrequented paths, and reached Medina safely on September 24, 622. This is the famous *hijrah* (Latin Hegira), which may be rendered "emigration," though the basic idea is the severing of kinship ties. It is the traditional starting point of Islāmic history. The Islāmic Era (AH or Anno Hegirae) begins on the first day of the Arabic year in which the *hijrah* took place; viz., July 16, 622, in the Western calendar.

Medina was different from Mecca. It was an oasis in which date palms flourished and cereals could be grown. Agriculture had been developed by several Jewish clans, who had settled among the original Arabs, and they still had the best lands. Later Arab immigrants belonging to the tribes of al-Aws and al-Khazraj, however, were in a stronger position. The effective units among the Arabs were eight or more clans, but nearly all of these had become involved in serious feuds. Much blood had been shed in a battle at about 618, and peace was not fully restored. In inviting Muḥammad to Medina, many of the Arabs there probably hoped that he would act as an arbiter among the opposing parties, and their contact with the Jews may have prepared them for a messianic religious leader, who would deliver them from oppression and establish a kingdom in which justice prevailed.

The Constitution of Medina

A document has been preserved known as the Constitution of Medina. In its present form it is a combination of at least two earlier documents and is probably later than 627, but its main provisions are almost certainly those originally agreed upon between Muḥammad and the Muslims of Medina. In form the document creates a confederation on traditional Arab lines among nine groups—eight Arab clans and the emigrants from Mecca. Muḥammad is given no special position of authority, except that the preamble speaks of the agreement as made between "Muḥammad the prophet" and the Muslims now resident in Medina, and it is stated that serious disputes are to be referred to him. The Jewish groups had refused to acknowledge Muḥammad as prophet and in the document appear in a secondary character as attached to various Arab clans. For at least five years, Muḥammad had no direct authority over members of other clans, but, in the closing years of his life, the prestige of his military successes gave him almost autocratic power. The revelations he received at Medina often contained legal rules for the community of Muslims, but they dealt with political questions only rarely.

The first five years at Medina. The first 18 months at Medina were spent in settling down. Muḥammad was given a piece of land and had a house built, which eventually held apartments grouped around a central courtyard for each of his wives. The Muslims often joined Muḥammad at prayers in his home, which, after his death, became the mosque of Medina. The emigrants (*muhājirūn*, the men from Mecca) were at first guests of brother Muslims in Medina, but Muḥammad cannot have contemplated this situation continuing indefinitely. A few emigrants carried on trade in the local market run by a Jewish clan. Others, with the approval of Muḥammad, set out in normal Arab fashion on razzias (*ghazawāt*, "raids") in the hope of intercepting Meccan caravans passing near Medina on their way to Syria. Muḥammad himself led three such razzias in 623. They all failed, probably because traitors betrayed the Muslim movements to the enemy. At last, in January 624, a small band of men was sent eastward with sealed orders telling them to proceed to Nakhlah, near Mecca, and attack a caravan from Yemen. This they did successfully, and in doing so they violated pagan ideas of sanctity—thereby making the Meccans aware of the seriousness of the threat from Muḥammad.

About the same time Muḥammad changed his general policy in important respects. One was his "break with the Jews"; instead of making concessions to the Jews in the hope of gaining recognition of his prophethood, he asserted the specifically Arabian character of the Islāmic religion. Hitherto the Muslims had faced Jerusalem in prayer, but a revelation now bade them face Mecca. Perhaps because of this change some Muslims of Medina were readier to support Muḥammad. In March 624 he was able to lead about 315 men on a razzia to attack a wealthy Meccan caravan returning from Syria. The caravan, led by Abū Sufyān, the head of the Umayyah clan, eluded the Muslims by devious routes and forced marches. Abū Jahl, the head of the Makhzūm clan, however, leading a supporting force of perhaps 800 men, wanted to teach Muḥammad a lesson and did not withdraw. On March 15, 624, near a place called Badr, the two forces found themselves in a situation, perhaps contrived by Muḥammad, from which neither could withdraw without disgrace. In the ensuing battle at least 45 Meccans were killed, including Abū Jahl and other leading men, and nearly 70 taken prisoner, while the Muslims lost 14 killed. To Muḥammad this appeared to be a divine vindication of his prophethood, and he and all the Muslims were greatly elated.

The Battle of Badr and its consequences

In the flush of victory some persons in Medina who had satirized Muḥammad in verse were assassinated, perhaps with his connivance. He also made a minor disturbance an excuse for expelling the Jewish clan, which ran the market. This weakened his most serious opponent there, the "hypocrite" (*munāfiq*), or nominal Muslim, 'Abd Allāh ibn Ubayy, whose allies these Jews were. The remaining waverers among the Arabs probably became Muslims about this time. Thus the victory of Badr greatly strengthened Muḥammad. At the same time he was using marriage relationships to bring greater cohesion to the emigrants. Of his daughters, Fāṭimah was married to 'Alī (later fourth caliph, or leader of the Islāmic community) and Umm Kulthūm to 'Uthmān (third caliph). He himself was already married to 'Ā'ishah, daughter of Abū Bakr (first caliph), and was now

espoused also to Ḥafṣah, daughter of ʿUmar (second caliph), whose previous husband was one of the Muslims killed at Badr.

In the same year Muḥammad led larger Muslim forces on razzias against hostile nomadic tribes and had some success. Presumably, he realized that the Meccans were bound to try to avenge their defeat. Indeed, Abū Sufyān was energetically mobilizing Meccan power. On March 21, 625, he entered the oasis of Medina with 3,000 men. One of the features of Medina was a large number of small forts that were impregnable to Arab weapons and tactics. Muḥammad would have preferred the Muslims to retire to these; but those whose cereal crops were being laid waste persuaded him to go out to fight. By a night march with 1,000 men, he reached the hill of Uḥud on the further side of the Meccan camp. On the morning of March 23 the Meccan infantry attacked and was repulsed with considerable loss. As the Muslims pursued, the Meccan cavalry launched a flank attack after the archers guarding the Muslim left had abandoned their position. The Muslims were thrown into confusion. Some made for a fort and were cut down, but Muḥammad and the bulk of his force managed to gain the lower slopes of Uḥud, where they were safe from the cavalry. The Meccans, because of their losses, were unable to press home their advantages and without delay set out for home, while Muḥammad the next day made a show of pursuing. The battle produced neither a clear victor nor loser. In Badr and Uḥud together, the Meccans had killed about as many men as they had lost; but they had boasted that they would make the Muslims pay several times over, and they had not shown the degree of superiority appropriate to their leading position in Arabia. Muḥammad, though he had lost above 70 men, realized that this was a military reverse, not a defeat; but the confidence of the Muslims and perhaps his own had been struck a serious blow. If the victory of Badr was a sign of God's support, did Uḥud indicate that he had abandoned the Muslims? Muḥammad's faith soon overcame any momentary doubts, and he was gradually able to restore the confidence of his followers.

The Battle of Uḥud

For two years after Uḥud, both sides prepared for a decisive encounter. In the razzias Muḥammad led or sanctioned, he seems to have aimed at extending his own alliances and at preventing others from joining the Meccans. In at least two cases a small party of Muslims was tricked or ambushed, and most of their lives were lost. And another Jewish clan was expelled from Medina. At length, in April 627 Abū Sufyān led a great confederacy of 10,000 men against Medina. On this occasion Muḥammad had ordered the crops to be harvested and a trench to be dug to defend the main part of the oasis from the Meccan cavalry. For a fortnight the confederates besieged the Muslims. Attempts to cross the trench failed, and fodder for the horses was scarce, while Muḥammad's agents among the attackers fomented potential dissensions. Then, after a night of wind and rain the great army melted away. The Meccans had exerted their utmost might and had failed to dislodge Muḥammad, whose position was now greatly strengthened.

The siege of Medina

For more than two years now there had been opposition to Muḥammad in Medina, chiefly from ʿAbd Allāh ibn Ubayy and other so-called hypocrites (*munāfiquin*) who had abandoned Muḥammad at Uḥud and who together had fostered disaffection. Shortly before the siege Muḥammad had a showdown with ʿAbd Allāh ibn Ubayy, who had joined in spreading slanders about Muḥammad's wife ʿĀ'ishah. This confrontation revealed that ʿAbd Allāh had little support in Medina, and he became reconciled to Muḥammad. After the siege of Medina, Muḥammad attacked the Jewish clan of Qurayẓah, which had probably been intriguing against him. When they surrendered, the men were all savagely executed and the women and children sold as slaves.

The winning of the Meccans. Muḥammad's farsightedness as a statesman is manifest in the policies he next adopted. He might have proceeded to crush the Meccans, and he indeed put economic pressure on them; but his main aim was to gain their willing adherence to Islām. He had already realized that, insofar as the Arabs became Muslims, it would be necessary to direct outward the energies expended on razzias against one another. There could be no question of Muslims raiding Muslims. It is noteworthy that his largest razzias, apart from the expeditions against the Meccans, were along the route to Syria followed by the Arab armies after his death (see ISLAM, HISTORY OF). He doubtless realized that the administrative skill of the Meccan merchants would be required for any expansion of his embryonic state.

In a dream Muḥammad saw himself performing the annual pilgrimage to Mecca, and in March 628 he set out to do so, driving sacrificial animals; but he was disappointed because no more than 1,600 men would accompany him. The Meccans were determined to prevent the Muslims from entering their town, so Muḥammad halted at al-Ḥudaybiyah, on the edge of the sacred territory of Mecca. After some critical days the Meccans made a treaty with Muḥammad. Hostilities were to cease, and the Muslims were to be allowed to make the pilgrimage to Mecca in 629. The orderly withdrawal showed how completely Muḥammad controlled his followers. Partly to reward this orderly conduct, Muḥammad two months later led the same force against the Jewish oasis of Khaybar, north of Medina. After a siege it submitted, but the Jews were allowed to remain on condition of sending half of the date harvest to Medina. Thus throughout 628 and 629 Muḥammad's power was growing, since success led more men to become Muslims, for the religious attraction of Islām was apparently supplemented by material motives.

The Treaty of al-Ḥudaybiyah

Meanwhile Mecca was in decline. Several leading men had emigrated to Medina and become Muslims. New leaders had taken over from Abū Sufyān but had accomplished little, although the treaty with Muḥammad had removed his pressure on their caravans. Shortly after the treaty, Muḥammad had married Umm Ḥabībah, a daughter of Abū Sufyān and a widow whose Muslim husband had died in Ethiopia. This led to an understanding with Abū Sufyān, who began to work for the peaceful surrender of Mecca. It was probably when he was in Mecca for the pilgrimage in March 629 that Muḥammad became reconciled with another uncle, al-ʿAbbās, and married his uncle's sister-in-law Maymūnah.

An attack by Meccan allies in about November 629 upon allies of Muḥammad led to the latter's denunciation of the treaty of al-Ḥudaybiyah. After secret preparations he marched on Mecca in January 630 with 10,000 men. Abū Sufyān and other leading Meccans went out to meet him and formally submitted, and Muḥammad promised a general amnesty. When he entered Mecca there was virtually no resistance. Two Muslims and 28 of the enemy were killed. A score of persons were specifically excluded from the amnesty, but some were later pardoned. Thus Muḥammad, who had left Mecca as a persecuted prophet, not merely entered it again in triumph but also gained the allegiance of most of the Meccans. Though he did not insist on their becoming Muslims, many soon did so.

The victorious entrance into Mecca

Muḥammad spent 15 to 20 days in Mecca settling various matters of administration. Idols were destroyed in the Kaʿbah and in some small shrines in the neighbourhood. To relieve the poorest among his followers, he demanded loans from some of the wealthy Meccans. When he marched east to meet a new threat, 2,000 Meccans went with him.

The closing years: the unification of Arabia. Ever since the *hijrah*, Muḥammad had been forming alliances with nomadic tribes. At first these were probably nonaggression pacts, but, when he was strong enough to offer protection, he made it a condition of alliance that the tribe should become Muslim. While in Mecca Muḥammad had word of a large concentration of hostile nomads, and he set out to confront them. A battle took place at Ḥunayn in which part of Muḥammad's army was put to flight, but he himself and some older Muslims stood firm. The enemy was finally routed, and their dependents and possessions were all captured. They were allowed to ransom wives and children, but their livestock was divided as booty.

Muḥammad was now militarily the strongest man in Arabia. Most tribes sent deputations to Medina seeking alliance. It is difficult to say how much of Arabia joined his alliance, for the inner politics of each tribe were complex, and in some cases the deputation might represent only a small section. Muḥammad benefitted from the defeat of the Persian Empire by the Byzantine (Christian) Empire (627–628), for, in the Yemen and in places on the Persian Gulf, minorities that had relied on Persian support against Byzantium now turned to Muḥammad instead.

March to the Syrian border. The greatest of all of Muḥammad's razzias occurred at the end of 630, when he took 30,000 men on a month's journey to the Syrian border. In this campaign he pioneered the invasion of Syria and made agreements that became models for treaty arrangements with captured peoples. Some of the tribes near Syria were Christian and adhered to the Byzantines; chiefly as a result of this, Muḥammad's earlier friendship for the Christians, notably those of Ethiopia, changed to hostility. Before his death, armed opposition to him appeared in one or two parts of Arabia, but the Islāmic state was strong enough to deal with this. Thus he left most of Arabia united and poised for expansion into Syria and Iraq.

Muḥammad personally led the pilgrimage to Mecca, in March 632, in a form according with Islāmic belief. But his health was poor. About the beginning of June, he ceased attending to business, and on June 8 he died. No arrangements had been made for the succession, but Abū Bakr had been nominated to lead the prayers, and a meeting of all the Muslims in Medina agreed that he should be caliph (*khalīfah*), or successor.

CHARACTER AND ACHIEVEMENTS

Few great men have been so maligned as Muḥammad. Christian scholars of medieval Europe painted him as an imposter, a lecher, and a man of blood. A corruption of his name, Mahound, even came to signify the devil. This picture of Muḥammad and his religion still retains some influence. The English author Thomas Carlyle in 1840 was the first notable European to insist publicly that Muḥammad must have been sincere, because it was ridiculous to suppose an impostor would have been the founder of a great religion. Even if Muslims have often idealized him, a mass of detail in the early sources shows that he was an honest and upright man who had gained the respect and loyalty of others who were likewise honest and upright men. His other alleged faults must be considered in the light of his time and environment. He was not condemned for them by his contemporaries, except where he abandoned practices connected with idolatry. In several respects he was a social reformer.

The background of Muḥammad's marital arrangements is the prevalence in Arabia of a matrilineal kinship system (in which descent, status, and inheritance are traced through the mother's side), which led to the neglect of paternity and permitted forms of polyandry (having many husbands at once) sometimes verging on promiscuity. In permitting a Muslim to have four wives, the Qur'ān did not restrict an unlimited polygamy but instituted a new system based on patrilineal kinship, to which many Arabs were apparently turning. A woman could have only one husband at a time, or, if divorced, she had to wait over three months before remarrying, so that there could be no doubt about the paternity of any child. Muḥammad himself left nine wives and a Christian concubine. For each of his marriages there was a political reason, the forming of closer links with the father or relatives of the woman. Some contemporaries were scandalized at Muḥammad's marriage to Zaynab, divorced wife of his "adopted son" Zayd ibn Ḥārithah. By pagan custom this was incestuous; but the relationship was fictitious, and Muḥammad was apparently attacking the pagan system of relationships. In this sphere, then, Muḥammad was not satisfying his lusts but acting as a reformer.

Muḥammad's connivance at assassinations and his approval of the execution of the men of a Jewish clan for treachery are matters that cannot be denied. This was the style of life of the times, and two or three unsuccessful attempts on Muḥammad's life are recorded. Internal security in Arabia still depended on the law of retaliation, a life for a life, but Muḥammad tried to mitigate this in practice by encouraging men to accept from fellow Muslims a blood penalty of 100 camels in place of a life, for the murder of a member of their family or clan.

Muḥammad was admired by his contemporaries for his courage, resoluteness, and impartiality, and for a firmness that was tempered by generosity. He won men's hearts by his personal charm; his pleasant smile is remarked on. He was gentle, especially with children. Though he was sometimes silent in thought, for the most part he was engaged in purposeful activity. He walked vigorously and spoke rapidly. He became for later Muslims an exemplar of virtuous character, and stories presented him as realizing the Islāmic ideal of human life.

Muḥammad's chief significance is as founder of a state and of a religion. In his lifetime he created a federation of Arab tribes, which, in less than 20 years after his death, defeated the Byzantine and Persian empires, occupied a vast territory from Libya to Persia, and then developed into the Arab, or Islāmic, Empire. This process was favoured by the social tensions in Mecca and Medina and by the exhaustion of the two older empires after a long war; but Arab energies would not have achieved what they did had it not been for Muḥammad's gifts in handling men and his farsightedness as a statesman.

He made the religion of Islām, which now has some 500,000,000 adherents of many different races, the basis of Arab unity. In 13 centuries Islām has produced great saints, and there has been a high level of religious devotion and moral performance among ordinary Muslims. Islāmic doctrine maintains that God is the founder of the religion, not Muḥammad, but the latter played an important part in fostering the nascent religion. His concern with ultimate questions, his mystical outlook, and his moral seriousness were important adjuncts to the preaching of the Qur'ānic message.

BIBLIOGRAPHY

Biographical works: W. MONTGOMERY WATT, *Muhammad at Mecca* (1953) and *Muhammad at Medina* (1956), the most recent full-scale treatment, summarized in *Muhammad: Prophet and Statesman* (1961), by the same author; FRANTS BUHL, *Das Leben Mohammeds* (1930), a German translation of a work published in Danish in 1903 that is still considered reliable; TOR ANDRAE, *Mohammed, sein Leben und Glaube* (1932; Eng. trans., *Mohammed: The Man and His Faith*, 1936), chiefly concerned with religious aspects; A. GUILLAUME, *The Life of Muhammad: A Translation of Isḥāq's Sīrat rasūl Allāh* (1955), the primary Arabic biography; NABIA ABBOTT, *Aishah: The Beloved of Mohammed* (1942), a scholarly work; JOHN BAGOT GLUBB, *The Life and Times of Muhammad* (1970), a popular account based on the author's familiarity with Arab life.

Primary sources: The Qur'ān, of course, contains basic contemporary materials on Muḥammad, but it is difficult to assess them without broader historical knowledge. The vast collections of Traditions (Ḥadīth), or anecdotes about Muḥammad's words and deeds, are historically disputable and, besides, seldom tell us anything significant about Muḥammad's career. The main sources of historical value are the early biographies (8th–9th century), especially the *Sīrah* of IBN ISHAQ, as edited by IBN HISHAM, and the *Maghāzi* ("Expeditions") of AL-WAQIDI, together with the supplementary materials recorded by his associate IBN SAʿD. The last item contains much material about the Companions (persons in contact with Muḥammad) and thus about the Prophet's relation to and work with them. Some contemporary documents are preserved in the early biographical works, the most important being the so-called "Constitution of Medina." The latter is in Guillaume's translation of Ibn Isḥāq cited above; other documents are in Watt's *Muhammad at Medina.*

(W.M.W.)

Muḥammad ʿAlī Pasha

The founder of the dynasty that ruled Egypt from the beginning of the 19th century until the middle of the 20th century, Muḥammad ʿAlī helped to create the conditions that made possible the emergence of a modern national state in Egypt.

Muhammad ʿAlī Pasha, lithograph by an unknown artist.
By courtesy of the trustees of the British Museum; photograph, J.R. Freeman and Co. Ltd.

He was born in 1769 in Kavala (Kaválla), a small Macedonian seaport on the coast of the Aegean Sea in what is now Greece; his ethnic background is unknown, though he was certainly a Muslim and an Ottoman subject. Since he later accompanied a small Albanian (Arnavut) contingent to Egypt, he may have been an Albanian. His father, Ibrahim Agha, the commander of a small provincial force maintained by the governor of Kavala, died when Muḥammad ʿAlī was a boy, and he was brought up by the governor. At 18 he was married to one of the governor's relatives, who became the mother of five of Muḥammad ʿAlī's 95 children. He became involved in the tobacco trade, an experience that may account for his later commercial interests.

In 1798, Egypt, at that time a semi-autonomous province of the Ottoman Empire, was occupied by a French force under Napoleon Bonaparte. Muḥammad ʿAlī went there as part of an Ottoman expeditionary force to oppose the French. With great political skill, he managed by 1805 to be named the *wālī*, the Ottoman sultan's viceroy in Egypt, with the rank of pasha.

Appointment as Egypt's viceroy

Nowhere in the Ottoman Empire was there greater opportunity for a total restructuring of society than in Egypt. The three-year French occupation (1798–1801) had disrupted the country's traditional political and economic structure. Continuing the task begun by the French, Muḥammad ʿAlī put an end to Egypt's traditional society. He eliminated the Mamlūks, the former ruling oligarchy; he expropriated the old landholding classes; he turned the religious class into pensioners of the government; he restricted the activities of the native merchant and artisan groups; he neutralized the Bedouins; and he crushed all movements of rebellion among the peasants. The task of rebuilding Egypt along modern lines now lay before him.

But, though Muḥammad ʿAlī had considerable native intelligence and great personal charm, he was a man of limited knowledge and of narrow horizons. He proved insensitive to the possibilities now open to him and governed generally according to Ottoman principles. No group within Egyptian society was capable of forcing fundamental changes upon him; elements that might have served as the instruments of change had been crushed at the outset of his regime. Neither was there an ideology capable of bringing together the ruler and the ruled in a great national effort. The nationalism that swept Europe in this period meant little to Muḥammad ʿAlī and nothing to Egyptians. Finally, Muḥammad ʿAlī had to devote much of his effort to resisting attempts by his Ottoman overlord to remove him from office. His policies were designed more to entrench himself and his family in Egypt as its hereditary rulers than to create a new society.

Reform of Egypt's administration and economy

To strengthen his position within Egypt and to increase his revenues, Muḥammad ʿAlī instituted sweeping changes. By 1815 most of Egypt's agricultural land had been converted into state land and profits from agriculture became available to the ruler. He improved Egypt's irrigation system, on which its agriculture depended; he introduced new crops, such as cotton, which promised high cash returns; he reorganized the administrative structure of the government to insure strict control of the economy. He also attempted to construct a modern industrial system to process Egypt's raw materials. Disbanding his mercenary army, he created a fleet and an army of Egyptians conscripted from the peasant class but commanded by Turks and others recruited from outside Egypt. To supply services for his armed forces, he created Western-style schools to train doctors, engineers, veterinarians, and other specialists. He also began sending educational missions to European countries for training in modern techniques.

His industrial experiments failed, largely because Egypt lacked sources of power, a native managerial class, and a trained working class. Even the agricultural sector declined ultimately because of administrative mismanagement, excessive taxation, military conscription of the peasantry, and his monopolization of trade. By the mid-1830s Muḥammad ʿAlī's policy of turning Egypt into a massive plantation for his own benefit had reached a point of diminishing returns. Furthermore, his financial requirements had greatly increased because of his military campaigns.

These campaigns fell into three categories. In the first half of his reign, he supported the Ottoman sultan in suppressing rebellion both in Arabia and in Greece, and also invaded the Nilotic Sudan in search of recruits for his army and gold for his treasury. Victorious in all three campaigns, until European intervention in Greece caused the destruction of his fleet at Navarino in 1827, Muḥammad ʿAlī felt that he was strong enough to challenge the sultan. His first war against the sultan (1831–33) gained him control of Syria as far north as Adana. In the second war (1838–41) the decisive defeat of Ottoman troops at the Battle of Nezib (June 24, 1839) and the desertion of the Ottoman fleet to Muḥammad ʿAlī led to intervention by the European powers. In July 1840 Great Britain, Russia, Austria, and Prussia agreed to end Egyptian rule in Syria, shattering Muḥammad ʿAlī's hopes for greater independence from the Ottoman Empire. In 1841 he and his family were granted the hereditary right to rule Egypt and the Sudan, but his power was still subjected to restraints, and the sultan's suzerain rights remained intact.

Attempts at expansion

The final years of Muḥammad ʿAlī's reign were uneventful. He retired from office in 1848 and died in Alexandria on August 2, 1849. Many of his reforms were abandoned after his death, having been too closely allied to his policies of self aggrandizement to survive the period of expansion. But Muḥammad ʿAlī cleared the path for the creation of an independent Egyptian state; that his successors failed to meet the challenge of the modern world may be traced in part to the inadequacies of his rule, but their failures remain largely their own.

BIBLIOGRAPHY. HENRY DODWELL, *The Founder of Modern Egypt* (1931, reprinted 1967), is the only work in English to survey the reign of Muḥammad ʿAlī in all its aspects. For a more critical view of the Pasha's achievements, see HELEN ANNE B. RIVLIN, *The Agricultural Policy of Muḥammad ʿAlī in Egypt* (1961), which also contains an extensive annotated bibliography.

(H.A.B.R.)

Muḥammad ibn Tughluq

Muḥammad ibn Tughluq extended the Delhi sultanate, the principal Muslim kingdom in north India in the late Middle Ages, to its farthest reach, and then, after exercising hegemony over the greater part of the subcontinent, as a result of misguided administrative actions and unexampled severity toward his opponents, lost his authority in the south. At the end of his reign, the sultanate had begun to decline in power.

Muḥammad was born about 1290, the son of the sultan Ghiyās-ud-Dīn Tughluq. Very little is known of his childhood, but he apparently received a good education. He possessed an encyclopaedic knowledge of the Qur'ān, Muslim jurisprudence, astronomy, logic, philosophy, medicine, and rhetoric. In 1321–22 his father sent him

Early years of his reign

against the city of Warangal in the Deccan, in which campaign, after initial reverses, he subdued the rebellious Hindu rajas. From his accession to the throne in 1325 until his death in 1351, Muḥammad contended with 22 rebellions, pursuing his policies consistently though ruthlessly. Ziyā'-ud-Dīn Baranī, his close companion and counsellor for 17 years, often advised him to abdicate, but Muḥammad disdainfully rejected his advice.

As his reign began, he attempted, without much success, to enlist the services of the *'ulamā'*, the Muslim divines, and the Ṣūfīs, the ascetic mystics. Failing to win the *'ulamā'* over, he tried to curtail their powers, as some of his predecessors had, by placing them on an equal footing with other citizens. The Sultan wanted to use the Ṣūfīs' prestigious position to stabilize his authority as ruler. Yet they had always refused any association with government and would not accept any grants or offices except under duress. Muḥammad tried every measure, conciliatory or coercive, to yoke them to his political wagon. Although he humiliated them, he could not break their opposition and succeeded only in dispersing them from the towns of northern India.

In the four pages of his so-called autobiography, Muḥammad's only surviving literary work, he confesses that he had wavered from traditional orthodoxy to philosophic doubts and then found his way to a rational faith. To still his own doubts, as well as to counteract the opposition of the Muslim divines, he obtained from the caliph in Cairo a *manshūr* (patent of royalty) legitimizing his authority.

Move to the south

The transfer of the capital in 1327 to Deogir (now Daulatābād) was intended to consolidate the conquests in southern India by large-scale, in some cases forced, migration of the people of Delhi to Deogir. As an administrative measure it failed, but it had far-reaching cultural effects. The spread of the Urdu language in the Deccan may be traced to this extensive influx of Muslims. He introduced several reforms in the monetary system, and his coins, in design as well as in workmanship and purity of metal, excelled those of his predecessors. His introduction of token currency, coins of baser metal with the face value of silver coins, however, failed dismally. A projected Khorāsān expedition (1327–28) that never materialized was intended to secure more defensible frontiers in the west. The Karajil (Garhwal-Kumaon) expedition (1329–30), an attempt to adjust the boundary dispute with the northern hill states then dominated by China, ended in disaster, but was followed by an exchange of emissaries between China and Delhi. The conquest of Nagarkot in the foothills of the Himalayas in northwest India was based on Muḥammad's policy of establishing secure frontiers.

Between 1328–29 the Sultan increased the land tax in the Doab—the land between the Ganges and Jumna rivers—but the taxpayers resisted it, especially because a severe drought coincided. Muḥammad was the first ruler to introduce rotation of crops, establish state farms, and extend cultivation and improve artificial irrigation by establishing a department of agriculture. When famine broke out in northern India (1338–40), he moved his residence to Swargdawari to supervise famine relief measures himself.

Muḥammad's last expedition against the rebel Ṭaghī ended with his death at Sondaha in Sind in 1351. He died with a smile on his face and verses of his own composition on his lips. In the words of a contemporary, "the Sultan was rid of the people and the people of the Sultan."

Assessment

Sultan Muḥammad was among the most controversial and enigmatic figures of the 14th century. A dauntless soldier, he was tolerant in religion, and was normally humane and humble, but these traits were vitiated at times by cruelty sometimes approaching the inhuman. He lived in constant conflict between faith and action, faith in the correctness of his policies and in the means by which he sought to implement them. A born revolutionary, he desired to create a more equitable social order by making Islām a religion of service rather than a means of exploitation. This end, he believed, could be achieved only by a strong centralized authority based on justice and patronage of the poor, the learned, and the pious and on the suppression of rebellions mainly of the privileged classes in a tradition-ridden society.

All contemporary historians based their assessment of Muḥammad on his administrative measures, which were neither vicious nor visionary. They failed because of the harshness of the Sultan in executing them, the challenge they posed to the privileged classes, the general lethargy and conservatism of his subjects, and the expansion of the empire with which Muḥammad's administrative machinery could not cope.

BIBLIOGRAPHY. No official history of the reign of Muḥammad ibn Tughluq has come down to us. All contemporary authorities speak of the Sultan as an elegant writer. The only literary work that has survived is a fragment (four pages) of what is considered by some writers as a fragment of his autobiography and by others as part of a letter the Sultan wrote to the Caliph at Cairo. Works by three of his most important contemporaries are: ZIA-UD-DIN BARANI, *Tārikh-i-Fīruz Shāhī*, Persian text (1860–62; partly translated in *The History of India As Told by Its Own Historians*, ed. by JOHN DOWSON, vol. 3, 1871), which includes a full-length account of Muḥammad ibn Tughluq's reign; *'Isāmī, Futuhus-salatin*, ed. by A.S. USHA (1948), text in Persian, a versified history of the Sultan's reign; and IBN BATUTAH, *Travels in Asia and Africa, 1325–1354*, trans. by H.A.R. GIBB (1929). Two modern works are ISHWARI PRASAD, *History of Qaranah Turks* (1936); and MAHDI HUSAIN, *The Rise and Fall of Muḥammad ibn Tughluq* (1938).

(S.A.R.)

Muḥāsibī, al-

The eminent Muslim mystic (Ṣūfī) and theologian al-Muḥāsibī owes his reputation chiefly to his psychological refinement of pietistic devotion and to his role as a precursor of the doctrine of latter Muslim orthodoxy. His name means "he who examines his conscience."

Abū 'Abd Allāh al-Ḥārith ibn Asad al-Muḥāsibī al-'Anazī was born in Basra toward the end of the 8th century. There is little historical information about his life. His parents apparently left for Baghdad shortly after his birth, perhaps attracted by the many opportunities afforded by the newly founded capital. His father had acquired some wealth, but Muḥāsibī is said to have refused it because of doctrinal differences. He led a normal life, owned a beautiful house, and liked sumptuous clothes. This image of the ordinary bourgeois, however, is qualified by a trait that Muḥāsibī imported from Basra: the spirituality of otherworldliness propagated by the famous Ṣūfī theologian al-Ḥasan al-Baṣrī (died 728).

Critique of Muslim asceticism and people

Muslim asceticism had developed some specific features: nightly recitals of the Qur'ān (the Muslim sacred scriptures), restrictions concerning the kind and quantity of food one should eat, and a special attire consisting of woollen clothing. These habits had been adapted from the life-style of Christian monks. But whereas Christian monks used to live in seclusion, a Muslim ascetic felt obliged to remain an active member of his community. Thus, Muḥāsibī came to realize that in his urban society with its inevitable public exposure, the practice of outward asceticism was open to ambiguity: although it could serve to suppress the normal sins of passion, it could also become a deceptive vehicle for inner vices like hypocrisy and pride. As soon as outward piety becomes a part of a person's image, it can act as a screen for the hidden intentions of the ego. Man has to recognize that sinful actions are frequently defined not by their objective reality but by the subjective attitude of the sinner. Outside the realm of the commandments and prohibitions of the Qur'ān, nothing should be labelled good or bad without restriction. The most commendable attitude is scrupulosity, although even this can be ambiguous, because it might result in spiritual paralysis. Asceticism is valuable as something additional, an act of supererogation, but it must always be tempered by attention paid to the inner and outer duties toward God (*ar-rī 'āyah li-ḥūqūq Allāh*, the title of Muḥāsibī's main work). The proper instrument for this is reason, the importance of which Muḥāsibī stressed far beyond the normal practice of mystics,

who often tended to emphasize irrationality and spiritual intoxication. The method he proposed was *muḥāsabah*, the anticipation of the Last Judgment through constant self-examination. This seems to have been an impediment to real mystical experiences; the ruthlessness of this psychological technique buried every attempt at ecstatic exaltation under an enormous inferiority complex.

Muḥāsibī propagated his ideas in didactic conversations, which he would record immediately afterward; his books still preserve this dialogical structure. His influence on posterity was immense, especially through his pupil Junayd. During his lifetime, however, he was regarded with suspicion, and his last years were embittered by persecution. He had joined a group of theologians who, led by ʿAbd Allāh ibn Kullāb (died 855), criticized the doctrines of the rationalist Muʿtazilite school dominant at that time. The discussion was focussed on the problem of the essence of God and the nature of his attributes. The Muʿtazilites, in stressing the unity of God, tended to reduce the attributes to mere nominal aspects; Muḥāsibī, in order to preserve their individual value, accentuated much more their independent status. And whereas the Muʿtazilites held the attribute of God's speech to be created, realized in temporal revelations like that of the Qur'ān, Muḥāsibī believed that it was also uncreated if seen under the aspect of the eternal Word of God. He did not go so far as to support the popular belief that the Qur'ān, too, was uncreated; he avoided this shibboleth used in the inquisition initiated in favour of the Muʿtazilites by the caliph al-Ma'mūn in 833. This diplomatic attitude became precarious when, in 850–851, a later caliph, al-Mutawakkil, put an end to the pro-Muʿtazilite policy of his predecessors and, two years later, prohibited rationalist theology altogether. Muḥāsibī's theological position was now viewed as treasonous by the former victims of the inquisition, precisely because he had been closest to them in their dogmatic outlook, for they considered the use of any rational theological method as heresy, regardless of the doctrine it supported. He was consequently forced to give up his public teaching and appears to have emigrated to Kūfah. Later on he was allowed to return to Baghdad, perhaps at the price of abandoning his theological convictions. Yet the boycott persisted: when he died there in 857, only four people attended his funeral.

Theological disputes and persecution

BIBLIOGRAPHY. MARGARET SMITH, *An Early Mystic of Baghdad* (1935), a biography of Muḥāsibī and an analysis of his mystical ideas; JOSEF VAN ESS, "Ibn Kullāb und die Miḥna," *Oriens*, 18–19:92–142 (1965–66), an article about Muḥāsibī's importance for the theology of his time.

(J.v.E.)

Mukden

Mukden, the Manchu name of the city known in Chinese as Shen-yang, capital of Liaoning Province, is the fourth or fifth most populous city in China and the industrial centre and largest city of the Northeast (formerly known as Manchuria). It is located on the Hun Ho (*ho*, "river"), a left-bank tributary of the Liao Ho, on the eastern edge of the Lower Liao Basin in the southern part of the central Northeast (Manchurian) Plain, about 100 miles (160 kilometres) north-northeast of the Gulf of Liaotung and about 140 miles west of the Yalu River, which forms the boundary between China and North Korea. In the early 1970s more than 3,000,000 persons were estimated to occupy the official municipal district, which covered 1,200 square miles.

History. *Early history.* The Lower Liao Basin, from the time of the Han dynasty (206 BC–AD 220), has been known as the Chinese Pale, an area of Chinese settlement, settled chiefly by immigrants from what are now the provinces of Hopeh and Shantung. The rest of Manchuria, not open to extensive Chinese settlement until relatively recent times, was under the control of various nomadic and tribal peoples of whom the Manchus became the most important. The Pale was largely set off from the rest of Manchuria by a discontinuous wall known as the Willow Palisade; and Mukden, then known as Shen, was a frontier post along the edge of the wall.

The Chinese Pale

The strategic and focal significance of its position was recognized at an early age. By the 10th century, it had become a major frontier settlement in the so-called Tung-ching Circuit of the Khitan kingdom, which had expanded across southern Manchuria into Korea from the steppelands of Jehol and western Manchuria. Its dominant peoples, also known as the Khitan, founded the Liao dynasty, which had its base in southern Manchuria and, for most of the period between 947 and 1125, controlled part of North China as well. Southern Manchuria was conquered by the Chin, or Juchen, peoples by 1122–23 and a century later by the Mongols, who by about 1280 had completed their conquest of all China. It was under the Mongols that the name Shen-yang was first applied to Mukden. Less than a century later, by 1368, the Ming dynasty had displaced the Mongols and extended its control over what had been the Chinese Pale of southern Manchuria, including the city that later became known as Mukden.

In the early 17th century, Nurhachi, a tribal chieftain, organized and united the various Manchu tribes in the eastern Manchurian highlands and established military and political control over all of Manchuria. With a view to the conquest of China, Nurhachi moved his capital down to the city on the plain of the Liao. Mukden, a Manchu name, proved an admirable organizing base for the conquest of China. In 1644, when the Manchus supplanted the Mings on the Imperial throne, the capital was transferred to Peking, and Mukden became known as Sheng-ching (Flourishing Capital). It retained its prestige as the older capital of the reigning dynasty, however; the treasures of the royal house were deposited in its palaces, and the tombs of the earlier emperors, particularly the mausoleum of Ch'ing T'ai Tsung (died 1643) at Pei Ling (North Mausoleum), about four miles from the city, are among the most famous monuments of China. In 1658 the provincial government of Feng-t'ien was established, and the city also became known as Feng-t'ien (Honoured Heaven). For 250 years thereafter, it grew steadily, especially in the last half of the 19th century when Chinese immigration to Manchuria reached flood proportions.

Use of Mukden as a base for the Manchu conquest

Modern history. In the period of struggle between Russia and Japan for dominance in Manchuria after 1895, Mukden was inevitably one of the key positions. From that time, when the Russians gained rights to build railroads in Manchuria, Mukden was a Russian stronghold; during the Russo-Japanese War (1904–05) it was the scene of the Battle of Mukden, which lasted from February 19 to March 10, 1905, when the city finally fell to the Japanese. After the war the Japanese concession in Mukden was one of their chief bases for the exploitation of southern Manchuria, but it was also the seat of government of the Chinese "viceroy" of the three Manchurian provinces, Feng-t'ien (Liaoning), Kirin, and Heilungkiang.

During the civil war and warlord period after World War I, the Chinese warlord Chang Tso-lin, a protégé of the Japanese, governed the Manchurian provinces from Mukden and participated with other warlords in the struggle for control of the shadowy government in Peking. The last warlord to resist the advance of the Kuomintang armies against Peking in 1928, he was killed in his retreat with his defeated troops.

Three years later, on September 18, 1931, another explosion touched off the Mukden Incident. An alleged Chinese bomb went off on the railway track near Mukden and gave the signal for a surprise Japanese attack on the Chinese Nationalist garrison and arsenal in Mukden. After protracted fighting, the Chinese forces were driven out of Manchuria. On September 24 the first puppet organization was set up in Mukden with the Committee for Preservation of Local Peace and Order. But, after the establishment of Manchukuo, a puppet state created out of Manchuria and Jehol, the capital was moved to Ch'angch'un, which the Japanese renamed Hsinking (Hsinching; New Capital). Under the puppet state of Manchutikuo (a later version of Manchukuo), Mukden was given its old Chinese name of Feng-t'ien. It also was the headquarters of the South Manchurian Railway (known

The Mukden Incident

familiarly as the Mantetsu), which was the most powerful corporate body in the new state and a direct arm of Japanese colonial policy. Although Hsinking was the capital, Mukden, which had grown rapidly with industrialization, was in fact the principal Manchurian city.

The Soviet Union declared war on Japan in early August 1945 and soon took Mukden. Several months after the surrender of Japan on August 14, 1945, Mukden was occupied by Chinese Nationalist troops (in March 1946), some of whom were flown in by U.S. transport planes. But, while Mukden had been occupied by Soviet troops, it and other Manchurian cities were stripped of much of their industrial plant. Chinese Communist forces also controlled large areas of Manchuria. During the ensuing civil war (1946–49) Mukden, a Nationalist stronghold, fell to Communist armies under Lin Piao on October 30, 1948. The city then served as a major base for the subsequent Communist conquest of the entire Chinese mainland in the following year and a half.

Between 1949 and 1954 Mukden, again under the name Shen-yang, served as the capital of the Northeast Administrative Region, one of six great administrative regions into which China as a whole was then divided. Under the leadership of Kao Kang, a leading Communist, the Northeast developed dramatically as a semi-autonomous, comparatively modernized region of China. In 1954 the regions were abolished, and Kao Kang was purged. Mukden then became the capital of Liaoning Province, one of three into which the Northeast was then divided. In 1956 it was reduced from a municipality of provincial rank to one of subprovincial rank.

The contemporary city. *Strategic location.* The site of Mukden commands the lower portion of the central Manchurian Plain, the largest erosional (instead of depositional) lowland in China and probably the most fertile. Mukden controls the main north–south corridor, which extends from the Liao Ho Basin to the south and north across the continental divide near Ssu-p'ing, Kirin Province, to the basin of the Sungari River, a tributary of the Amur, which drains much of the northern region of the Northeast. The Liao Ho Basin opens onto the Gulf of Liaotung, the northern arm of the Gulf of Chihli, a branch of the Yellow Sea. The Gulf of Chihli is a shallow, partly enclosed sea, across which movement from North China, including the Shantung Peninsula, to the Liao lowland has long been undertaken by the Chinese people. In addition, the gulf has also served as a trade route. In fact, most of the population of the Manchurian Plain had its ancestral origins in the northern Chinese provinces of Hopeh and Shantung, which border the Gulf of Chihli. The Hun Ho Valley has long provided Mukden with another major route: eastward toward the Yalu River and North Korea; and to the southeast of Mukden another major routeway leads across the east Manchurian hills to the mouth of the Yalu and again North Korea.

Climate. The climate of the city is similar to that of Minneapolis, Minnesota, or of Moscow, but the summers are somewhat warmer. January mean temperature is 9° F (–13° C); July mean is about 75° F (about 24° C). The winters are long and hard, the summers short and hot. Annual precipitation averages just over 26 inches (660 millimetres), most of it concentrated in the summer months, although winter snowfall and snow cover are common.

City plan. Mukden lies on a level site at an altitude of about 185 feet (56 metres). Its structure reflects its diverse origins and complex history. The core of the city is the small old Chinese city, most of whose original 10-mile-long surrounding earthen wall has either been removed or is in a state of disrepair. Within the old city are the former Manchu Imperial Palace and its associated structures, modelled after the Forbidden City in Peking. North of the old city are the tombs of the Manchu emperors, as at Pei Ling. To the east is the former Chinese arsenal district. To the west lies the extensive Japanese-built city, laid out on an irregular rectangular grid pattern but with circles and diagonals reminiscent of other examples of Japanese Imperial urban design. Grand boulevards, shaded by trees and divided into separate routeways for fast and slow traffic, are typical of this area, which came into being after 1905 and which was greatly expanded after 1931. Here also are the main administrative buildings and facilities such as the main post office, the larger hospitals, and the greater number of theatres and libraries. Parks and recreational areas also are common. To the far west is the newest part of the metropolis, T'ieh-hsi-chu (West of the Railroad District), laid out in a regular grid pattern oriented to the northeast. It was constructed after 1935 as a workingman's residential area, as the industrial base of the city expanded, and has since become the most rapidly growing district. The official municipal district of Mukden covers about 1,200 square miles (3,100 square kilometres), about 60 percent of which is built-up area; the remainder is essentially rural in character. This overbounding means in effect that the city has no suburbs. But the expansion of the urbanized area within the district promises to convert almost all of these open lands into urban-land uses as the population and functions of the city expand.

Traditional neighbourhoods and regions

Transportation. Mukden was the railway hub for the southern area of the Northeast and remains, next to Peking, the leading railway centre of modern China. Railways lead northward to Ch'ang-ch'un and Harbin; southward to Dairen and Port Arthur (later merged to form the municipality of Lü-ta); southeastward to Tan-tung and the mouth of the Yalu; and southwestward to Chinchou, Shan-hai-kuan and the North China Plain, including Peking and Tientsin. Large railway marshalling yards were associated with manufacturing facilities for rolling stock. The South Manchurian Railway, administered from Mukden, was self-sufficient in the production of rolling stock and locomotives, which were manufactured in Dairen. A basic road network also was constructed to converge on the growing city. Mukden's railway station, with an associated bus terminal, lies on the margins of the Japanese-built part of the city. Bus services link the various parts of the metropolis. A large airport, constructed by the Japanese, is on the northern outskirts.

Demography. Mukden municipality (including the old city, the new city developed by the Japanese, the old arsenal district, and the T'ieh-hsi-chu) was believed to have a population of well over 3,000,000 in 1970, as compared with about 2,300,000 in 1953, a little under 2,000,000 in 1945, and about 421,000 in 1931. When its immediate hinterland—which includes several of the major industrial cities of China—is taken into account, the total population under the domination of Mukden approaches 10,000,000. This makes it comparable in size as well as function to such diverse metropolitan complexes as Chicago, Ōsaka, and Calcutta.

Economic life. During the period of Japanese hegemony in southern Manchuria, Mukden was developed as the effective regional capital of the Northeast and served as its chief industrial centre. Within a radius of about 50 miles three great industrial cities were developed: An-shan, 50 miles to the south-southwest; Pen-ch'i, 30 miles to the southeast; and Fu-shun, 20 miles to the east on the Hun Ho.

Since 1950 Mukden has continued as the hub of the heavy industrial complex of the southern region of China's Northeast, the greatest in China. It is estimated that about a fourth of all the machinery production and metal fabricating in China is done there. Rolling stock, machine tools, wire and cables, cement, electrical equipment, chemicals and chemical fertilizers, and pharmaceuticals are produced there. The complex also includes oil-seed-processing plants, flour mills, paper plants, soap and leather factories, textile mills, and glass factories. A small local steel plant produces finished metal for some of these industries, using local scrap and drawing pig iron from the mills at An-shan. The smelting of metals such as copper, zinc, lead, and manganese also is an important part of the industrial mix of the city. Coal to fuel its industries comes from Fu-shun and Fou-hsin.

Industrial products

For its industrial and domestic power needs, the city is tied into the southern electric power grid of the North-

east, which depends for its power sources on a great dam on the Sungari near Kirin (city) and on an even larger high dam at Shui-feng (Sup'ung) on the Yalu, the power of which is shared with North Korea. In addition, Mukden has a large, 600,000-kilowatt coal-fired generating plant, the first two units of which were opened in 1959.

The city obtains its water supply from a reservoir on the Hun Ho, a few miles upstream of Fu-shun.

Education and culture. Mukden has long functioned as the education and cultural centre of the Northeast. The city houses a major university, Liaoning University (formerly Feng-hang University), Northeastern China Technical University, Northeastern Engineering College, Northeastern Institute of Finance and Economics, and a medical college. In addition to theatres and libraries (the Liaoning Library has more than 1,300,000 volumes), it is the home of the Northeastern Institute of Fine Arts and the Northeastern Music Institute. It is also the site of the Northeast Museum and of the North Mausoleum and other Chinese historical monuments. One of China's regional television stations also is located in Mukden.

(N.S.G.)

Muller, Hermann Joseph

One of the founders of genetics and a Nobel laureate, Hermann Joseph Muller is chiefly remembered for his idea of the gene as the basis of life, his demonstration that X rays produce mutations, and his advocacy—as a means of directing human evolution—of free choice from sperm banks for artificial insemination. Muller's main conceptual contribution to biology is that evolution is implicit in the two properties unique to the genetic material: that of acting as template in its own replication and that of undergoing occasional changes ("mutations") transmissible to the replicas.

By courtesy of Indiana University, Bloomington

Muller.

Muller was born on December 21, 1890, in New York City. He attended Columbia University (1907–09) on a scholarship supplemented by part-time jobs.

Work at Columbia University

At Columbia his interest in genetics was fired first, as an undergraduate, by E.B. Wilson, the founder of the cellular approach to heredity, and later, in a postgraduate course, by T.H. Morgan, who had just introduced the fruit fly *Drosophila* as a tool in experimental genetics. The possibility—indeed, the necessity, in Muller's later view—of consciously guiding the evolution of man was the initial motive in Muller's scientific work and social attitudes. His early experience at Columbia convinced him that the first necessary prerequisite was a better understanding of the processes of heredity and variation.

A laboratory assistantship in zoology in 1912 allowed him to spend part of his time doing research on *Drosophila* at Columbia. He produced a series of papers, now classic, on the mechanism of crossing-over, obtaining his Ph.D. in 1916. The work of the *Drosophila* group, headed by Morgan, was summarized in 1915 in the book *The Mechanism of Mendelian Heredity*. This book is a cornerstone of classical genetics.

After three years at the Rice Institute, Houston, Texas, and an interlude at Columbia as instructor, Muller in 1920 became associate professor (later professor) in the department of zoology, University of Texas, Austin, where he remained until 1932.

The 12 years at Austin were scientifically the most productive in Muller's life. The study of the process of mutation led Muller to form a picture of the nature of genetic material and later to the experimental induction of mutations (the basis of his Nobel Prize in 1946). In 1931 he was elected to the U.S. National Academy of Sciences.

Despite Muller's progress, the pressure of work, his despair at social conditions during the Depression, the constraints on the free expression of his Socialist views in Texas, and domestic difficulties (his marriage of 1923 to the mathematician Jessie Marie Jacobs came to an end in 1932), all combined to produce a nervous breakdown. As a diversion, he spent one year at the Kaiser Wilhelm Institute in Berlin, where, in conjunction with a group of colleagues, he originated the biophysical attack on the nature of the gene. In September 1933 he moved to Leningrad at the invitation of N.I. Vavilov, head of the Institute of Genetics there. Shortly after, the Institute was transferred to Moscow, and Muller moved with it. For Muller, the opportunity of working in genetics in a Socialist society was a fulfillment of a long-held wish. The inspiration and help of Vavilov and a group of able, enthusiastic young collaborators provided a fertile environment for Muller's research.

Muller's work in the Soviet Union

Genetics in the Soviet Union at this time could have made great strides. But the false doctrines of the biologist T.D. Lysenko became politically powerful, bringing to an end the activities of all flourishing research groups. Soviet biology was consequently left unprepared for the next leap toward molecular biology.

Muller fought Lysenkoism whenever possible, but ultimately had to leave the U.S.S.R. He spent three years at the Institute of Animal Genetics in Edinburgh, returning to the United States in August 1940. The period in Edinburgh was very fruitful. He had a profound impact on genetics in Great Britain and helped to organize the VIIth International Congress of Genetics. At the Institute he met Dorothea Johanna Kantorowitz, a German refugee, who became his second wife in 1939.

On returning to the United States, Muller obtained temporary positions at Amherst College in Massachusetts (1941–45), and, finally, a professorship in zoology at Indiana University, Bloomington.

The award of the Nobel Prize to Muller in 1946 increased both his opportunity and his sense of obligation regarding social activity. At the onset of the atomic age he was foremost in promoting awareness of the dangers of radiation to future generations. He also became more actively involved in discussions on the evolutionary future of man in relation to the relaxed processes of selection operating in modern society. His concern for man's future and the positive action that he had advocated for many years are summarized in the clearest terms in his address ("What genetic course will man steer?") to the International Congress of Human Genetics in 1966. Muller died on April 5, 1967.

A never-ending curiosity, a great breadth and depth of imagination, and a deep humanity combined with uncanny experimental dexterity made Muller a commanding figure in the biology of the 20th century.

BIBLIOGRAPHY. G. PONTECORVO, "Hermann Joseph Muller 1890–1967," *Biogr. Mem. Fellows R. Soc.*, 14:349–389 (1968), contains the list of some manuscripts and the complete list of about 370 publications, and includes both detailed biographical data and an evaluation of Muller's work. Publications, manuscripts, and correspondence are deposited in the Lilly Library, Indiana University, Bloomington. *Studies in Genetics: The Selected Papers of H.J. Muller* (1962), contains a selection of papers, and of excerpts from papers, adequate for an appreciation of Muller's work and ideas. The selection was made by Muller himself.

(G.Po.)

Müller, Johannes Peter

Johannes Müller was one of the great natural philosophers of the 19th century. In German medicine and biology he marked the transition from naturalistic speculation to precise observation and physiological experimentation. In physiology, anatomy, and zoology he established new concepts and trained outstanding students who broadened the scope of medicine by introducing the techniques of natural science.

By courtesy of the Institut fur Geschichte der Medizin der Freien Universitat Berlin

Müller, lithograph by Louisa Corbaux, after a drawing by S. Lawrence.

Müller was born in Koblenz on July 14, 1801, the son of a shoemaker. In 1819, he entered the University of Bonn, where the faculty of medicine was permeated with a naturalistic romanticism, which the young Müller eagerly espoused. His thesis *Über die Bewegungsgesetze der Tiere* (1822; "On the Laws of Movement of Animals") was written in the naturalistic manner. He continued his studies at the University of Berlin, where he came under the influence of the sober, precise anatomist Karl Rudolphi and thereby freed himself from naturalistic speculation.

In 1824 he was granted a lectureship in physiology and comparative anatomy at the University of Bonn. In his inaugural lecture, "Physiology, a science in need of a philosophical view of nature," he outlined his approach to science, maintaining that the physiologist must combine empirically established facts with philosophical thinking. Two years later he was appointed associate professor, and in 1830 he became a full professor.

In the meantime, his voluminous *Zur vergleichenden Physiologie des Gesichtssinnes* . . . (1826; "the Comparative Physiology of Visual Sense . . . ") brought Müller to the attention of scholars by its wealth of new material on human and animal vision; he included the results of analyses of human expressions and research on the compound eyes of insects and crustaceans. His most important achievement, however, was the discovery that each of the sense organs responds to different kinds of stimuli in its own particular way or, as Müller wrote, with its own specific energy. The phenomena of the external world are perceived, therefore, only by the changes they produce in sensory systems. His findings had an impact even on the theory of knowledge.

Significance of his early work in physiology

Müller's monograph "On Imaginary Apparitions" was also published in 1826. According to this theory the eye as a sensory system reacts not only to external optical stimuli but can also be excited by internal stimuli generated by the imagination. Thus, persons who report seeing religious visions, ghosts, or phantoms may actually be experiencing optical sensations and believe them to be of external origin, even though they do not in fact have an adequate external stimulus.

Because of his research Müller became the best known figure in the Bonn faculty of medicine. Maintaining an almost incredible level of output, he examined many problems in physiology, evolution, and comparative anatomy. He studied the passage of impulses from afferent nerves (going to the brain and spinal cord) to efferent nerves (going away from the same centres), further elucidating the concept of reflex action. By careful experiments on living frogs, he confirmed the law named after Charles Bell and François Magendie, according to which the anterior roots of the nerves originating from the spinal cord are motor and the posterior roots are sensory. He investigated the nervous system of lower animal species, the intricate structure of glands, and the process of secretion. When tracing the development of the genitalia, he discovered what is now known as the Müllerian duct, which forms the female internal sexual organs. He contributed to our knowledge of the composition of the blood and lymph, the process of coagulation, the structure of lymph hearts of frogs, the formation of images on the retina of the eye, and the propagation of sound in the middle ear.

Studies in physiology and pathology in Berlin

In 1833 Müller was called to Berlin to succeed Rudolphi. In his new post he again carefully explored many problems concerning animal function and structure. His early years in Berlin were devoted mainly to physiology. His *Handbuch der Physiologie des Menschen für Vorlesungen* (2 vol., 1834–40; Eng. trans., *Elements of Physiology*, 2 vol., 1837–42) was recognized throughout the world and established a beneficial interchange between physiology and hospital practice in Germany. It stimulated further basic research and became a starting point for the mechanistic concept of life processes, which was widely accepted in the second half of the 19th century.

Inspired by the vast Berlin anatomical collection, Müller became interested again in pathology. After the demonstration by his assistant, Theodor Schwann, that the cell was the basic unit of structure in the animal body, he concentrated on the cellular structure of tumours with the aid of a microscope. In 1838 his work *Über den feineren Bau und die Formen der krankhaften Geschwülste* (*On the Nature and Structural Characteristics of Cancer, and of Those Morbid Growths Which May Be Confounded With It*, 1840) began to establish pathological histology as an independent branch of science. Müller also distinguished himself as a teacher. His students included the renowned physiologist and physicist Hermann Helmholtz and the cellular pathologist Rudolf Virchow.

Later work in taxonomy and morphology

Beginning in 1840 Müller increasingly focussed his research on comparative anatomy and zoology, in so doing becoming one of the most respected scholars in these subjects. He was a master at collecting and classifying specimens; he devised an improved classification of fish, and, based on an ingenious analysis of vocal organs, did the same for singing birds. For several years he concentrated on the lowest forms of marine vertebrates, the Cyclostomata and Chondrichthyes. He painstakingly described the structures and complex development of members of various classes of the invertebrate phylum Echinodermata. His last research activities were concerned with the marine protozoans Radiolaria and Foraminifera.

In 1827, 1840, and 1848, Müller suffered periods of depression that rendered him incapable of working for months on end. They may perhaps be attributed—as his periods of explosive productivity—to a manic-depressive disposition. It may also be regarded as the cause of his death, on April 28, 1858, in Berlin. Some scholars have assumed that he died by his own hand.

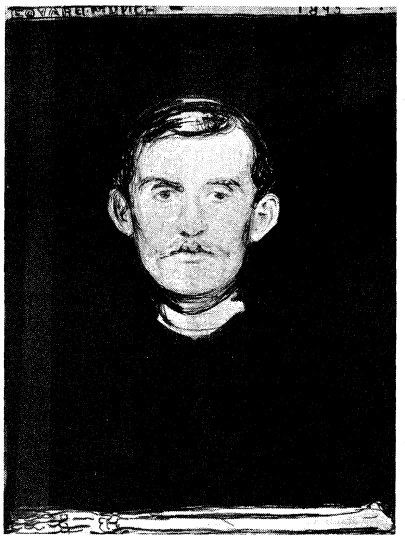
BIBLIOGRAPHY. ULRICH EBBECKE, *Johannes Müller, der grosse rheinische Physiologe* (1951), includes a new edition of Müller's work *Über die phantastischen Gesichtserscheinungen*. See also WILHELM HABERLING, *Johannes Müller: Das Leben des rheinischen Naturforschers* (1924), an extensive biography with numerous letters to relatives, friends, and colleagues; GOTTFRIED KOLLER, *Das Leben des Biologen Johannes Müller 1801–1858* (1958), a biography that does justice to Müller's zoological achievements; JOHANNES STEUDEL, *Le Physiologiste Johannes Müller* (1963), a brief biography; and WALTHER RIESE and GEORGE E. ARRINGTON, JR., "The History of Johannes Mueller's Doctrine of the Specific Energies of the Senses: Original and Later Versions," *Bull. Hist. Med.*, 37:179–183 (1963).

(J.Ste.)

Munch, Edvard

Revolting against the stilted academicism of 19th-century naturalistic painting, the Scandinavian painter and printmaker Edvard Munch developed a highly personal art that was the forerunner of 20th-century Expressionism. When he was 26 he wrote, "I shall paint living people who breathe, feel, suffer, and love. The sacredness of this will be understood and people will take off their hats as though they were in church." The emotional power of these expressions of the human condition made him not only the greatest Norwegian artist but one of the most noted figures in the early development of modern art.

By courtesy of the Albertina, Vienna



Munch, self-portrait, lithograph, 1895. In the Albertina, Vienna. 45.7 × 32 cm.

Born on December 12, 1863, at Löten into a family prominent for its contribution to the political and cultural life of Norway, Munch was the second eldest of five children. His mother died of tuberculosis when he was five; ten years later his eldest sister died of the same disease, and another sister became mentally ill. Both his father and brother also died while Munch was a youth. The spectre of death and terminal illness that filled his childhood was to preoccupy him throughout his life, becoming one of the principal themes in his work. "The Sick Child" (1886; Nasjonalgalleriet, Oslo), "The Death Chamber" (1892; Munch-museet, Oslo), "By the Death Bed" (1895; Rasmus Meyers Collection, Bergen), and "The Dead Mother" (1899; Kunsthalle, Bremen, West Germany) are among the many paintings and prints that reflect his macabre childhood impressions.

The death theme in Munch's work

At an early age, Munch showed a flair for drawing but received little formal training. For a short time, he attended the School of Art and Handcraft in Christiania (Oslo). His student works were criticized by Christian Krohg (1852–1925), the important Norwegian naturalistic painter, who introduced Munch to French Impressionism. After his first trip to Paris in 1885, such pictures as "The Evening Hour" were executed in an Impressionistic manner. Subsequent exposure to contemporary French art effected a major change in his style, when (*c.* 1890) he was influenced by the Postimpressionist work of such French painters as Paul Gauguin and Henri de Toulouse-Lautrec, as well as by Art Nouveau design. The Impressionistic "Spring Day on Karl Johan Street" (1891; Gallery of Painting, Bergen) contrasted to the Expressionistic "Evening on Karl Johan Street" (1892; Rasmus Meyers Collection, Bergen) dramatically marks this stylistic transition.

The Christiania of Munch's youth was a provincial town, ruled by what he regarded as a narrow-minded and self-righteous gentility, whose cultural interests were limited and conventional. Munch early became associated with a Bohemian element involved with new ideas of social behaviour, morals, and art. The effect of this milieu was emancipating, and Munch's art became increasingly concerned with sex and social conditions.

Munch was to count among his friends many important literary personalities, including the Scandinavian playwrights Henrik Ibsen and August Strindberg, the Norwegian novelist Knut Hamsun, and the French Symbolist poet Stéphane Mallarmé. He had a particular affinity to the social realism of Ibsen's works and designed sets for productions of *Hedda Gabler* and *Ghosts*, as well as illustrating an edition of *The Pretenders*.

In 1892, Munch was invited to exhibit in Germany by the Verein Berliner Künstler (Union of Berlin Artists), a society of Berlin artists. The exhibit of 45 works opened on November 5 and closed a week later. The violent emotionalism and unconventional imagery of these works created a bitter controversy. The "Munch Affair" was hotly debated in the German press, developing into a struggle over the freedom of artistic expression. Munch's expressionism was soon accepted, however, and he became a European celebrity, living mainly in Berlin and Paris. In 1894, influenced by the revival of the graphic arts in Germany during the 1890s, he began to make prints. Using the mediums of etching, lithography, and woodcut, Munch was to become one of the most significant modern printmakers. Noted for its simplicity and directness, the imagery in his prints was characterized by the same ruthless subjectivity seen in his paintings.

Controversy in Berlin

The 16 years between Munch's first Berlin exhibition and a nervous breakdown in 1908 were the most important of his career. He drove himself with a frightening, neurotic compulsiveness. In a tortuous linear style, he expressed the misery and anxiety of human existence. "The Cry," perhaps Munch's most famous work, is a symbol of modern man's spiritual anguish. This work clearly points to the evocative violence of German Expressionist painting of the early 20th century and to the convulsive reality of another major modern symbol, Pablo Picasso's "Guernica" (1937; Museum of Modern Art, New York).

After 1910, Munch resided permanently in Norway. The self-consuming and introverted aspects of his earlier works, in which the individual is victimized by passion and the malevolence of nature, are replaced by the joy of life, and the abundance of nature and man's affinity with it. Direct contact with nature, always important to Munch, is expressed in numerous landscapes painted in a free, vigorous, and broad manner. The brighter and richer colour of these paintings is a further expression of his more positive frame of mind. During this period, his most monumental works were the murals he painted for the Aula, or festival hall, of Oslo University.

Abandoning printmaking, Munch spent his last years in solitude at his home near Oslo, where he continued to paint with unflagging energy until his death on January 23, 1944. Munch bequeathed his estate and all the paintings, prints, and drawings in his possession to the city of Oslo, which erected the Munch Museum in 1963.

MAJOR WORKS

PAINTINGS: "The Sick Child" (1886; Nasjonalgalleriet, Oslo); "The Evening Hour" (1888; Munch Museum, Oslo); "Inger Munch" (1892; Nasjonalgalleriet, Oslo); "Evening on Karl Johan Street" (1892; Rasmus Meyers Collection, Bergen); "The Cry" (1893; Nasjonalgalleriet, Oslo); "Vampire" (1893–94; Munch Museum, Oslo); "Ashes" (1894; Nasjonalgalleriet, Oslo); "Puberty" (1894; Nasjonalgalleriet, Oslo); "In Hell, Self-Portrait" (*c.* 1895; Munch Museum, Oslo); "Self-Portrait with a Cigarette" (1895; Nasjonalgalleriet, Oslo); "Jealousy" (1895?; Rasmus Meyers Collection, Bergen); "The Dance of Life" (1899–1900; Nasjonalgalleriet, Oslo); "The Red Vine" (1900; Munch Museum, Oslo); "White Night" (1901; Nasjonalgalleriet, Oslo); "Village Street" (1905; Munch Museum, Oslo); "Marat's Death" (1906; Munch Museum, Oslo); "The Bathers Triptych" (1907–13; Munch Museum, Oslo); murals for the Aula of Oslo University (1910–16); "Self-Portrait—Spanish Influenza" (1919; Nasjonalgalleriet, Oslo); "Model by the Armchair" (1929; Munch Museum, Oslo); "Between Clock and Bed, Self-Portrait" (1940; Munch Museum, Oslo).

BIBLIOGRAPHY. GUSTAV SCHIEFLER, *Edvard Munchs graphische Kunst* (1923), represents a nearly complete catalog of the artist's graphic opus; S. WILLOCH (ed.), *Edvard Munchs raderinger* (1950), a complete collection of repro-

ductions of the artist's etchings. Most of the biographical works on the artist and his works are written in German or Scandinavian languages. In English the following books give a representative introduction to the subject: ARVE MOEN, *Edvard Munch: Age and Milieu* (1956), *Edvard Munch: Women and Eros* (1957), *Edvard Munch: Nature and Animals* (1958), illustrated with practically all major works in colour and in black and white, with a bibliography of books also in German and Scandinavian languages; J.H. LANGAARD and REIDAR REVOLD, *Masterpieces from the Artist's own Collection* (1959); OTTO BENESCH, *Edvard Munch* (1960); PAL HOUGEN, *Edvard Munch and the Munch Museum* (1968); ROLF STENERSEN, *A Close up of a Genius* (1969); and WERNER TIMM, *The Graphic Art of Edvard Munch* (1969).

(A.S.M.)

Munich

Sometimes called "Germany's secret capital," Munich (German München) is, after Berlin and Hamburg, the third largest city in the German Federal Republic and the largest city of the *Land* (state) of Bayern (Bavaria). It is an exceptionally attractive city and is, indeed, considered by many to be one of the finest in Europe because of its wealth of Baroque, Rococo, and neoclassical architecture. Situated some 30 miles from the edge of the Alps, its unique character might perhaps be described as a combination of German practicality with Alpine romanticism, of urban thinking with a rural way of life. Although it is thoroughly modernized, it retains something of the genial sociability (*Gemütlichkeit*) of the old royal Bavaria. This, in a high-spirited form, is happily shown in the autumn of each year, when the people of Munich, particularly its 30,000 students, celebrate with beer, music, and traditional costumes their famous folk festival, the Oktoberfest.

Origin of name

History. Munich, or München (a word meaning the Home of the Monks) as it was originally called, traces its origin to the Benedictine monastery at Tegernsee, which was probably founded in AD 750. Henry the Lion, duke of Bavaria during the latter half of the 12th century, in 1157 granted the monks the right to establish a market where the road used for transporting salt from Salzburg met the Isar River. A bridge was built the following year, and the marketplace was fortified. From this small beginning, Munich thus grew and flourished, and Duke Henry may be regarded as its true founder.

Munich's subsequent history was influenced by two important factors: the increase in the flow of trade in Europe and the succession of Otto of Wittelsbach to the Bavarian duchy in 1180. For over 700 years the Wittelsbach family, whose home it became from 1255, was closely connected with the town's destiny. Three members, in particular, of this great German dynasty were pre-eminent.

In the early 14th century the emperor Louis IV, called the Bavarian, erected a second city wall and, after a disastrous fire in 1327, rebuilt and expanded the town to the size at which it remained up to the end of the 18th century.

Under the elector Maximilian I (1597–1651), a powerful and effective ruler who was also a patron of the arts, Munich increased in wealth and size and saw the flowering of its Renaissance architecture. The Thirty Years' War, however, brought suffering. It was occupied briefly by the Swedes under Gustavus Adolphus in 1632, and two years later a plague epidemic killed off one-third of its population of about 20,000.

The third Wittelsbach who left his mark on the town was Louis I, king of Bavaria from 1825 to 1848. It was he who planned and created modern Munich (see below) and whose architects established in their public buildings the style of Romantic Classicism, which still gives the city its characteristic aspect.

The 19th century was Munich's greatest period of growth and development. In the early part, when what had been a purely Roman Catholic town admitted Protestants to citizenship for the first time, the population was 40,000. In 1854 it became a "city" with 100,000 inhabitants, and by 1900 the population exceeded 500,000. In the second half of the century, Munich's cultural importance in Europe was particularly enhanced when Louis II, by his championing of the composer Richard Wagner, revived its fame as a city of music and the stage.

The rule of the Wittelsbach dynasty finally ended with the abdication of Louis III in November 1918, and in the immediate aftermath of World War I, Munich was the scene of severe revolutionary fighting, especially in the spring of 1919, when the Räterepublik—the Socialist Republic of Workers' and Soldiers' Councils—was proclaimed. More significantly, the city witnessed the political activity of Adolf Hitler in the 1920s. It was there that he joined the National Socialist (Nazi) Party and became its leader, held vast open-air meetings, and organized his first party army. The beer cellar where he held meetings that led to the *Putsch* ("rising") against the Bavarian authorities on November 8, 1923, can still be seen.

The Munich *Putsch*

In World War II, Munich suffered heavily from bomb raids, with 43 percent of its buildings destroyed, but most of the damage has been repaired.

The contemporary city. *Location and boundaries.* Munich is situated in Upper Bavaria, 1,700 feet above sea level, some 30 miles north of the edge of the Alps, and about 65 miles east-southeast of Augsburg. The city area covers 120 square miles (311 square kilometres), extending 13 miles from north to south and 16 miles from west to east. The Isar River flows through the middle of the town.

Architectural features. The old town, clustered around the ancient crossroads of the marketplace in the Marienplatz, has become more and more of a business centre and has lost much of its ancient character. Among the old buildings that still stand, though damaged in World War II, are three of the seven town gates—Karlstor, Sendlingertor, and Isartor, all dating from the 14th century—and medieval buildings such as the cathedral, the Frauenkirche (1468–88), whose massive cupola-capped towers are conspicuous landmarks, and the old town hall (1470–80), in the Marienplatz, with its beautiful wooden barrel vault. Nearby, the Peterskirche, Munich's oldest church (1169), which was completely destroyed in World War II, has been rebuilt almost in its original form. The former arsenal of the town at Jakobsplatz is now the municipal museum; among its treasures are the ten original Maruska (Morisken) dancers carved in wood in about 1480 by Erasmus Grasser for the town hall.

The Wittelsbach residences

The oldest of the Wittelsbach residences, the 15th-century Alte Hof, has to a great extent retained its medieval appearance. A later residence, the Neue Veste (New Fortress), lies between the Hofgarten and Max-Josef-Platz. The building dates from 1385, but the style of every period up to the 19th century was tried out on it with superb effect; it was destroyed in World War II but has been reconstructed. Of the Maxburg, built in 1590 by Duke William V, only one tower remains; a modern steel and concrete building now stands in its place. Other survivals from this period are the Münzhof, at one time the Duke's art gallery and stables (1563–67), and the Michaelkirche (1583–97), which is considered to be the most important Renaissance church in Germany and one of the most beautiful in central Europe.

The old town is predominantly Baroque and Rococo. Italian models inspired the late-17th-century Theatinerkirche, the early-18th-century Dreifaltigkeitkirche, and Bürgersaal. St. Michael at Berg am Laim, the Johann-Nepomuk-Kirche, and the Residenztheater (the court theatre) were completed between 1743 and 1753. The last named, a magnificent Rococo theatre, in the Neue Veste, designed by the French architect François de Cuvilliés, was largely destroyed in World War II, but much of its interior decoration had been evacuated and saved, and the theatre was reconstructed in its original form and reopened in 1958. Outside the town walls were built summer castles: Schleissheim and Nymphenburg, the latter with its hunting lodge the Amalienburg Pavilion, another Rococo masterpiece by Cuvilliés. The Bavarian landed gentry and court nobility built their houses around the court at the same time early in the 18th century. Of these the Preysing palais, the Palais Pocci, and the Palais Holnstein still remain.

Munich with the towers (from left to right) of Peterskirche, Frauenkirche, the town hall, and the Heiliggeistkirche, now the municipal museum.
D.P.A.—Pictorial Parade

Developments under King Louis I

Munich received the first great impetus to expand beyond the old town through the magnificent building and artistic development under King Louis I in the 19th century. Between the Siegestor (Victory Arch) and the Feldherrnhalle (Hall of Generals), he had the monumental Ludwigstrasse built and constructed there the state library, the Ludwigskirche, and the university. Other work commissioned by Louis I were the Königsplatz with the Glyptothek, which houses a collection of ancient and modern sculpture, the Propyläen, the Staatsgalerie, the Alte Pinakothek (containing one of the great European collections of paintings), the king's apartments, and the court banqueting hall. Louis' son Maximilian II (1848-64) built the Maximilianstrasse with the Maximilianeum, today the seat of the Bavarian Landtag (Diet).

The people. Munich's population in 1970 was approximately 1,300,000, 11 percent of whom were non-German. The metropolitan region, comprising an additional 140 communities, has about 1,800,000, inhabitants. The annual increase in population is 30,000, which is due not to an increase in the birthrate but to immigration. This trend may be explained by the attractions of a cultural centre near the Alps but especially by the changes in the economic structure of the German Federal Republic, which has made Munich an increasingly important centre of trade.

Economic life. Munich has in the past suffered economically because of its distance from seaports and from the Ruhr coal mines. But this situation improved when fuels other than coal came into more general use. There was a shift from heavy to light industry, to the manufacture, for example, of precision instruments and of optical and electrical appliances and to the production of foods, cosmetics, and clothing.

Munich beer

The brewing industry is traditional in Munich, with seven well-known breweries—Löwenbräu, Pschorr, Paulaner, Spaten, Hacker, Hofbräu, and Augustiner—producing about 88,000,000 gallons of beer a year. Also important are book publishing and printing and motion-picture production, mainly television, industrial, and educational films. Tourism and the hotel business, for long a profitable line, especially because of the number of trade fairs, exhibitions, and congresses taking place in Munich, were given particular prominence in the city's preparations for the 1972 Olympic Games.

Munich is a centre of great banking houses and, because of its geographical position, possesses one of the largest wholesale markets in Europe for fruit, vegetables, and animal produce.

Transportation. Munich is connected by rail to all the main towns of Germany and Austria, and three *Autobahnen* (expressways)—from Stuttgart, Nürnberg, and Salzburg—converge on the city. The airport at Riem on the eastern outskirts was, because of its limited capacity and nearness to the city, due to be replaced in the 1970s by a larger airport near Erding to the northeast. In the city, traffic congestion has become a serious problem, and to ameliorate the situation, an underground railway (subway) begun in 1968 was completed in 1971.

Government. Munich is administered by a lord mayor, two mayors, and some 40,000 civil servants. This organization is controlled by the city parliament of 60 citizens elected for six years. As capital of Bavaria the city is also the seat of the Bavarian government and parliament.

Education and cultural life. The university, called Ludwig-Maximilians-Universität, was moved from Landshut, on the Isar River, to Munich in 1826. In 1970 it had more than 23,000 students. Among other important educational establishments are the Technische Hochschule München (technical college), founded in 1868, and academies of art and music.

Munich has long been a centre for the arts. Music, especially opera, has flourished. In addition to the Bavarian State Opera Company and the Munich Philharmonic and other orchestras, there are the Bavarian State Dramatic Company and 20 private theatres. Important libraries include the Bavarian State Library, with more than 2,000,000 volumes, the University Library, and the library of the Deutsches Museums (a science library).

There are some outstanding museums and art galleries. In addition to the Glyptothek, the Staatsgalerie, and the Alte Pinakothek (see above), the Bavarian National Museum contains German art and applied art since the Middle Ages; the Residenzmuseum has paintings, sculpture,

furniture, and ceramics; and the Deutsches Museum on Museum Island in the Isar River is a huge and comprehensive museum of science, engineering, and technology. There is also a Modern Art Museum and numerous small art galleries.

The Englischer Garten

Recreation. Munich has many parks and gardens, of which the best known is the Englischer Garten, covering 600 acres near the Isar. This imitation of an English park was laid out in the late 18th century by the British-American scientist Sir Benjamin Thompson (Count von Rumford), who spent 11 years in the Bavarian civil service and was created a count of the Holy Roman Empire for his services.

BIBLIOGRAPHY. For a comprehensive survey, see LUDWIG HOLLWECK, *München: Staatsgeschichte in Jahresporträts* (1968); the same author's compilation, *Unser München* (1967), describes the city in the 20th century. The KREIS DER FREUNDE ALT-MUNCHENS, *München im Wandel der Jahrhunderte*, 2nd ed. (1958), provides good historical coverage. Important for an understanding of the 19th-century political and cultural history of Munich is RUPERT HACKER (ed.), *Ludwig II von Bayern in Augenzeugenberichten* (1966). LUDWIG SCHROTT, *Münchner Alltag in acht Jahrhunderten* (1969), is a good social history describing the daily lives of the city's inhabitants through eight centuries.

(F.Bu./Mi.Sc.)

Müntzer, Thomas

Thomas Müntzer, a leading radical reformer of the 16th-century Protestant Reformation, remains a figure of deep religious and social significance in modern European history. As an exponent of the supremacy of the inner light of the Holy Spirit as against the authority of Scriptures, Müntzer was said by Luther to have swallowed the Holy Spirit, "feathers and all" (the symbol of the Holy Spirit is the dove). Marxists in the 20th century view Müntzer as one of the first major figures in the struggle for a classless society.

By courtesy of the Herzog Anton Ulrich-Museum, Braunschweig

Müntzer, engraving by an unknown artist, 1525.

Early life and career

Thomas Müntzer (Münzer; Monczer; Latin Monetarius) was born sometime before 1490, the son of a burgher in Stolberg in the Harz Mountains. Of his childhood and youth very little is known. Traces of his family, who presumably had to leave Stolberg because of civil disturbances involving heretical ideas, are to be found in Aschersleben, Halberstadt, and Quedlinburg, where he may have attended the Latin school. His name appears in the 1506 register of Leipzig University, and in 1512 he attended the University of Frankfurt an der Oder, later earning the academic ranks of master of arts and bachelor of theology. As a linguistic specialist in Latin, Greek, and Hebrew and an accomplished scholar of ancient and humanistic literature—particularly the Old and New Testaments—he belonged to the most learned and intellectually significant group of his time. He was an assistant teacher in Halle (Saale) in 1513 and a clergyman as well as a teacher in Aschersleben in 1514 and 1515. In these capacities he represented the interests of the middle class in its striving for church reforms. He initiated secret alliances between various groups in order to achieve the reforms.

From 1516 to 1517 Müntzer worked as a prior at Frohse monastery at Aschersleben; in 1517–18 he taught at the Braunschweig Martineum (city secondary school) until, in 1518, he was attracted to Martin Luther and his ideas of reform. The designation Martinian was first applied to Müntzer in 1519 after he spoke out against the Franciscan monastic order, the Roman Catholic ecclesiastical hierarchy, and the veneration of the saints. He early showed himself to be an independent thinker. After occasional participation in debates between Luther and the German theologian Johann Eck (1486-1543), in Leipzig, he pursued intensive literary studies at the monastery of Beuditz at Weissenfels (1519–20). There he developed, especially under the influence of mysticism, his own view of the Reformation. From an action-hungry conspirator in local burgher plots he became a reformer who began to see the work inaugurated by Luther as a fundamental change in both ecclesiastical and secular life, and therefore as a revolution. He henceforth judged Luther's position by this criterion.

In Zwickau (1520–21), Müntzer prospered as a pastor in the socially tense condition that existed between the upper classes and early miners' guilds. In this work, he sided with the common people, who seemed to him to be the executors of the divine law and will on earth. More and more he found himself opposed to Catholic practices as well as to the reform concepts of the Lutherans. More and more he found himself holding to the sectarian view that true authority lay in the inner light given by God to his own, rather than in the Bible, a view taught by Nikolaus Storch, a leader of a radical reform group known as the "Zwickau prophets." Driven away from Zwickau, Müntzer sought on trips to Saaz (Žatec) and Prague (1521) to win to his viewpoint the Taborites, a Bohemian group that followed the teaching of Jan Hus, a 15th-century reformer. At Nordhausen (1522), Müntzer became fully aware of his opposition to Luther, where, in a struggle against Luther's supporters, his theological differences of opinion with them became more pronounced. For the first time it was the Lutherans who were to effect his expulsion from a city.

Although he began his religious revolt by following Luther's theological doctrines, Müntzer soon went his own way. Believing that teachings came from the spirit, he placed them in opposition to the Lutheran doctrines of justification (justification by faith alone) and of the authority of Scripture (Scripture as the exclusive source of divine truth). The revolutionary aspect of Müntzer's theology lay in his linking his concept of the inevitable conquest of the anti-Christian earthly government with the thesis that the common people themselves, as the instruments of God, would have to execute this change. He believed that the common people, because of their lack of property and their unspoiled ignorance, would disclose the will of God.

Müntzer arrived in Halle at the end of 1522, travelling by way of Weimar and Wittenberg. By his preaching in Glaucha, he won numerous disciples. Here he may also have met his later wife, the nun Ottilie von Gersen. From this union there were two children. Before Easter of 1523, Müntzer found employment as pastor of a Saxon community in Allstedt, near the Mansfeld mining area.

Important writings and radical reform activities

His most important religious, liturgical, and political writings originated here. They included *German Church Office, German-Protestant Mass, Protestation or Defense . . . Regarding the Beginning of the True Christian Faith and Baptism, Of Written Faith,* and *Precise Exposure of False Belief.* Here too, he drafted a speech, "Motivation for Defense," and delivered his "Princes' Sermon," in which he unsuccessfully tried to mold those princes of Saxony who were inclined toward Reformation principles into a league for resistance against what he called the

anti-Christian nobility and for the protection of his own preaching. Built upon the idea of "Christian unification" and also as a self-defense organization, the Allstedt alliance originated in 1524 and remained the centre of his doctrine until the fall of 1524, when he left Allstedt.

In Mühlhausen he organized the working classes into a group called the "Eternal Covenant of God" and lent a resolute character to the innerstate movement (based on the rule of the elect who are known by their inner experience of conversion) that had lasted since 1523. After another expulsion he went to Nürnberg, where his main political writings were published. He then went on to Hegau and Klettgau, the area where the Peasants' Revolt (an abortive revolt in 1524–25 against the nobles over rising taxes, deflation, and other grievances) was beginning, and stayed through the winter in Griessen. His experience with the rising insurrection impelled him to go back to Mühlhausen, which became the centre of the middle German revolt (after the overthrow of the governing council and the formation of what the insurgents called an "eternal council" in March 1525). Following Müntzer's dogmatic program, the common people triumphed in April–May 1525 over the religious and civil authorities. Cities and even some of the lesser nobility joined the alliance. Müntzer and his followers lent determination and consistency to the revolt. They were not, however, capable of overcoming the local and regional narrow-mindedness of the people. In the Battle of Frankenhausen, May 15, 1525, they were defeated by the superior strength of the princes.

Out of egoistical, material, and provincial motives, Müntzer dismissed resistance to his movement as a revolt against God. He believed that only if the common people were to realize the law of God within themselves, and place the interests of the group above those of the individual, would they be capable of demonstrating the will of God externally for the transformation of society. Müntzer's work was mainly concerned with the religious and ethical training of the peasants, but in expecting them to comprehend his concept of a future society without social and legal distinctions, he demanded too much. During the rebellion, Müntzer attempted to relate the battle of the peasants, tradesmen, and commoners about immediate concerns with that of the liberation of all Christendom and adapted himself to the everyday concerns of the various groups. The collapse of the revolt seemed to him to be the judgment of God on the as yet unpurified people, but not synonomous with the defeat of his idea of a new society. Müntzer was taken prisoner, tortured, and on May 27, at the princes' camp at Mühlhausen, was tried and executed.

Symbolism of Müntzer

After the bloody suppression of the Peasants' Revolt, Müntzer's program became a symbol for revolutionary movements, in which there would be a free people, which tolerated no governmental authority and itself wielded the sword. According to some scholars, especially Marxists, the revolutionary elements of Müntzer's thought were suppressed by those who opposed his views; and his character was hidden under a thick web of falsifications, fiction, and vulgarities, allowing only an apparition, the devil incarnate, to remain.

Even 100 years after his death, Müntzer's works appeared on the papal list of forbidden books. The Müntzer legend, which dates back to Luther, had preserved itself with great tenacity and could not be suppressed even by the 18th-century rationalistic Enlightenment. Not until the 19th century did Müntzer become an object of scholarly investigation. His personality and work were reinterpreted under the influence of the French Revolution, Hegelian concepts of history as a dialectic of opposing ideas, and democratic political movements of the mid-19th century, as well as the Marxist view of history. Marxist historical writing treats Müntzer as a starting point in the revolutionary tradition that led through the class struggles of the 19th and 20th centuries to Socialism, in which it sees Müntzer's basic concerns fulfilled. Efforts have been made in some Protestant historical writing to bring the heretic Müntzer back into the church. As an example, Mennonite historical writing has preserved an interpretation of Müntzer that is different from most Protestant and Roman Catholic views. Thus, the picture of Müntzer remains a matter of controversy, due to the markedly political relevance of Müntzer's ideas.

BIBLIOGRAPHY. *Schriften und Briefe von Thomas Müntzer* (1968), is the first critical complete edition that comprises all known written works by Müntzer. Critical text editions are *Politische Schriften* (1950), with commentary by CARL HINRICHS; and *Politische Schriften, Manifeste, Briefe 1524/25* (1970), with preface and commentary by MANFRED BENSING and BERND RUEDIGER. Müntzer's reformation concept and revolutionary action, discussed by M.M. SMIRIN in *Die Volksreformation des Thomas Münzer und der grosse Bauernkrieg*, 2nd ed. (1956; orig. pub. in Russian, 1947), strongly stimulated the Marxist Müntzer research. Luther's territorial reformation and Müntzer's people reformation are examined by ALFRED MEUSEL in *Thomas Müntzer und seine Zeit* (1952). MANFRED BENSING, *Thomas Müntzer und der Thüringer Aufstand 1525* (1966), deals with the zenith and the conclusion of the life of Thomas Müntzer in the German Peasants' War. ERNST BLOCH, *Thomas Müntzer als Theologe der Revolution* (1921, reprinted 1960); ANNEMARIE LOHMANN, *Zur geistigen Entwicklung Thomas Müntzers* (1931); CARL HINRICHS, *Luther und Müntzer, ihre Auseinandersetzung über Obrigkeit und Widerstandsrecht* (1952); and WALTER ELLIGER, *Thomas Müntzer* (1960), deal with Müntzer's ideological development. The presently known biographical knowledge and contemporary images of Müntzer are found in MANFRED BENSING, *Thomas Müntzer* (1965). MAX STEINMETZ, *Das Müntzerbild von Martin Luther bis Friedrich Engels* (1971), examines revolutionary Müntzer traditions as well as the falsification of the Müntzer image in the 16th to 19th centuries. All the above works are in German. Very little has been published on Müntzer in English. See, however, ERNEST GORDON RUPP, "Thomas Müntzer: Prophet of Radical Christianity," *Bulletin of the John Rylands Library*, 48:467–487 (1966); and "Thomas Müntzer, Hans Huth and the 'Gospel of All Creatures'," *ibid.*, 43:492–519 (1961); and ERIC W. GRITSCH, *Reformer Without a Church: The Life and Thought of Thomas Muentzer, 1488?–1525* (1967).

(M.Be.)

Muscle Contraction

Motion is often said to be a characteristic of all living things. The more closely various living organisms and cells are examined, the more evidence accumulates to support this idea. The most obvious and striking movements, such as walking and running by higher animals, are produced by the action of muscles—organs composed of contractile tissue—which make up a large part of the bodies of these animals. There are, however, many examples of nonmuscular movements of cells and of primitive organisms.

For a comparative discussion of the muscle systems of animals, see MUSCLE SYSTEMS. The muscle system of man and diseases of muscles are dealt with in MUSCLE SYSTEM, HUMAN; and MUSCLE DISEASES.

This article is divided into the following sections:

I. General features
  Primitive contractile systems
  Muscle in higher animals
II. The nature and properties of striated muscle
  Structure and organization
  The neuromuscular junction: nerve–muscle communication
  Mechanical properties
  Energy transformations
  Metabolic pathways producing ATP
  Molecular components
  Molecular mechanisms of contraction
III. The nature and properties of cardiac muscle
  Structure and organization
  Mechanical properties
  Metabolic pathways
IV. The nature and properties of smooth muscle
  Vertebrate smooth muscle
  Invertebrate smooth muscle

I. General features

In all types of primitive motility described below, there are at least two points of similarity to muscle contraction —the requirement for an energy-yielding reaction and the presence of contractile protein.

PRIMITIVE CONTRACTILE SYSTEM

Cilia and flagella. Motion can be caused by the action of cilia or flagella, microscopic hairlike structures that extend out of the cell surface.

Cilia occur in large numbers on cell surfaces and move in a coordinated way. The pattern of motion during the beat of an individual cilium is fairly simple. During the power stroke, the cilium is slightly curved and rigid as it pushes through the surrounding medium, an action somewhat like the arm stroke of a swimmer; the cilium then becomes limp and remains flexible, thus reducing friction as it goes through the recovery stroke and returns to its original position. Unicellular organisms such as the paramecium, a protozoan that lives in freshwater ponds and streams, propel themselves by the action of cilia. But there are ciliated cells within the vertebrate body as well. Their cilia cause fluid and mucus to flow along the surface of interior passages such as the lining of the respiratory tract.

Flagella (singular, flagellum) are structurally similar to cilia, except that they are longer (sometimes up to 50 times longer) than cilia and usually number only one or two per cell. The stroke pattern by which a flagellum causes motion is somewhat more complicated than that of a cilium. In a flagellated cell, waves propagated along the length of the flagellum cause a helical, or spinning, movement of the cell as it is propelled forward. Sperm cells of most higher organisms move by the action of flagella. Many types of unicellular algae and protozoans use flagella in swimming through the water in which they live.

Electron-microscopic examination of both a cilium and a flagellum shows that each contains a regular pattern of tubules extending along their length. Cross sections of cilia and flagella reveal an outer ring of nine pairs of tubules and a central pair of tubules. Each tubule is made up of filaments that consist of a string of globular subunits. Like the contraction of a muscle, the movement of a cilium or a flagellum requires energy, which is obtained from the breakdown of the chemical compound adenosine triphosphate (ATP). The action of these cellular extensions does not depend on some process that takes place inside the cells, since either of them, when separated from a cell, will beat as usual upon the addition of ATP to the surrounding medium.

Some types of bacteria have flagella that differ somewhat from the type described above. Their motion seems to depend completely on a cellular particle called the basal granule, to which the flagellum is attached. Such flagella do not move when detached from their cells and provided with ATP.

Movement on "false feet"

Amoeboid motion. Cells capable of amoeboid motion have flexible surfaces, or boundaries. Motion occurs as an extension of the cytoplasm, called a pseudopod ("false foot"), flows outward and deforms the cell surface, and is followed by the rest of the cell. Many pseudopods may be formed at the same time, and their action does not seem to be coordinated.

The cytoplasm, or cell fluid, inside amoeboid cells is in two forms. The inner part, called the endoplasm, is thin and watery. It flows forward inside the advancing pseudopod, and at the leading end it flows in a fountainhead pattern, changing into a stiffer outer gellike form called the cortical layer. The cortical layer flows to the base of the pseudopod, where it converges toward the centre and is transformed back into endoplasm. It seems likely that the force responsible for amoeboid motion is generated by either the endoplasm-to-gel transformation at the leading end of the pseudopod or by the gel-to-endoplasm transformation at the base.

Amoeboid motion is, of course, characteristic of the amoeba, a one-celled protozoan, but it is also found in other cells. Slime molds, for example, move by amoeboid motions, as do certain cells in higher animals (*e.g.*, embryonic cells, types of white blood cells). A chemical compound in slime molds is very similar to actin, one of the contractile proteins in the muscles of higher animals. (In muscle contraction, actin interacts with myosin, the other contractile protein, and at one stage they combine with each other.) Experimentally, the actin-like compound from slime molds can combine with the myosin from rabbit muscle, suggesting a similar contractile mechanism for both types of organisms.

Even in stationary cells there is always some turbulence or slow streaming of the cytoplasm. This motion, called cyclosis, is important for the transport of various chemicals within the cell. It is like amoeboid motion, except that it does not deform the boundary of the cells and does not result in cell movement.

MUSCLE IN HIGHER ANIMALS

In higher animals such as vertebrates, most movements are caused by skeletal, cardiac, and smooth muscles. The skeletal muscles, which have visible stripes, or striations, when viewed with the microscope, usually contract rapidly and are under the voluntary control of the animal. Cardiac muscle, which makes up the heart, is unique in being striated but not under voluntary control; it continues pumping automatically throughout life. Smooth muscle, which does not have microscopic striations, usually contracts quite slowly and often undergoes rhythmic pulsations. It comprises the muscles of the intestinal walls and other parts of the internal organs (viscera). The action of most smooth muscle in vertebrates is not under voluntary control.

More is known about the operation of muscle at the molecular level than about any other tissue. One factor that has contributed to this knowledge is the close relationship between the structure and function of muscle—the study of the structure of muscle has added enormously to understanding its operation.

The levels of muscle structure

The structure of muscle can be considered on many levels of size ranging from the whole muscle, which can be examined with the unaided eye, to the arrangement of individual molecules in the contractile proteins, some of which are in the nanometre (1 nm = 10^{-9}, or one-billionth, metre) size range and must be magnified hundreds of thousands of times to be visible.

II. The nature and properties of striated muscle

STRUCTURE AND ORGANIZATION

Whole muscle. Striated, or striped muscle constitutes a large fraction of the total body weight. The movement of the limbs depends on the contraction of striated muscles, and posture is maintained by them. Both ends of most striated muscles are attached to the skeleton and thus are often called skeletal muscles. They are attached to the bones by tendons, which have some elasticity provided by the proteins collagen and elastin, the major chemical components of tendons.

Each striated muscle has blood vessels and nerves associated with it. The vessels transport blood to and from the muscle, supplying oxygen and nutrients and removing carbon dioxide and other wastes. The signals that initiate contraction are sent from the central nervous system to the muscle via the motor nerves. Each muscle also has important sensory structures called stretch receptors, which monitor the state of the muscle and return the information to the central nervous system. Stretch receptors are sensitive to the velocity of the movement of the muscle and the change in length of the muscle. They complete a feedback system that allows the central nervous system to assess muscular movement and to adjust motor signals in light of the movement.

The muscle fibre. Living muscle is composed of many long, cylindrical-shaped fibres from 0.02 to 0.08 millimetre in diameter (Figure 1). In some muscles the fibres run the entire length of the muscle, up to several centimetres long. In others a tendon extends along each edge, and the fibres run diagonally across the muscle between the tendons. Considerable variation can be found among the different skeletal muscles, the actual arrangement of the fibres depending on the function of the muscle.

The fibres in living muscle appear transparent when viewed in an ordinary optical microscope. Special types of microscopes, however, reveal a high degree of organization within the living fibre, a series of alternately dark

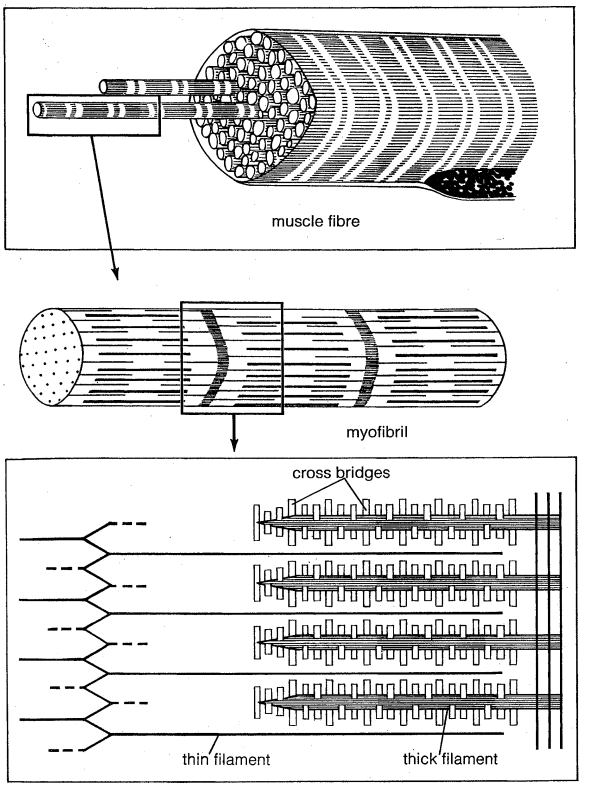


Figure 1: The spatial relationships of a striated muscle fibre, myofibrils, and sarcomeres (see text).
From E.P. Hammel, V.M.D., and K.M. Friedenberg, J. Vet. Med., M.R.C.V.S., in *Equine Medicine and Surgery*, ed. by E.J. Catcott (1972); American Veterinary Publications

The appearance of muscle fibres

and bright bands. Each band extends across the fibre perpendicular to the length of the fibre. This cross-banded pattern can also be seen in chemically prepared and stained sections of muscle. (The structures responsible for the banding pattern are dealt with in a later section; see below *The myofibril.*)

Microscopic examination of stained sections of fibres reveals many features in addition to the cross bands. Each fibre is surrounded by a rather complex, multilayered structure, the sarcolemma. The outermost layer is a fine network of fibrils, which, at the ends of the muscle, extend into the tendons and form the structural link with them. The next layer of the sarcolemma is a foundation, or basement, membrane. The innermost layer is a plasma membrane like the ones that surround most cells; in electron micrographs it shows the usual form of two dark lines with a light one between them. This three-layered plasma membrane, which consists of lipid (fat) and protein molecules, maintains the electrical potential that plays a major role in stimulating muscle contraction.

Sarcoplasm is the special name for the cytoplasm of a muscle fibre. It is a water solution containing adenosine triphosphate (ATP), compounds called phosphagens, and all the enzyme, reactant, intermediate, and product molecules involved in many reactions, including one of the sequences of reactions (glycolysis) that produce ATP. The major ions (charged components) present in the sarcoplasm are magnesium (Mg^{++}), potassium (K^{+}), and sodium (Na^{+}). During contraction an appreciable amount of calcium ion (Ca^{++}) is also present in the sarcoplasm. Large structural units are suspended in the sarcoplasm as well.

Unlike most other cells in the body, including smooth or cardiac muscle cells, each striated muscle cell, or fibre, contains many nuclei. This is the result of the fusion of singly nucleated cells that occurs during the embryological development of striated muscle. After fusion, the cells never again divide.

Another prominent feature of muscle fibre is the presence of cell components called mitochondria. They contain the enzymes involved in the Krebs cycle and oxidative phosphorylation, two series of chemical reactions that produce the ATP necessary to muscle contraction. These reactions produce large amounts of ATP at a low rate and can fulfill the energy requirements of continuously active muscle. Muscles that operate at a steady, low level of activity are reddish, as a result of cytochromes, molecules involved in oxidative phosphorylation, and myoglobin, an oxygen-carrying molecule in the sarcoplasm. Striated muscles that work in bursts of activity, however, have fewer mitochondria and depend more heavily on reactions that do not require oxygen to make ATP. Such muscle fibres contain fewer molecules of cytochromes or myoglobin and constitute white muscles.

Prominent subcellular features

Small darkly stained granules can be seen in the sarcoplasm of muscle prepared for viewing with the electron microscope. The granules are aggregates of glycogen, a storage form of carbohydrate. The breakdown of glycogen and the metabolism of the individual units of carbohydrate through glycolysis, the Krebs cycle, and oxidative phosphorylation are important sources of ATP.

The myofibril. Electron micrographs of very thin sections of muscle fibres reveal groups of filaments oriented with their axes parallel to the length of the fibre. Two sizes of filaments, thick ones and thin ones, are present. Each array of filaments, called a myofibril, is shaped like a cylindrical column. Along the length of each myofibril alternate sets of thick and thin filaments (Figure 1) overlap, or interdigitate, presenting alternate bands of dark regions (with thick filaments and overlapping thin ones) and light regions (with only thin filaments). Within a fibre all the myofibrils are in register, so that the regions of similar density lie next to each other, giving the fibre the characteristic striated appearance it shows in the phase-contrast microscope and in stained sections viewed with a light microscope.

Each myofibril is about one or two micrometres, or microns (μ; $\mu = 10^{-6}$ metre), in diameter and extends the entire length of the muscle fibre. The number of myofibrils per fibre is quite variable. At the end of the fibre the myofibrils are attached to the sarcolemma.

Forty to 80 nanometres usually separate adjacent myofibrils in a fibre. This space contains two distinct systems of membranes involved in different aspects of the activation of muscle contraction (Figure 2). One system is a

From *Journal of Cell Biology* (1965)

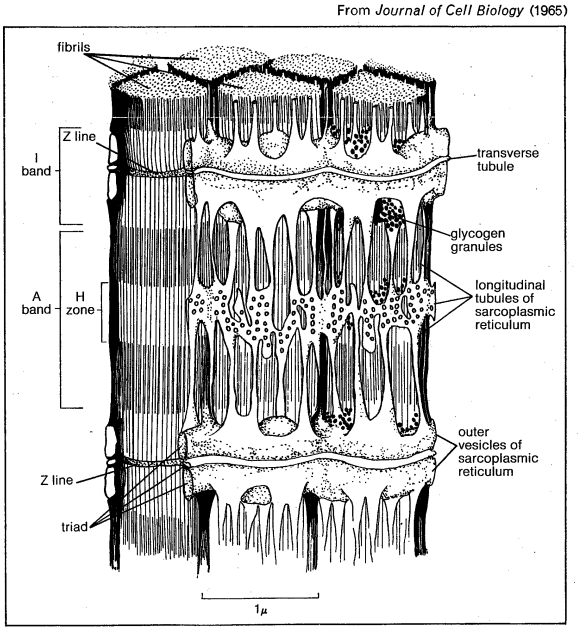


Figure 2: Ultrastructure of a group of myofibrils, showing the sarcoplasmic reticulum and transverse tubules, which constitute the two membrane systems within a muscle fibre.

series of channels that open through the sarcolemma to the extra-fibre space. These channels are called the transverse tubules because they run across the fibre; in other words, they lie on circular planes perpendicular to the fibre length. The transverse tubular system is a network

of interconnecting rings, each of which surrounds a myofibril. It provides an important communication pathway between the outside of the fibre and the myofibrils, some of which are located deep inside the fibre. The exact spatial relationship of the tubules to the filaments in the myofibril depends on the species of animal.

Transverse tubules and the sarcoplasmic reticulum

The other membrane system that surrounds the myofibril is the sarcoplasmic reticulum, a series of closed sac-like membranes. Each myofibril has a cufflike layer of sarcoplasmic reticulum surrounding it and extending between the transverse tubules. Near the transverse tubule, relatively simple enlarged sacs called terminal cisternae form from the sarcoplasmic reticulum. The walls of the transverse tubule and the terminal cisternae (singular cisterna) lie close together but with no direct physical continuity between them. In most vertebrates, each transverse tubule has two cisternae closely associated with it, forming a three-element complex called a triad. In fish and crustaceans, however, only one cisterna is associated with each transverse tubule, thus forming a dyad. The parts of the sarcoplasmic reticulum farther away from the transverse tubule are called longitudinal tubules; these narrow channels run parallel to the length of the myofibril. The sarcoplasmic reticulum controls the level of calcium ions in the sarcoplasm. The terminal cisternae apparently are the sites from which the calcium ions are released when the muscle is stimulated, and the longitudinal tubules are the sites at which calcium ions are bound again and effectively removed from the sarcoplasm.

The myofilament. As mentioned earlier, the myofibril is a column-like array of filaments. Figure 2 is a diagram of a longitudinal section through a myofibril. In the centre of the I band—the light band, of low density—there is a prominent dense line called the Z line. The area between two Z lines, a sarcomere, can be considered to be the primary structural and functional unit directly responsible for muscle contraction. The myofibril can thus be thought of as a stack of sarcomeres.

The A band, which contains thick filaments partly overlapped with thin filaments, appears dark. At very high magnification, small bridgelike structures can be seen on the thick filaments extending toward the thin filaments in the overlap region (Figure 1). They are called cross bridges and are believed to be responsible for the movement and force developed during contraction. In the middle of the A band, where only thick filaments are present, is a region called the H zone; it looks somewhat lighter than the overlap region of the A band. Also in the A band is a narrow, lightly stained region that contains bare thick filaments without cross bridges and is called the pseudo-H zone. In the centre of the A band is a narrow, darkly stained region called the M band, in which occur fine bridges between the thick filaments. The bridges differ from the cross bridges between the thick and thin filaments and are in fact composed of an entirely different protein.

Cross bridges to the thin filaments

If cross sections of the myofibril at different levels of the sarcomere are examined in the electron microscope, the filaments can be seen end-on, and the three-dimensional nature of the lattice of filaments can be appreciated (Figure 3). The I band contains only thin filaments, with a diameter of six to eight nanometres. In the A band, in the overlap region, the thin filaments appear with thick ones (diameter of 12 nanometres) in an extremely regular pattern, or lattice. In vertebrates, the thick filaments are arranged in a hexagonal lattice, and the thin ones are located at the trigonal positions of the lattice; in other words, a thin filament is located in the centre of each triangular-shaped area formed by considering three thick filaments to be the corners of a triangle. Sections through the H zone contain only thick filaments arranged in the same hexagonal pattern they form in the overlap region. In the M band, the hexagonal array of thick filaments can be seen with M bridges running between them.

The thick filaments are composed of about 200 molecules of the protein myosin. Myosin has two outstanding roles—a structural one, as the building block for the thick filament; and an enzymatic one, as the catalyst of the breakdown of ATP during contraction. The individual myosin molecule has a molecular weight (based on the weight of a hydrogen atom as 1) of about 500,000; it contains two major protein chains and a few small ones, the entire molecule being about 160 nanometres in length and asymmetrically shaped. The rodlike tail region, about 120 nanometres long, consists of two chains of protein wound together to form a helix (coil). At the other end of the molecule, the two protein chains form two globular headlike regions which have the ability to combine with the protein actin (see below *Molecular components*).

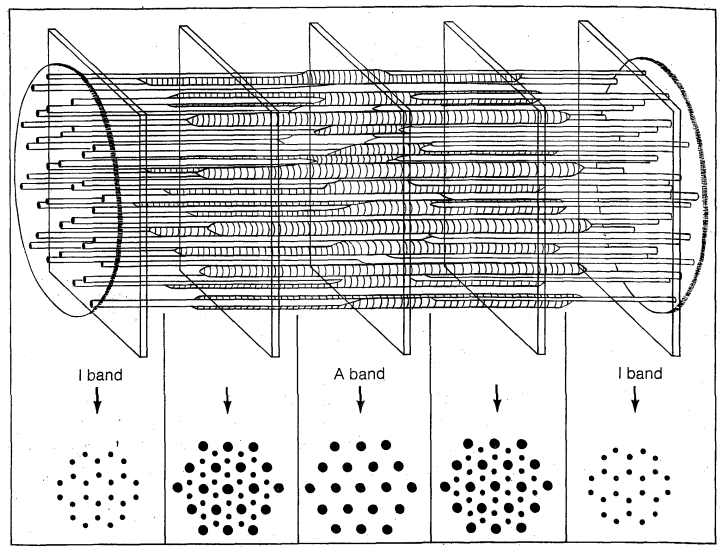


Figure 3: Muscle fibril, showing perspective model of the sarcomere.

From H.E. Huxley, *The Contraction of Muscle* (Copyright © 1958); Scientific American, Inc., all rights reserved

In the centre of the thick filament the molecules are assembled in a tail-to-tail fashion. Along the rest of the filament they are arranged head to tail. The tail parts of the molecules form the core of the filament; the head portions project out from the filament. This kind of arrangement explains, on a molecular basis, the appearance of the thick filaments in the A band of sectioned muscle. The cross bridges are actually the globular head regions of myosin molecules extending outward from the filament, and the smooth pseudo-H zone is the region of tail-to-tail aggregation, in which there are only tails and no heads.

Molecular arrangement in thick filaments

During contraction the thin filaments move into the A band, and the overlap of the thick and the thin filaments increases. The direction of movement at the ends of a thick filament is different. If a longitudinal section of the sarcomere is considered, the thin filaments on the left side of the A band would move to the right into the A band, and the filaments on the right of the A band would move to the left into the A band. The basis for the directionality of the motion partly results from the structural polarity of the thick filament itself, since, at opposite ends of the filament, the myosin molecules themselves are oriented in opposite directions.

The precise three-dimensional arrangement of the cross bridges projecting from the thick filament cannot be seen easily in electron micrographs but can be determined from X-ray diffraction study of living muscle (Figure 4). Two bridges project from opposite sides of the filament every 14.3 nanometres along the length of the filament. Each successive set of bridges is located at a position rotated 120° farther around the filament. The pattern of six bridges—three sets of two bridges—repeats itself every 42.9 nanometres along the thick filament.

Proteins of the thin filaments

The thin filaments contain three different proteins—actin, tropomyosin, and troponin. The actin in the thin filaments is in the form of beadlike units in two long strands, twisted around each other to form a helix. Each of the beadlike units is about 5.5 nanometres in diameter and consists of a single molecule of actin. There are 13 of these units in each complete turn of one strand around

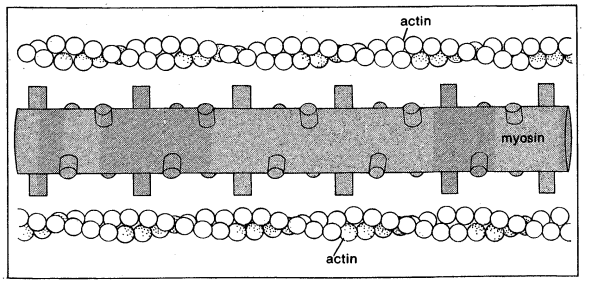


Figure 4: Arrangement of the cross bridges projecting from the thick filaments and the globular units of actin in the thin filaments of a muscle fibril.
From *Hospital Practice* (June 1970)

the helix. The double-stranded structure repeats itself every 36.5 nanometres along the thin filament. Actin is believed to be directly involved in the process of contraction since the cross bridges can become attached to it. Although it is not apparent from the shape of the actin units, these molecules are polarized.

Tropomyosin, a rod-shaped molecule about 40 nanometres long, with a molecular weight of about 70,000 combines with actin and is believed to lie in the groove between the two strands of actin in the thin filament. Troponin, a globular protein, with a molecular weight of about 50,000, does not combine directly with actin but, instead, forms a complex with tropomyosin, which attaches to filaments of actin. Tropomyosin lies along the entire length of the thin filament, and a troponin molecule is located every 40 nanometres along the filament. Troponin and tropomyosin are both involved in the regulation of contraction and relaxation of muscles.

The thin filaments do not simply run through the Z line; they begin on one side of the Z line just opposite the space between two thin filaments located on the opposite side of the Z line. The exact structure of the Z line has not yet been completely established, but it is thought that four filaments extend from the end of each thin filament to four different thin filaments on the other side of the Z line. This type of arrangement would give the Z line a three-dimensional structure resembling a series of pyramid-like units.

The fine M bridges are composed of a protein that has not yet been fully characterized, but it has been determined that it is not myosin, actin, tropomyosin, or troponin. The M bridges run between sites on the thick filaments at which myosin molecules are joined tail to tail. It may be that the M bridges keep the filaments in register during contraction.

THE NEUROMUSCULAR JUNCTION: NERVE–MUSCLE COMMUNICATION

Nervous control of muscle action

Nature of the communication. The activity of the nervous system is largely responsible for the control of muscle action. In vertebrates, nerve impulses travel along the surface of the motor nerve cells as waves of electrical changes, or depolarization and subsequent repolarization, and cause the contraction of the voluntary muscles.

Each nerve cell and muscle fibre is surrounded by its own membrane, but between the membrane of the nerve ending and the membrane of the muscle fibre is a gap filled with a solution of ions in water. When the electric signal reaches the nerve ending, it must cross the gap in order to stimulate the muscle. It does so by chemical rather than electrical transmission. Synapse is the general name for the site of this type of chemical communication between separate, electrically excitable cells; the special name for the synapse between nerve and muscle is the neuromuscular junction.

Anatomy of the junction. The vertebrate muscle fibres that contract in response to a single motor nerve impulse are striated fibres called "twitch" fibres. Each fibre has one nerve ending associated with it, forming one neuromuscular junction per muscle fibre (Figure 5B). Near the neuromuscular junction, the motor nerve branches and loses its fatlike, myelin covering but is partly surrounded by Schwann cells, which elsewhere cover the neuron and produce myelin. The nerve terminal—the end of the nerve cell—lies close to a specialized region of muscle fibre called the end plate. The muscle membrane at the end plate is arranged in a folded pattern: one long fold runs parallel to the long axis of the muscle fibre. The nerve terminal lies in this fold. A set of small folds, junctional folds, in the muscle membrane run perpendicular to the long one (Figure 5B). Within the gap that lies between the nerve membrane and the folded muscle membrane (Figure 5C) lies the basement membrane, which does not connect the two types of cells. In the nerve terminal are mitochondria, which produce the ATP used in the energy conversions of the cell, and structures that are more directly involved in the transmission process: vesicles, or bubble-like structures, about 50 nanometres in diameter, which congregate opposite the junctional folds in the muscle membrane. The vesicles contain a certain amount, or quantum, of acetylcholine, the chemical transmitter substance in vertebrate motor nerves.

The nerve terminal and the end plate

The transmission process. The function of the neuromuscular junction is to transfer the information from the nerve to the muscle so that muscle contraction results. This transfer is accomplished in several steps. The nerve impulse arrives at the nerve terminal and accelerates the secretion of the transmitter substance, acetylcholine, from the nerve endings into the space next to the junctional folds. (It has been suggested that transmitter release occurs as a large number of vesicles come in contact with activated areas of the nerve membrane; each vesicle joins with the membrane and empties its contents into the gap.)

Once released, the acetylcholine molecules flow across the gap to the junctional folds in the muscle membrane, where they act on special receptor sites in the end-plate membrane and produce an end-plate potential. The permeability of the membrane to certain small ions is increased, and depolarization, or decrease in the electrical potential, across the end-plate membrane results. After acetylcholine has acted on the receptor sites, it is destroyed by the action of acetylcholine esterase, an enzyme located in the muscle end-plate region. This is a very important process, because it turns off the nerve-to-muscle signal and allows the muscle to relax. Choline, which is a product of the breakdown of acetylcholine, can be reabsorbed by the nerve ending and used to make more of the active transmitter.

Meanwhile, the end-plate potential gets weaker and fades out as it travels over the end-plate region; *i.e.*, it is nonregenerative. (An analogy would be the spreading and weakening waves that appear on a pond surface after a pebble is thrown into it.) This nonregenerative end-

From B. Katz, *Nerve, Muscle and Synapse* (Copyright 1966); reproduced by permission of McGraw-Hill Book Co., Inc.

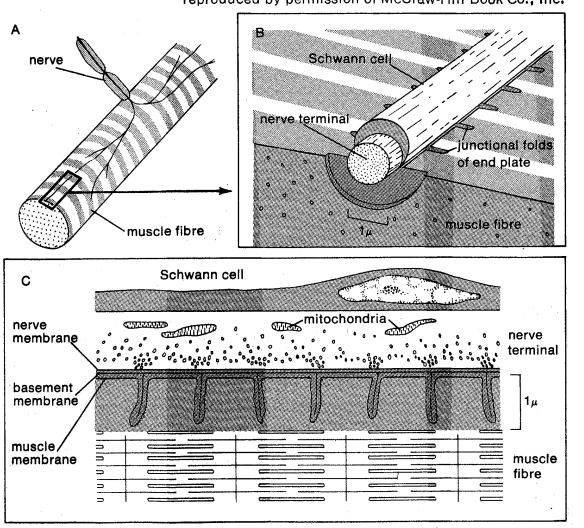


Figure 5: (A) Motor nerve ending and muscle fibre. (B) Enlarged section of small rectangle in (A), showing end plate and nerve terminal. (C) Longitudinal section through the neuromuscular junction.

plate response, however, initiates a different type of potential change across the sarcolemma, the membrane covering the rest of the muscle fibre. This electrical impulse, or action potential, travels over the entire surface of the muscle fibre and leads to contraction.

The action potential

The muscle action potential differs from the end-plate potential in two important ways. First, unlike the end-plate potential, the action potential is self-regenerating, which means that the wave of depolarization does not fade out but is maintained at the same magnitude as it travels along the muscle fibre. Second, the magnitude of the action potential, or degree of depolarization of the sarcolemma, is the same regardless of the strength of the nerve stimulus as long as it is above a certain minimum, or threshold, value. This is often referred to as the "all-or-none" character of the action potential. The special aspects of the end-plate potential and the action potential described above depend on the special properties of the flow of ions across membranes and are typical of twitch muscle fibres.

In other types of muscle, the events that follow nerve stimulation are somewhat different. Slow, or tonic, muscle fibres in some frog muscles, for example, require several stimulations to produce contraction; a single nerve stimulus is not sufficient. These muscle fibres do not have a mechanism for producing self-regenerating action potentials and do not show all-or-none behaviour. Tonic fibres, however, have many neuromuscular junctions distributed along the fibre surface, and the depolarizations produced at these junctions accumulate as they travel along the fibre surface. Thus the size of the depolarization of the muscle can have a variety of values depending on the strength of the nerve stimulation.

In the motor nerve and voluntary muscles, both fast and slow, of vertebrates, the nerve stimulus results in muscle contraction. Neuromuscular transmission, however, can communicate other messages in other systems. One example is the special nerves and muscles in the legs of crustaceans; nerve impulses cause the muscles to relax rather than contract. The nerve fibres in such a case are called inhibitory fibres; they work along with excitatory fibres, which cause contraction, and the two types coordinating the movements of the crab's legs.

Neuromuscular transmission also varies according to the type of transmitter substance. As mentioned above, acetylcholine is the transmitter in the well-studied vertebrate motor nerve–voluntary muscle system. Some of the autonomic nerves—those taking part in involuntary functions and responses of the body—are likewise mediated by the release of acetylcholine. Such autonomic nerves are called cholinergic nerves on the basis of their transmitter substance. Some examples of the muscular responses initiated by these nerves are: dilation of blood vessels in the skin, contraction of the walls of the digestive tract, and constriction of the pupils of the eye. Other autonomic nerves, however, act by releasing epinephrine (adrenaline) and norepinephrine; they are called adrenergic nerves. The actions induced by these nerves generally oppose those induced by cholinergic nerves: constriction of the blood vessels in the skin, relaxation of the smooth muscles of the digestive tract, and dilation of pupils. The transmitter substances have not yet been identified in some nerve–muscle systems, for example in the asynchronous flight muscles of insects and the inhibitory nerve fibres of crustaceans.

Cholinergic and adrenergic nerves

MECHANICAL PROPERTIES

The physical aspects. Vertebrates are able to move about and to exert and bear forces because of the contraction of the striated, or skeletal muscles, which constitute a large part of the body. These activities usually involve several structures operating in different ways. The skeleton with the muscles attached to it operates as a lever system. When a muscle shortens, it moves the joints that it spans. In addition, in coordinated movement usually several muscles contract in different ways. As some muscles shorten, others develop a force while at a fixed length, and still others may be lengthened by an external force even as they contract.

The force that a muscle develops is a "pulling" force, never a "pushing" force. The direction of movement, however, is determined by the size of the load exerted on the muscle. If the load is small enough, the muscle can shorten and produce a pulling motion (an isotonic condition). If the load is just equal to the maximum force the muscle can develop, the muscle will remain stationary (an isometric condition). An even larger load will stretch the muscle; the substantial force it exerts could be described as a pulling force that does not succeed in producing a pulling motion.

Factors modifying contraction

The size and the rate of the mechanical responses to stimulation, whether by a nerve in the body or by direct electrical shocks of an isolated muscle, depend on the muscle and the temperature. In a frog sartorius muscle (of the leg) at 0° C (32° F), the action potential reaches its peak of depolarization about 1.5 milliseconds (0.0015 second) after the stimulus. The first seven milliseconds after the stimulus is the latent period, during which time the tension in the muscle does not change. Until about 11.5 milliseconds the tension falls slightly, a response called latency relaxation. From 11.5 to 15 milliseconds, the tension again builds up to its initial starting level. From then on, the progress of tension development depends on the load on the muscle.

The very early tension changes require much more rapid and sensitive measuring and recording instruments than are necessary for studying other aspects of the contraction process. The latent period is the amount of time needed for the electrical signal, which appears as an action potential at the surface membrane, to be translated and travel to the contractile apparatus within the muscle fibres. The explanation for latency relaxation, however, is not so clear. It may be related to a change in shape of the sarcoplasmic reticulum. This structure makes up about 15 percent of the fibre volume and releases a large amount of calcium ions at about the time latency relaxation occurs. This ion flow may alter the shape of the reticulum and thus temporarily decrease the tension in the fibre as a whole.

Twitch and tetanus responses. Skeletal muscles respond to a single electric shock of sufficient magnitude by rapid, intense contractions called phasic contractions. If the ends of a frog sartorius muscle (at 0° C) are fixed to prevent shortening, the tension increases for about 200 milliseconds and then begins to decrease, at first rather rapidly and then more slowly. More happens during this mechanical response to a single stimulation, called a twitch, than the tension record suggests (Figure 6).

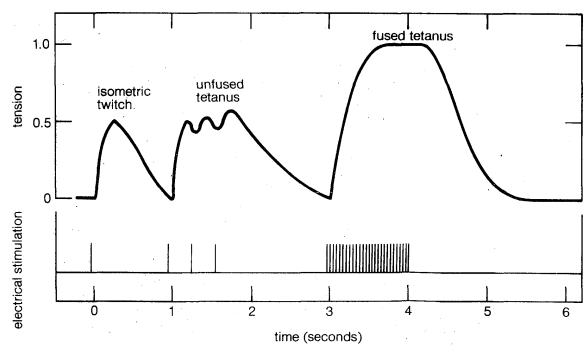


Figure 6: The effect of the rate of stimulation on the development of force in a frog muscle (at 0° C).

The mechanical response to repeated stimulation depends on the rate of the stimulation. Muscle, like other excitable tissues, has a period following its action potential during which the membrane will not respond to stimulation regardless of the strength. This absolute refractory period in the frog sartorius at 0° C lasts about ten milliseconds after stimulation. Therefore, a second pulse coming within that time span will not elicit any response. If, however, the pulses are 300 milliseconds apart, the muscle will be relaxing when the second pulse is given, and the tension will appear in waves in phase with the stimulation, causing an unfused tetanus. It is possible to stimu-

Absolute refractory period

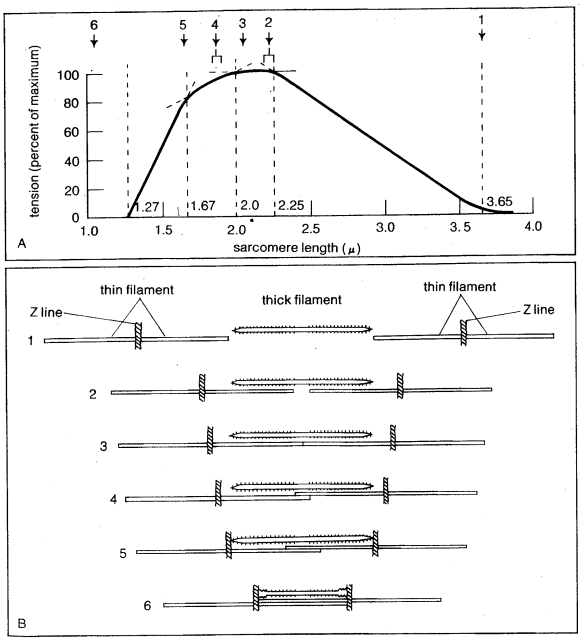


Figure 7: (A) The amount of force developed during an isometric tetanus of a single muscle fibre at various sarcomere lengths. (B) Relative positions of the thick and thin filaments at various sarcomere lengths.
From *Journal of Physiology* (1966)

late the muscle at a frequency between these extremes so that the tension developed by the muscle remains constant. This latter type of contraction is called a fused tetanus, and the rate of stimulation that produces it is called the fusion frequency. The exact rate depends upon the particular muscle and the temperature.

Usually, the maximum tetanus tension is from 1.2 to 1.8 times greater than the maximum tension during a twitch. This sizable difference has a rather subtle origin and points out the complexity of contraction even under isometric conditions. Within the muscle many elastic structures, connected in series with the contractile elements, are stretched during contraction. The attachment of the muscle fibres to the tendons at the end of the muscle, the attachment of the thin filaments to the Z lines, and even the actin–myosin cross bridge links probably contribute to this elastic component. In the frog sartorius muscle, the elasticity of the series elastic elements represents only a small percentage of the resting length of the muscle. Full maximum tension is not apparent at the end of the muscle until the contractile elements have shortened enough to stretch the elastic elements—somewhat like taking up the slack in rope before a pull on one end can be felt at the other end. In a twitch, the activity of the muscle is so brief that the contractile elements cannot extend the elastic elements completely before relaxation begins; as a result, the tension at the ends of the muscle does not reach the maximum possible level. During a tetanus, on the other hand, the activity of the contractile elements is maintained, and they can eventually shorten enough to extend fully the series elastic elements. When this has been accomplished, the maximum tension is apparent at the ends of the muscle.

Effects of series elasticity

Length–tension relationship. The force developed by a muscle, whether it is contracting or resting, is strongly dependent on the length of the muscle (Figure 7). Resting skeletal muscle does not exert any force at lengths less than the normal length of the resting muscle in the body (l_o). When resting skeletal muscle is extended somewhat beyond l_o, however, a passive force begins to assert itself. The exact length at which this passive force occurs depends on the particular muscle. This force is characterized as passive because it is developed in noncontracting or inactive muscles by the elastic elements of the muscle.

Skeletal muscles do not develop any force when they are stimulated at lengths of less than about 50 percent l_o. The amount of force developed, however, increases during isometric contractions at lengths from 50 to 100 percent of l_o. Once beyond l_o, the passive force becomes a factor. Thus, for contracting muscle at lengths greater than l_o, the total force has two components, a passive one, resulting from elasticity, and an active one, resulting from contraction. For isometric contractions at successively greater lengths beyond l_o, the active force developed during contraction decreases; at about 150 percent l_o, the muscle fails to develop any active force.

The structural basis for the dependence of the magnitude of active force on the length of skeletal muscle has been established in experiments with single fibres rather than whole muscles. With this more exact technique, details of the relationship between active force and sarcomere length have been found, and the active force has been correlated with the relative position of thick and thin filaments during contraction at each sarcomere length. It was determined that tension increases as the sarcomere length shortens from 3.65 to 2.25μ (Figure 7A, 1 and 2) because more sites for actin-myosin cross-bridge interaction become available as the overlap of the thick and thin filaments increases (7B, 1 and 2). The force is constant from sarcomere length 2.25 to 2.00 μ (7A, 2 and 3) as the thin filaments move over the bare region devoid of cross bridges in the centre of the thick filaments (7B, 2 and 3). As the sarcomeres shorten further, from 2.00 to 1.67 μ (7A, 3 to 5), the thin filaments overlap each other, preventing effective interaction with myosin in the thick filaments (7B, 3 and 5) and consequently diminishing the force. Finally, from sarcomere length 1.67 to 1.27 μ (7A, 5 to 6) the thick filaments run up against the Z lines (7B, 6) and internal resistance causes an even greater diminution of the force.

Contractions involving large movement. When a muscle shortens and lifts a constant load as it is stimulated, it is contracting under isotonic conditions (Figure 8). The

From D.R. Wilkie, *Muscle;* Edward Arnold (Publishers) Ltd.

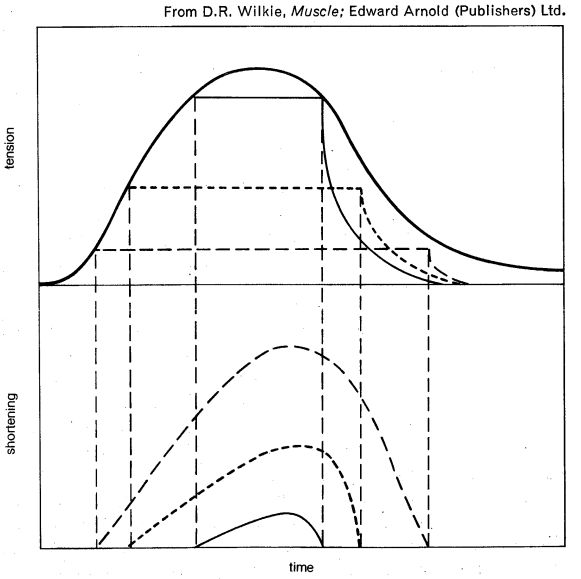


Figure 8: Force development (tension) and change in muscle length during an isotonic contraction (see text).

size of the load determines the magnitude of the time delay between the beginning of stimulation and the beginning of shortening. After stimulation starts, the force increases, just as in an isometric contraction, and, when the force is equal to the load, the muscle begins to shorten and lifts the load. The heavier the load, the longer it takes for the muscle to develop the required force and hence the delay before shortening begins. When the activity of both the muscle and the force in it begin to decline, the load stretches the muscle back to its initial length. The tension in the muscle is equal to the load during the shortening and the lengthening of the muscle, except during brief periods of acceleration as the muscle begins to move. Only after the muscle is back at its initial length does the tension begin to diminish. The size of the load also determines the velocity of shortening, and this relationship be-

Effects of the load on muscle contraction

tween load and velocity also applies to cardiac and smooth muscle. It can be mathematically described by Hill's equation:

$$V = \frac{(P_o - P)b}{P + a},$$

in which V is the velocity of shortening; P_o is the maximum force developed under isometric conditions; P is the force developed at a particular muscle length; and b and a are physical constants. A graphic interpretation of these values results in a rectilinear hyperbola called the force–velocity curve (Figure 9). From this curve can be determined both the velocity of shortening when the load is known and the force developed to overcome the load when the velocity is known.

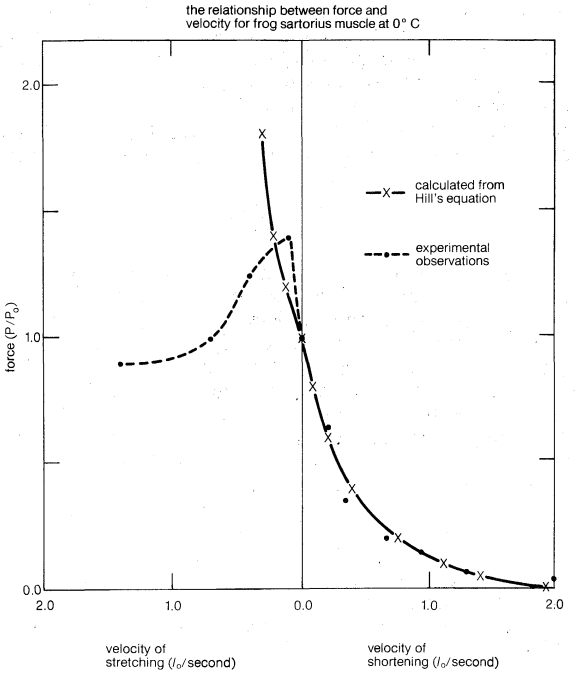


Figure 9: The force-velocity curve: relationship of the velocity of movement to the amount of force developed during tetanic stimulation (see text).

ENERGY TRANSFORMATIONS

When a chemical reaction occurs, energy is absorbed or released. In a contracting muscle, chemical reactions release energy that appears either as mechanical work or as heat. The first law of thermodynamics, or the law of conservation of energy, states that the heat and work produced must equal the energy released by the chemical reactions. The muscles that shorten and do external work liberate more energy as heat and work than do those that contract under isometric conditions and do not shorten or do external work. In light of the law of conservation of energy, this finding means that the amount of chemical reaction that takes place during contraction depends on the type of contraction performed by the muscle. In other words, the flow of energy is subject to regulation.

The practical work of muscles

The muscle's ability to do work is of great practical importance, since it enables animals to move their limbs and perform other actions. The efficiency of the process of muscle contraction depends on the fate of the free energy released in chemical reactions—*i.e.*, whether it is converted primarily into work or degraded into heat. Although the production of heat can detract from the efficiency of working muscle, energy that appears as heat is not always wasted. In warm-blooded animals, for example, the heat released by muscles maintains a constant body temperature regardless of the environmental temperature. When an animal shivers in the cold, a large amount of heat is generated in the muscles. The muscles alternately contract and relax, releasing energy chiefly as heat.

The energy-yielding reaction. Muscles use the free energy released by chemical reactions by coupling the chemical reaction to physical changes in the contractile proteins. The exact molecular details of this fundamental coupling process are not yet completely known. Of the reactions that have been identified, the splitting of adenosine triphosphate (ATP) is the energy-yielding reaction nearest to the contractile event. Water participates in this reaction in which ATP is broken down to adenosine diphosphate (ADP) and phosphate (P_i); the reaction that occurs in the muscle, during which chemical free energy is converted into work, can be written as follows:

$$\text{ATP} + H_2O + \text{contractile elements} \rightarrow \text{ADP} + P_i + \text{contractile elements} + \text{work} + \text{heat}.$$

This equation emphasizes the obligatory role of the contractile elements and the coupled nature of the reaction that produces work.

ATP is involved in the contraction of muscles in many types of animals. In some the ATP reaction can also be coupled to systems that convert the free energy from the ATP reaction into other kinds of energy. Some fishes, for example, convert it into electricity, and fireflies convert it into light. Within an animal, the net formation or breakdown of ATP is coupled to a large number of chemical reactions.

The ATP-building reaction. In a resting vertebrate muscle, the available supply of ATP can sustain maximal muscle work for less than one second. The muscle, therefore, must continuously replenish its ATP store, and this is done in many different ways. One mechanism for the formation of ATP operates so rapidly that for a long time scientists were unable to detect any change in the amount of ATP in the muscle as a result of contraction. This immediate rebuilding of ATP is accomplished by the reactions of compounds called phosphagens. All of them contain phosphorus in a chemical unit called a phosphoryl group, which they transfer to ADP to produce ATP. Whereas different phosphagens occur in various animals, all of them contain two common chemical units—a guanidino group and a phosphoryl group.

Phosphagens

The reaction for the rebuilding of the ATP that is utilized during contraction in vertebrate muscle can be written as follows:

$$\text{ADP} + \text{PCr} \rightarrow \text{ATP} + \text{Cr}.$$

The phosphagen, phosphorylcreatine (PCr), gives rise to creatine (Cr) upon the formation of ATP. This reaction is catalyzed by the enzyme ADP: creatine phosphoryl transferase.

Resting skeletal muscle has about ten times more phosphagen compound than ATP. During rapid and intense contraction, the phosphagen can be utilized to rebuild ATP rapidly and maintain its level in the muscle only so long as the phosphagen lasts, which in a maximally working human muscle is just a few seconds. The phosphagen level is replenished after contraction by a reversal of the reaction given previously. In this reaction, ATP is utilized to form phosphagen (PCr) from creatine (Cr); ADP is also formed.

$$\text{ATP} + \text{Cr} \rightarrow \text{ADP} + \text{PCr}$$

This occurs at the same time that ATP is produced by various reactions that are described in a later section.

As mentioned earlier, changes in the ATP level during contraction are difficult to measure. Such changes can be observed, however, when the muscle is treated with the chemical 2,4-dinitro-fluorobenzene (DNFB), which inhibits the enzyme ADP:creatine phosphoryl transferase. In DNFB-treated muscles, the ATP level decreases during contraction; the ADP and phosphate (P_i) levels increase, and phosphorylcreatine and creatine levels remain constant. The mechanical properties of DNFB-treated muscle during maximal contraction are the same as those of untreated muscles, except that DNFB-treated muscle runs out of ATP sooner and therefore contracts for a shorter time.

The amount of phosphagen is higher in skeletal muscle than it is in cardiac or smooth muscle. This correlates

with the type of activity of the muscles. Skeletal muscle operates in bursts of activity, whereas cardiac and smooth muscle contract in a regular pattern. Skeletal muscle needs an immediate supply of a large amount of ATP, which is provided by the phosphagen reaction; cardiac and smooth muscle, which use ATP at a lower rate, rely on slower reactions to fill their energy requirements.

METABOLIC PATHWAYS PRODUCING ATP

ATP from the sugar glucose or a derivative

In skeletal muscle most of the ATP is produced in metabolic pathways involving reactions of the sugar glucose or some other carbohydrate derived from glucose. During contraction, for example, glucose is made available for these reactions by the breakdown of glycogen, the storage form of carbohydrate in animal cells. During the recovery period following relaxation, the glycogen supply is replenished by synthesizing glycogen from glucose brought to the muscle in the blood.

The glycolytic pathway. In all but the most brief, intense contractions—during which phosphagen is utilized to form ATP—additional ATP is supplied by a set of chemical reactions called the glycolytic pathway, or the Embden-Meyerhof-Parnas pathway. These reactions are especially important when the muscle is doing work at a high rate. The enzymes that catalyze these reactions are located in the sarcoplasm, as is glucose, the major starting material. The glycolytic pathway can be summarized as follows: reactions of glucose to lactate are coupled to formation of ATP from ADP and phosphate.

$$\underset{\text{glucose}}{C_6H_{12}O_6} + 2\ \text{ADP} + 2\ P_i \rightarrow \underset{\text{lactate}}{2C_3H_6O_3} + 2\ \text{ATP} + 2H_2O$$

As already indicated, ATP is broken down in the contractile reaction to form ADP and P_i, which can then be recycled into the pathway. The lactate, a waste product, diffuses out of the muscle and is transported by the blood to the liver. The removal of lactate from the muscle is slower than its production; if the muscle contracts for an extended period, the accumulation of lactate in the muscle causes fatigue and ultimately muscle cramps.

Most of the lactate taken to the liver is converted to glucose, from which glycogen is formed in a series of reactions that require ATP to provide more ATP. About one-sixth of the lactate is oxidized to carbon dioxide and water through metabolic pathways such as the Krebs cycle and oxidative phosphorylation. Because these reactions require oxygen after the exercise is done, a so-called oxygen debt accumulates during brief intense exercise. Animals breathe hard after intense exercise to supply the oxygen needed for the reconversion of lactate.

The relative speed of glycolysis

The glycolytic pathway can operate up to 1,000 times faster in a contracting muscle than it does in a resting muscle. Two regulatory mechanisms play roles in the increase in the rates of these reactions. First, reacting compounds (reactants) accumulate; for example, more glucose is available because glycogen breakdown increases during contraction, and ADP and P_i accumulate. The second type of regulation depends on the activity of the enzyme phosphofructokinase, which catalyzes one of the early steps in glycolysis. This enzyme can unite with, or bind, either ATP or AMP. When ATP is bound, the activity of the enzyme is low; that is the reaction it catalyzes proceeds slowly. When AMP is bound, however, the enzyme has a high activity. During contraction the ATP level drops and that of ADP increases; ADP, in turn, produces AMP as well as ATP, in a reaction catalyzed by the enzyme adenylate kinase, or myokinase. The AMP then can bind with the phosphofructokinase, thereby accelerating glycolysis and the production of ATP. The formation of ATP in the myokinase reaction enables the muscle to use the energy in the ADP formed during contraction.

The oxidative pathways. A large amount of ATP is produced by the complete oxidation of carbon-containing compounds (*e.g.*, glucose) to carbon dioxide and water in reactions of the Krebs cycle and oxidative phosphorylation. These reactions take place inside the membrane-bound components of cells called mitochondria. Carbohydrate can enter the oxidative pathways at a branch point in the glycolytic pathway. At the branch point, two kinds of reaction can occur: pyruvate, a three-carbon compound, can either react to form lactate or lose a carbon dioxide molecule to form a two-carbon unit, which then enters the Krebs cycle. If the muscle is working hard and using ATP rapidly, lactate is formed. If it is working slowly, a larger fraction of the carbon enters the oxidative pathways.

During reactions of the Krebs cycle, carbon dioxide and hydrogen atoms are removed from the compounds that enter the cycle. The carbon dioxide diffuses out of the muscle and is carried by the blood to the lungs, where it is exhaled. The electrons from the hydrogen atoms are passed through an electron-transport chain consisting of the series of reactions involving cytochrome molecules. These events occur in the mitochondria, in which ADP is also combining with phosphate (P_i) to form ATP. In the last step of this reaction sequence, the hydrogen atoms combine with oxygen to form water. The net reaction for each glucose molecule that enters the glycolytic pathway and proceeds through the Krebs cycle and oxidative phosphorylation is summarized below.

$$\underset{\text{glucose}}{C_6H_{12}O_6} + 6\ O_2 + 38\ \text{ADP} + 38\ P_i \rightarrow 6\ CO_2 + 38\ \text{ATP} + 44\ H_2O$$

Efficiency of oxidative pathways

Although the oxidative pathways result in the formation of more ATP from each glucose molecule than does the glycolytic pathway, the complete process is much slower than glycolysis. The rate of ATP production by the oxidative pathways is limited by the rate at which starting materials can be delivered to the mitochondria. The slowest steps are those involving the passage of the carbon-containing compounds from the sarcoplasm into the mitochondria and the delivery of oxygen, which ultimately comes from the air that is breathed. In red blood cells oxygen is combined with hemoglobin, a protein containing four identical subunits, during its transfer from the lungs to the muscles. Muscle contains myoglobin, which has a structure similar to a single subunit of hemoglobin. Myoglobin, which combines with and can store oxygen, is responsible for transporting much of the oxygen through the sarcoplasm to the mitochondria. Myoglobin is especially important in the heart and "red" muscles, which rely heavily on oxidative metabolism for the production of ATP.

During intense exercise of skeletal muscle, ATP is supplied almost entirely by carbohydrate metabolism. During rest or very light exercise, however, skeletal muscle depends largely on the oxidation of stored fats for the production of ATP. Two-carbon units and hydrogen atoms are removed from fats in a stepwise fashion. The two-carbon units enter the Krebs cycle, as do the identical units derived from carbohydrates, as the compound acetyl co-enzyme A. The hydrogen atoms from the breakdown of the fat and from the reactions of the Krebs cycle proceed through oxidative phosphorylation, and ATP is produced.

Heart muscle, which operates constantly but at a lower level of intensity than does skeletal muscle, also depends on oxidative reactions for ATP. Most of the two-carbon units, however, are derived from either fatty acids or ketones, components of fat, brought to the heart by the blood (see also METABOLISM for further detail on metabolic pathways).

MOLECULAR COMPONENTS

The myofilaments, which are the structural elements responsible for muscle contraction, are composed of several different proteins, constituting about 50 percent of the total protein in muscle. The other 50 percent consists of the proteins in the Z line and M band, the enzymes in the sarcoplasm and mitochondria, collagen, and the proteins in membrane structures. Of the myofilament proteins, myosin and actin are known to play a direct part in the contractile event. Troponin and tropomyosin, which are located in the thin filaments, regulate contraction by controlling the interaction of myosin and actin.

Myosin. Myosin makes up more than half of the myofilament protein and is the major component of the thick filaments in muscle. Each molecule of myosin has a mo-

lecular weight of about 500,000 and contains two long protein chains and two or three short chains. Each of the long chains has a molecular weight of 225,000 and contains an estimated 1,800 amino acids; they are, in fact, the longest protein chains that have yet been studied.

The myosin molecule can be split to produce two fragments, light meromyosin (LMM) and heavy meromyosin (HMM).

LMM and HMM fragments of myosin

The LMM portion, about 80 nanometres long, contains most of the tail region of the original myosin molecule. In it each of the two long chains of protein is in a helical conformation, and the two chains are wound around each other to form a so-called superhelix. The LMM portion forms the backbone of the thick filament in the muscle. The HMM portion, also about 80 nanometres long, consists of two globular head regions attached to a part of the superhelical tail region of the myosin molecule. It forms the cross bridges that protrude from the thick filaments in muscle and interact with the thin actin-containing filaments. HMM contains the enzymatic site that catalyzes the splitting of ATP.

There are some structural differences in myosins isolated from different muscles within the same animal and from muscles at different stages of development. Myosins from muscles in the fetus and the adult of the rabbit, for example, contain different amounts of unusual amino acids (*e.g.*, methylhistidine) and have different degrees of sensitivity to trypsin, an enzyme used to split the myosin molecule.

In addition to forming an important structural part of the thick filaments in muscle, myosin has enzymatic activity: it catalyzes the splitting of ATP in muscle. Of the known chemical reactions that occur during contraction, this one is most closely associated with the actual contractile event. Isolated myosin splits ATP, but its properties are modified by the presence of actin and other myofilament proteins in the muscle.

Several factors influence the enzymatic activity of myosin—*i.e.*, the rate at which the breakdown of ATP proceeds. The presence of other muscle proteins is one factor. The activity of myosin is also very strongly dependent on the amount of magnesium ion or calcium ion present.

Actin. Actin, which constitutes about 25 percent of the protein of myofilaments, is the major component of the thin filaments in muscle. An individual molecule of actin is a single protein chain coiled to form a roughly egg-shaped unit. Actin in this form, called globular actin or G-actin, has one calcium ion and one molecule of ATP bound to it. Under the proper conditions G-actin is transformed into the fibrous form, or F-actin, that exists in the thin filament in muscle. In F-actin the globular units are in two long strands twisted around each other. When the G-to-F transformation takes place, the ATP bound to G-actin breaks down, releasing inorganic phosphate (P_i) and leaving an ADP molecule bound to each actin unit. The transformation of G-actin to F-actin is a complex reaction in which specific sites on the molecules must interact during the transformation process.

Transformation of G-actin to F-actin

Actomyosin. Mixtures of isolated myosin and actin are used to study the relationship between the ATP breakdown reaction and the interaction of myosin and actin. The ATP reaction can be followed by measuring the change in the amount of a reactant, or product, present in the solution. The myosin–actin interaction also changes the physical properties of the mixture. If the concentration of ions in the solution is low, myosin molecules aggregate into filaments, and the solution is viscous. As myosin and actin interact, they squeeze out water and form a tight compact gel mass; the process is called superprecipitation.

Purified actin and myosin, in the presence of magnesium ions and ATP, superprecipitate and split ATP. If troponin and tropomyosin are also present, however, the actin and myosin do not interact, and ATP is not broken down. This inhibitory effect corresponds to the state of relaxation in the intact muscle. When calcium ions are added, they combine with troponin, inhibition is released, actin and myosin interact, and ATP is broken down. This corresponds to the state of contraction in intact muscle. The exact mechanism by which troponin, tropomyosin, and calcium ions regulate the myosin–actin interaction is not yet known. In the relaxed muscle, each thin filament contains one troponin and one tropomyosin molecule for every seven actin units.

A somewhat different scheme of regulation operates in the muscle of mollusks. As in vertebrate muscles, calcium ions act as the initiator of contraction. The difference is that the component that binds calcium ions in the mollusc muscle is associated with myosin rather than with the actin-containing thin filaments.

MOLECULAR MECHANISMS OF CONTRACTION

Excitation–contraction coupling. The nerve impulse that ultimately results in muscle contraction appears as an action potential at the sarcolemma, the membrane that surrounds the muscle fibre. This electrical signal is communicated to the myofilaments inside the fibre in the following way. When the action potential reaches the opening of the transverse tubules (channels that open through the sarcolemma to the space outside the fibre; see above *Structure and organization: The myofibril*) in the surface of the fibre, it travels down into the fibre along the tubular membranes, which are continuous with the surface membrane, to within a fraction of a micron of each functional contractile unit (Figure 2). In frog muscle the transverse tubules surround the myofibril at the level of each Z line, and in mammalian muscles they are located at the edge of the A bands and I bands. At the triads (the three-element complex consisting of one transverse tubule and two cisternae, which are enlarged saclike membranes) the transverse tubule walls are very close to the membranes of the terminal cisternae of the sarcoplasmic reticulum.

By some as yet unknown mechanism, the change in the electrical properties of the transverse tubule during an action potential causes the very rapid release by the terminal cisternae of relatively large amounts of calcium ion into the sarcoplasm. As the concentration of calcium ions increases in the sarcoplasm, they become bound to the troponin in the thin filaments. This releases (or removes) the troponin–tropomyosin-mediated inhibition of the myosin–actin interaction. As the stimulation of the muscle continues, the terminal cisternae continue to release calcium ions. At the same time, however, some of the calcium ions are being removed from the sarcoplasm by another portion of the sarcoplasmic reticulum, the longitudinal tubules. Once the calcium ions are inside the lumen (cavity) of the longitudinal tubules, many of them slowly diffuse back to the terminal cisternae. Some step in the process of removal of calcium ions from the sarcoplasm by the sarcoplasmic reticulum is an energy-requiring one. The breakdown of ATP is the chemical reaction that supplies the energy, and two calcium ions are apparently removed from the sarcoplasm for each ATP molecule that is split. The exact energy-requiring process is not yet known, but is likely to involve the transport of calcium ions through the membrane into the lumen of the sarcoplasmic reticulum.

The effect of calcium ions

The development of muscular force. Until the early 1950s, most theories to explain muscle contraction were based on the idea that a single structural element, which extended the entire length of the muscle, folded up or changed shape in some way, causing the muscle to shorten. In the early 1950s, however, it became clear that there were two types of filaments in muscle and that neither extended the entire length of the muscle. Furthermore, it was found that the overall length of the filaments did not change during contraction. Those findings ruled out most of the then existing theories about the mechanism of muscle contraction and provided the basis for the now widely accepted sliding-filament theory.

According to the sliding-filament theory, movement involves the two sets of filaments, thick ones and thin ones, which are arranged in an overlapping array in each sarcomere (see Figure 1). Shortening of the entire muscle occurs as the thin filaments on both sides of the A band slide further into the A band (see Figure 7B). As the sliding progresses, the areas of the sarcomere containing only one type of filament—that is, the I band and the H zone—decrease in size because more and more of the thin and

thick filaments overlap each other. On the other hand, the A band remains the same length because the thick filaments do not change in length except in extreme shortening. The distance from the Z line to the edge of the H zone also remains virtually the same length because it is determined by the length of the thin filaments.

Although the sliding-filament theory seems basically correct, it must be expanded to explain certain aspects of contraction: how the force that moves the filaments past each other is generated; how ATP takes part in the process of contraction. A major difference between the various expanded versions of the theory is that some depend on the operation of the cross bridges, while others do not.

Expanded cross-bridge theories

In the cross-bridge theories, the sliding of the filaments is thought to result from the interaction of the cross bridges with the thin filaments during contraction. Each time a bridge links to a thin filament and operates in a specific way (see below), it causes a small movement of the thin filament along the thick filament. Since muscles are able to shorten considerably, requiring sliding of the filaments through large distances, there must be repeated cycles of interaction between a given cross bridge and successive sites on the thin filament.

According to one theory, the cross bridge contains a flexible peptide (*i.e.*, a very short chain of amino acids) with a partly ionized molecule of ATP attached to it. When the bridge is not attached to the thin filament, the negatively charged ATP is repelled by another negative charge in the myosin molecule near the attachment of the peptide to the rest of the myosin molecule. This repulsion maintains the peptide in an extended form, which is partly a random coil and partly the fully extended form (β form). When the muscle is activated by calcium ions and the cross bridge attaches to actin in the thin filament, the charge on ATP is neutralized and repulsion ceases. As a result, the peptide takes on a much shorter folded form (α helix). The formation of hydrogen bonds in this form provides the energy necessary to pull the thin filaments farther into the A band. When the peptide has shortened, the ATP attached to it comes into contact with the myosin ATP-splitting enzyme (ATP-ase) located at the base of the peptide; ATP is split, breaking the link between the thick and thin filaments. The charge repulsion is restored when the ATP is reformed. The peptide is then re-extended into the partly β form and is ready to link with the next site on the thin filament.

In a different model of cross-bridge operation, the cross bridges attach to the thin filament and then tilt the angle of their attachments by an undescribed mechanism. This tilting of the cross bridges pushes the thin filament, producing a sliding motion. ATP is split at some unspecified stage during this process.

Non-cross-bridge mechanisms

In the non-cross-bridge mechanisms, the force responsible for the sliding of the filaments is attributed to electrical forces between the thick and thin filaments—not to the cross bridges or to any other physical link between the filaments. It has been proposed that both sets of filaments (thin and thick) are negatively charged during rest. Following the release of calcium ions, the thin filaments become positively charged and tend to slide into the array of the thick filaments.

III. The nature and properties of cardiac muscle

The heart is the pump that keeps blood circulating throughout the body and thereby transports nutrients, breakdown products, antibodies, hormones, and gases to and from the tissues. The heart consists mostly of muscle, the myocardial cells (collectively termed the myocardium), arranged in several ways that set it apart from other types of muscle. The outstanding characteristics of the action of the heart are its contractility, which is the basis for its pumping action, and the rhythmicity of the contraction.

STRUCTURE AND ORGANIZATION

Because the gross design of the heart as a pump depends on the type of respiratory and circulatory system of the animal, a variety of hearts occur in the animal kingdom. In all hearts, however, the pumping action results from the contraction of at least one chamber. The four chambers of the mammalian heart—the two atria and the two ventricles—are composed primarily of contractile tissue. The atria act chiefly as chambers for receiving the blood. The ventricles, with thicker walls, do most of the pumping.

The myocardial cells. The sheets, or bands, of muscle fibres (myocardial cells) of the ventricles are arranged in a complex pattern. Some encircle only the cavity of the left ventricle. Others surround both ventricles in a helical pattern. This arrangement is advantageous because, when the ventricular muscle contracts, the blood is virtually wrung, rather than pushed, out of the heart.

Intercalated disks

Myocardial cells are branched and form a netlike pattern, with occasional dense regions called intercalated disks at sites where two myocardial cells abut. Each disk shows two distinct regions. One region, which runs perpendicular to the long axis of the contractile filaments, occurs only at the level of the Z line in the sarcomere. The contractile filaments themselves never cross the disk region from one cell to another. The thin filaments branch at this region of the disk, producing a dense region that can be seen in the light microscope. At the intercalated disk the membranes of the two abutting cells are highly folded, providing a large area of contact between the cells and perhaps also adding to the mechanical strength of the tissue. The second region of the intercalated disk occurs at the point at which the membranes of two adjacent cells lie next to each other parallel to the contractile filaments. In this special area, called the nexus, the membranes of the two cells appear to fuse and form a tight junction. It is thought that the nexus areas present a much lower resistance to intercellular electrical communication than usually exists between separate cells. This property enables the heart to respond to stimulation as a single functional unit even though it is composed of many individual myocardial cells.

Other aspects of the fine structure of myocardial cells are similar to those of skeletal muscle cells. Cardiac muscle, however, does not have as extensive a system of sarcoplasmic reticulum as does skeletal muscle. Cardiac muscle seems to depend on the influx of calcium ions from either the extracellular space or some other area during contraction, to supplement the calcium ions released from the sarcoplasmic reticulum.

The pacemakers of the heart. In addition to the cells that look and act like typical muscle cells, the heart contains cells that have very few myofibrils. These cells constitute the pacemaking areas, which are primarily responsible for the rhythmic nature of the contraction and for conducting the activating signal through the heart.

Self-activation of the heart

The autorhythmicity, or self-activating character, of the heart clearly sets it apart from skeletal muscle. Each animal species has a characteristic heart rate determined by specialized cells that constitute the pacemaker of the heart.

In the normal human heart, a group of cells called the sino-atrial node, located on the inner wall of the right atrium, acts as the pacemaker. The node operates in conjunction with a special conduction system that spreads the excitation over the heart, coordinating the progress of its contraction. The system's components are the atrioventricular node and the Purkinje fibres. The atrioventricular node is located in the lower part of the right atrium, close to the septum between the right and left parts of the heart. From the atrioventricular node the network of Purkinje fibres proceeds to the ventricles and then branches out over the surface, penetrating the ventricular walls.

Although the atrioventricular node and the Purkinje fibres have the ability to act as pacemakers—initiating rhythmic contraction of the heart muscle—they do not do so normally. Some unknown mechanism permits the cells with the highest rate of pacemaker activity to dominate those with lower rates of activity. In humans the sino-atrial node, with its intrinsic rate of 70 to 80 excitations per minute, acts as the pacemaker under normal conditions. If it fails to operate for some reason, such as disease, the pacemaker role may be taken over by the

atrioventricular node, whose rate of excitation is 40 to 60 per minute, or the Purkinje fibres, whose rate is 15 to 40 excitations per minute.

MECHANICAL PROPERTIES

The initiation of systole

The contraction–relaxation cycle. The contraction of the heart, or systole, is initiated by the action of the pacemaker, whose cells have special membrane properties that cause them to generate electrical signals spontaneously. When the membrane potential reaches a certain level, an action potential is propagated and is communicated directly to the atrial cells adjacent to the sino-atrial node. The action potential spreads over the atria and activates the myocardial cells, causing them to contract and force blood into the ventricles.

The action potential is then transmitted from the atria to the ventricles via the atrioventricular node and Purkinje fibres. Since the node fibres are smaller than the atrial cells, the progress of the action potential slows at this point, delaying contraction of the ventricular muscle until the atria have emptied their contents into the ventricles. When the action potential has been transmitted through the atrioventricular node, it reaches the larger Purkinje fibres, through which conduction is faster than through the ventricular tissue itself. The Purkinje fibres branch out over the surface of the ventricles and into their thick walls. The widespread distribution of the Purkinje fibres, as well as their speed of conduction, causes the entire ventricular muscle mass to contract almost simultaneously. This contraction forcibly wrings the blood out of the right ventricle into the pulmonary artery toward the lungs and out of the left ventricle into the aorta toward the rest of the body.

The source of the action potential and the conduction pathways for it are especially suited to the function of the heart. The action potential itself has a special characteristic: long duration. It is 20 to 50 times longer than the action potential in skeletal muscle.

Just as in skeletal muscle, depolarization is followed by a time of decreased excitability. The first part of this time is the absolute refractory period, during which the membrane will not conduct an impulse regardless of the strength of the stimulus applied to it. Following this is the relative refractory period, during which the membrane will respond, but the stimulus must be much stronger than normal. In cardiac muscle, the refractory periods outlast the action potential and are almost as long as the contractile response. This ensures that the contractile elements have time to relax before the membrane has recovered its normal degree of excitability. An important consequence of the long refractory period is that cardiac muscle cannot be tetanized to produce constant tension, a safety factor for an organ that depends on the repetition of complete cycles of contraction and relaxation to operate as a pump.

Adaptability of the heart

Regulation of the heartbeat. The fact that the heart beats faster during and after hard work and during periods of excitement or fright exemplifies the remarkable ability of this organ to adapt to the changing demands placed on it by the body. One mechanism responsible for this adaptiveness operates in the heart itself and can be expressed as Starling's law of the heart: the quantity of blood that fills the ventricles during relaxation, or diastole, regulates the force of the subsequent contraction of the ventricles during systole. The more blood that enters the heart, the more vigorous the subsequent contraction. This kind of response helps avoid a backup of blood in the veins and tissues. A similar situation occurs in skeletal muscle.

The pacemaker cells are the sole source of the excitation of the heart and set the normal rate of rhythmic contraction and relaxation. The nerves that serve the heart do not keep the heart beating but act only to modify the rate of the heartbeat; they originate in the medulla oblongata, the basal part of the brain, and consist of two types: inhibitory (the vagus nerves) and stimulatory (sympathetic) nerves.

The vagus nerves innervate the atrial portions of the heart only. They act by releasing acetylcholine, which increases the electrical potential across the cell membranes, and thus slow down or, in extreme cases, stop transmission of impulses in the sino-atrial and atrioventricular nodes. If transmission is stopped, the Purkinje fibres, which are located in the ventricles and therefore are not under the influence of the vagus nerves, may begin to act as pacemaker cells. The sympathetic nerves, found in all parts of the heart, act by releasing norepinephrine. Stimulation of the nerves leading to the sino-atrial node increases the rate of electrical discharge of the node and thus speeds the heartbeat.

METABOLIC PATHWAYS

The myocardium converts chemical energy, derived from nutrients, into the mechanical energy that propels the blood through the body and a certain amount of heat. The molecule used by the contractile elements in the energy-conversion process is ATP (adenosine triphosphate).

One starting material for the production of ATP in the heart is glycogen, which is broken down into glucose units and then metabolized through the steps of the glycolytic pathway (see above *Energy transformations*); three molecules of ATP are formed for each glucose unit. Although the other product, pyruvate, can react to form the compound (acetyl co-enzyme A) that can enter the tricarboxylic acid, or Krebs, cycle, most acetyl co-enzyme A in the heart is formed from free fatty acids brought to the heart in the blood. In the Krebs cycle, the carbon atoms in acetyl co-enzyme A are oxidized to carbon dioxide; the hydrogen atoms formed take part in oxidative phosphorylation, and react with oxygen to form water. During this process three molecules of ATP are formed.

Conversion of glucose to ATP in the heart

Normally, the heart depends on the metabolic processes that take place in the presence of oxygen. If carbohydrate (as glucose) is the starting material, 38 molecules of ATP are produced for each glucose unit metabolized. If fats are the starting material, 12 molecules of ATP are produced for each two-carbon unit derived from free fatty acids. If the oxygen supply to the heart is limited, some ATP can be formed through the glycolytic pathway, by the metabolism of glucose to pyruvate and then to lactate. Under completely anaerobic conditions (no oxygen at all), however, the amount of ATP produced is inadequate to keep the heart functioning for very long.

IV. The nature and properties of smooth muscle

VERTEBRATE SMOOTH MUSCLE

The term smooth muscle derives from the uniform appearance of the cytoplasm of stained sections when examined under the low magnification of a light microscope. Although many cellular structures can be seen in the electron microscope, they are not nearly as striking as the internal organization of skeletal muscle cells.

Many organs of vertebrates contain smooth muscle: the stomach, intestines, bile ducts, bladder and urinary tracts, uterus, and blood vessels. Smooth muscle exhibits much variation and is often separated into two major groups—visceral and multi-unit. Visceral muscles operate as a syncytium (*i.e.*, as a single functional unit), and many of them also have pacemaker, or self-stimulating activity. A multi-unit smooth muscle is a group of cells, each of which is subject to nervous control and can act individually in the same manner as individual fibres in a skeletal muscle. Actually, there is probably a wide spectrum of smooth muscle types between these extremes.

Some of the details of the structure and operation of muscle in the intestine and in the vasa deferentia (singular vas deferens, a duct leading from the testes to the penis) are described in this section. These tissues contain muscles that embody the major characteristics of the two categories: intestine representing visceral smooth muscle and vas deferens representing multi-unit smooth muscle.

Visceral, or single-unit, muscle. *Structure and arrangement.* The individual smooth muscle cell of the intestine is often roughly spindle-shaped, wide in the middle and tapering to the ends; some are ribbon shaped or rod shaped. In the thickest middle section the diameter is usually five to ten microns (μ), ten to 20 times smaller

Size of smooth muscle cells

than skeletal muscle fibres. The length of a smooth muscle cell ranges between 50 and 500 μ. How closely packed the cells are depends on the tissue and is correlated with its operation. In syncytial-type muscles of the intestinal walls, the cells are often arranged in branched groups, or bundles, of a few hundred cells. Adjacent cells in a bundle are usually separated by 50 to 80 nanometres of basement membrane. The bundles of cells are separated from each other by layers of collagen fibrils and elastin. Blood vessels and nerves may also occur in the extracellular spaces.

Smooth muscle cells have an ellipsoidal nucleus in about the middle of the cell. Mitochondria, endoplasmic reticulum, and, in some cells, glycogen granules are present. Thin filaments have a diameter of 50 nanometres and contain the proteins actin, tropomyosin, and troponin, just as do the thin filaments of skeletal muscles. Recent studies have disclosed thick filaments as well (these apparently were destroyed by the preparation procedures used in earlier electron-microscopic studies).

Several features of the arrangement of the filaments within smooth muscle set it apart from other types of muscle. The thick filaments, surrounded by many thin ones, are not arranged in a regular array as they are in skeletal and cardiac muscle. Neither does smooth muscle contain Z lines; however, there are scattered dense regions that may represent a structure analogous to the Z line. In some cells, thin filaments converge at the dense regions; other thin filaments are attached to the cell membrane. Despite the relatively low degree of order in the arrangement of the filaments, the sliding-filament mechanism is thought to operate during contraction of smooth muscle.

Like cardiac muscle, some smooth muscle has regions of close contact between individual cells. The membranes of adjacent cells occasionally come into very close contact forming nexuses, areas of cell-to-cell communication that enable the muscle to function as a syncytium. There are also small bridges between cells and fingerlike projections at the ends of cells.

Contraction in the form of peristaltic waves

Nature of contraction. In most of the tubular organs, smooth-muscle contraction is in the form of peristaltic waves. A small area of the contraction partly encircles the tube and moves along it for only a short distance before fading.

Unlike other excitable tissues, such as nerve and skeletal muscle, the membrane potential is constantly drifting or fluctuating rather slowly. A burst of action potentials, however, accompanies contraction. These signals originate in many regions of the intestinal muscle that act as pacemakers. Specific pacemaker cells such as are found in the heart apparently do not occur; instead pacemaker activity continuously moves to different smooth muscle cells of the intestine.

In pacemaker regions the membrane slowly and spontaneously depolarizes. These slow potentials, or slow waves, are small and spread in a diffuse way. They do not directly cause contraction of the muscle; however, when the slow potential reaches the threshold value for the membrane, action potentials are propagated. These pass from cell to cell by way of the nexuses between cell membranes. In this way all the cells in a bundle (of perhaps a few hundred cells) get the signal to contract. There also seems to be some communication between the bundles. A wave of contraction (which may include thousands of cells) then progresses partly around the intestine and then along it for a distance of a few centimetres. Intestinal smooth muscle, therefore, operates as a syncytium. To varying degrees, many other vertebrate smooth muscles, especially in the digestive tract, operate in the same manner.

Variable excitability of intestinal muscle

The electrical, or membrane, excitability of intestinal muscle cells is quite variable. Both the level of the resting potential and magnitude of the threshold potential can change. The activity of the nerves supplying the muscle is one factor that influences membrane properties. The structural relationship of the nerves and muscles differs from that in skeletal muscles. In the taenia coli, a muscle forming a particular part of the large intestine, a bundle of small autonomic nerves surrounded by Schwann cells penetrates the muscle. Occasional portions of the nerves are bare (*i.e.*, not surrounded by Schwann cells), forming areas probably analogous to the synapse region of the motor nerves. Unlike motor nerves, however, these autonomic nerves end at sites about 100 nanometres away from the intestinal muscle, and there appear to be no receptive structures in the muscle membrane opposite the nerve ending.

Despite this lack of highly organized nerve–muscle junction, the nerves do influence membrane properties and intestinal-muscle action. Two types of nerves are involved: the sympathetic nerves, which inhibit contraction (an effect mimicked by the chemical epinephrine), and the parasympathetic nerves, which intensify contraction (an effect mimicked by acetylcholine). Other syncytial smooth muscles, such as the ureter, operate without neural influence; waves of contraction are coordinated by the muscle itself.

The mechanical state of the muscle is another factor that can influence its excitability. When intestinal muscle is stretched, the passive, or elastic, tension in it increases, and the resting potential becomes less negative; *i.e.*, it changes in the direction of depolarization, and the rate of discharge of action potentials increases. The mechanical properties of smooth muscle depend to some extent on the large amount of elastic material in the tissue, which probably contributes significantly to the regulation of contractile activity and to the overall operation of the tissue.

Multi-unit muscle. The other type of vertebrate smooth muscle is the non-syncytial, or multi-unit, muscle, as exemplified in the vas deferens. The vasa deferentia do not have pacemaker activity but depend entirely on the nerves for activation. The nerves lie in depressions in the muscle, a much closer spatial relationship between nerve and muscle than occurs in syncytial smooth muscles. Every muscle cell of a vas deferens has at least one nerve ending associated with it. The nerve endings contain two types of vesicles: granular vesicles, which contain compounds resembling adrenaline that probably act as transmitters, and agranular vesicles, which contain acetylcholine, the chemical transmitter in vertebrate skeletal muscle. When a single nerve impulse occurs, it results in a small change in the membrane potential of the muscle cell. This change, called a junctional potential, is probably caused by the direct influence of a transmitter substance on the muscle membrane. Junctional potentials resulting from a series of nerve impulses can accumulate, progressively depolarize the membrane, and attain the threshold, at which point action potentials are propagated. The transmitter apparently is not rapidly destroyed since it seems to diffuse to other cells in addition to the one closest to the nerve ending.

Vesicles in nerve endings

INVERTEBRATE SMOOTH MUSCLE

Catch muscle. One of the most interesting types of muscle, invertebrate "catch" muscle, contains the protein paramyosin. The muscles are present in clams, scallops, and oysters, animals with two shells. The paramyosin-containing fibres are in the adductor muscles that hold the shells closed, an action accomplished by maintaining tension against a springy ligament at the hinge point of the shells. When the adductor muscles relax, the ligament is released, and the shells open by spring action. The remarkable aspect of the adductor muscles is that they can remain contracted continuously for weeks. This type of contraction is called the "catch" state.

Structure and arrangement. The cells of the catch muscles are very long and contain two types of filaments. One type, similar to the thin filaments in striated skeletal muscle, contains the protein actin and has the same dimensions as a skeletal thin filament. The other kind of filament, however, is unique to catch muscles. It contains both paramyosin and myosin: the core of the filament is paramyosin and the covering is myosin. It ranges in diameter from 30 to 150 nanometres—about ten times greater than a thick, myosin-containing filament found in striated skeletal muscle or heart muscle. The paramyosin-contain-

ing filaments are about 30 nanometres long—about 20 times longer than the myosin-containing filaments in skeletal muscle.

Nature of contraction. Evidence suggests that contraction in catch muscles is based on actin–myosin interaction and a sliding-filament mechanism. Paramyosin may modify this interaction and confer the special properties of the catch state. It may also be responsible for the great mechanical strength of the filaments in these muscles.

The catch state resembles the tetanized state of skeletal muscle in that a constant amount of tension is maintained by the muscle. The relationship of the membrane activity to contraction in the two cases is, however, quite different. A skeletal muscle can be tetanized by stimulating it electrically so that action potentials are continuously propagated, and relaxation is impossible. If the rate of stimulation is too low, tetanus does not result, and the tension will rise and fall in phase with the membrane potential. In the catch state, on the other hand, bursts of action potentials appear intermittently, but the tension in the muscle is not synchronized with them.

Since catch muscle can remain contracted even when the membrane is quiescent, how is the contraction turned off? What induces relaxation? Experiments on muscles removed from animals suggest that certain nerves may cause relaxation from the catch state by releasing serotonin, a chemical found in muscle, which acts as an inhibitory transmitter substance.

Experimental stimulation of catch muscle

Mechanical properties. If an isolated catch muscle is stimulated with direct electrical current or by application of acetylcholine, the muscle goes into the catch state, a steady tonic contraction. The same kind of muscle can be made to contract and relax rapidly like a vertebrate skeletal muscle by stimulation with alternating electrical current. This twitchlike response is called a phasic contraction. The mechanical properties of the muscle are quite different in these two kinds of contractions. If a muscle is stretched while in the catch state, it resists the stretch with a much greater force than when it is stretched during a phasic contraction. In other words, the muscle has greater stiffness in tonic contraction than it has in phasic contraction. At any given moment in the catch state, there are apparently more cross-bridge links than in a phasic contraction.

Metabolic pathways. In catch muscles, as in the muscles of vertebrates, ATP is the molecule believed to be directly involved in the energy-transforming processes. During contraction ATP is broken down; it can be reconstituted rapidly by the transfer to ADP of a phosphoryl group from phosphorylarginine, a phosphagen compound. Although catch muscles break down an appreciable amount of ATP during the development of tension, the maintenance of contraction requires only a small amount of ATP breakdown. This is certainly an important factor underlying the ability of catch muscle to remain contracted for a long time without fatiguing.

Quick action muscle. The muscles that operate the wings of some insects contract very rapidly indeed and, weight for weight, can produce more work per second than any other type of muscle. These asynchronous flight muscles are found in insects (*e.g.*, bees, flies, and mosquitoes) that have great manoeuverability during flight. Insects such as butterflies have synchronous muscles, similar to those of other animals. One major difference between the two types of muscle is in the relationship of the nerve impulses to the contractions of the muscle. In synchronous muscle, contraction is produced in direct response to, or is synchronized with, the arrival of nerve impulses at the muscle. In asynchronous muscle, on the other hand, contraction and relaxation proceed at a rate much faster than that at which nerve impulses reach the muscle, more than 1,000 cycles per second in the flight muscles of gnats, for example.

Asynchronous flight muscles

The basis for the differences between these muscles is not completely known. Although asynchronous muscle has a highly organized internal structure similar to that of synchronous muscle, it lacks an extensive sarcoplasmic reticulum, and so the calcium released during activation of the muscle is not removed rapidly from the sarcoplasm. Asynchronous flight muscles thus remain activated during a series of contraction–relaxation cycles. Rapid shortening of the muscle causes an immediate loss of force. The subsequent stretching of the muscle is caused by the elastic properties of the thorax of the insect and results in the redevelopment of tension.

Another striking example of rapid movement is the jumping of an insect such as the flea. The muscle in this action is not a particularly rapid acting muscle; it contracts powerfully and stretches a tendon, thus bending the insect's leg. The quick relaxation of a complementary muscle allows the leg to extend extremely rapidly, using the elastic energy of the stretched tendon.

BIBLIOGRAPHY. J.R. BENDALL, *Muscles, Molecules and Movement: An Essay in the Contraction of Muscles* (1969), with numerous references to original papers; EUGENE BRAUNWALD, JOHN ROSS, and EDMUND H. SONNENBLICK, *Mechanisms of Contraction of the Normal and Failing Heart* (1968), includes information on structure, metabolism of heart muscle, operation of the heart, and heart failure; GEOFFREY BOURNE (ed.), *The Structure and Function of Muscle*, 2nd ed. (1972), a multivolume collection of important articles on most aspects and types of muscles; R.E. DAVIES, "On the Mechanism of Muscular Contraction," in P.N. CAMPBELL and G.D. GREVILLE (eds.), *Essays in Biochemistry*, vol. 1, pp. 29–55 (1965), a review of the history and current knowledge of chemical changes in muscle during contraction, with a brief summary of the author's theory of muscle contraction; HUGH DAVSON, *A Textbook of General Physiology*, 4th ed., 2 vol. (1970), an excellent college-level textbook, with extensive references to original publications; A.V. HILL, *Trails and Trials in Physiology* (1965), includes chapters on specific aspects of muscle and nerve physiology and a bibliography of the author's publications and other major papers from 1909–64—helpful for those interested in original experiments and major developments in the field since 1910; H.E. HUXLEY "Muscle Cells," in JEAN BRACHET and ALFRED E. MIRSKY (eds.), *The Cell*, vol. 4 (1960), an article on the structure of muscle, especially the ultrastructure as seen with an electron microscope; PETER V. KARPOVICH, *Physiology of Muscular Activity*, 6th ed. (1965), includes many aspects of exercise, including metabolism, fatigue, and physical fitness; BERNARD KATZ, *Nerve, Muscle, and Synapse* (1966), an excellent brief work; ALBERT L. LEHNINGER, *Biochemistry* (1970), an excellent college-level biochemistry textbook; DAVID S. SMITH, "The Flight Muscles of Insects," *Scient. Am.*, 212:76–88 (1965); GRAHAM HOYLE, "How Is Muscle Turned On and Off?" *ibid.*, 222:84–93 (1970); H.E. HUXLEY, "The Mechanism of Muscular Contraction," *ibid.*, 213:18–27 (1965); and KEITH R. PORTER and CLARA FRANZINI-ARMSTRONG, "The Sarcoplasmic Reticulum," *ibid.*, 212:73–80 (1965), four well-illustrated short articles written for the general reader; D.R. WILKIE, *Muscle* (1968), devoted entirely to muscle contraction.

(R.E.Da./N.A.C.)

Muscle Diseases

Muscles are bodily tissues that by the contraction of their fibres perform mechanical functions such as those concerned with movement, sustaining weight, balancing, or with narrowing the bore of blood vessels. The voluntary muscles, also called the skeletal muscles, are the muscles of the arms, legs, trunk, neck, and face that are subject to conscious control. It is these muscles that are referred to in the term muscle diseases. Disorders affecting the involuntary muscles, the muscles not subject ordinarily to conscious control, such as those in the walls of blood vessels or of the intestines, are considered to be among the diseases or disorders of the bodily systems of which the muscles are a part—the digestive system, the respiratory system, or the cardiovascular system, for example. (The disorders and diseases of the third type of muscle, that of the heart, are also dealt with in discussions of cardiovascular diseases rather than of diseases of muscle.)

Causes of muscle disease

Voluntary muscles constitute a major part of the human body and are involved in a wide variety of disease states. Muscle disorders may stem from local disorders, such as injury, local disturbance of the blood circulation, local infection, or disuse, or from the general effects of serious systemic disease, including cancer, chronic infections, diseases involving the endocrine (ductless) glands and nutritional deficiencies. There are also primary diseases of muscle (*e.g.*, muscular dystrophy; see below *The muscular dystrophies*), which have been a field of major re-

search in recent years. Nerve and muscle operate as a functional unit, and disease of the motor neurons results in muscular atrophy (wasting) and paralysis. (Neurons are nerve cells; sensory neurons are stimulated by impulses from sensory organs, such as those in the skin or the nose; motor neurons transmit impulses to the muscles and glands, controlling their functioning.)

Dealt with in this article are the manifestations of muscle disease with a note on the abnormalities found in diseased muscle tissue. Included are brief discussions of some of the more important varieties of muscle disease, including some information on their causes, how they are identified, the type of treatment employed, and the outlook for recovery.

COMMON INDICATIONS

Muscular wasting (atrophy) and weakness are among the most common indications of muscular disease. Though the degree of weakness is not necessarily proportional to the amount of wasting, it usually is so if there is specific involvement of nerve or muscle.

Pain is ordinarily present in muscle disease due to defects in blood circulation or injury or involving inflammation of the muscle. It is rare, except as a result of abnormal posture or fatigue, in muscular dystrophy—a familial disease characterized by progressive wasting of the muscles. Cramps are associated with disease of the nerve cells, either motor or sensory; with biochemical disorders (*e.g.*, they may occur in hypercalcemia, a condition in which the blood level of calcium is abnormally high); when the muscle tissues are affected by some form of poisoning; and as a result of disease of the blood vessels. But cramps may also occur when untrained persons exercise, particularly when they are cold.

Enlargement of muscles—muscular hypertrophy—occurs naturally in athletes. Hypertrophy not associated with exercise occurs in two unusual forms of muscular dystrophy, known to doctors as myotonia congenita and hypertrophia musculorum vera. (The latter condition combines increased muscle size with strength; in myotonia congenita muscle cramping is an additional feature.) Pseudohypertrophy, muscular enlargement with weakness, occurs in other forms of muscular dystrophy, particularly the rapidly developing variety called the Duchenne type.

Another manifestation of muscular disease, persistent weakness that is made worse by exercise, is the primary characteristic of the disease myasthenia gravis.

Tetany, tetanus, and fasciculation

Tetany is the occurrence of intermittent spasms, or involuntary contractions, of muscles, particularly of the arms and legs and the larynx, or voice box; it results from low levels of calcium in the blood and from alkalosis, an increased alkalinity of the blood and tissues. Tetanus, a similar term, is a state of continued muscle spasm, particularly of the jaw muscles, caused by toxins produced by a particular bacillus, *Clostridium tetani*. Tetanus is also called lockjaw.

The twitching of muscle fibres controlled by a single motor nerve cell, a process called fasciculation, may occur in a normal person but may be an important indication that muscular atrophy is due to disease of motor nerve cells in the spinal cord. Fasciculation is seen most clearly in muscles close to the surface of the skin.

Intermittent claudication (*i.e.*, a temporary lameness accompanied by cramps and pain) results from insufficiency of blood supply to the legs. The pain and cramps, usually centred in the calves and caused by arteriosclerosis (hardening and thickening of the vessel walls) of the arteries of the legs, tend to occur after a period of walking and to disappear after a few minutes of rest.

Characteristic changes in diseased muscle tissue are evident when specimens of the tissue are examined under a microscope. In muscular dystrophy there is gross variation in fibre size, with splitting of fibres and presence of increased fat and connective tissue within the muscle. In atrophy due to nervous-system disease, groups of atrophied muscle fibres can be seen. In inflammatory disease, fibres that are injured or dead are present, and large numbers of the cells that ordinarily swarm to areas of inflammation can be seen.

TYPES OF MUSCLE DISEASE

Secondary effects of systemic disease. Mention has already been made of the effects produced on the muscles by diseases of the nervous system and of the blood vessels. Other conditions involving the muscles include a group of conditions of unknown cause in which nonbacterial inflammation of the muscles is a prominent feature; muscle disorders caused by endocrine and metabolic disease; disorders due to nutritional deficiencies or intoxication (poisoning); and those associated with cancer.

Inflammatory disease of unknown cause. A group of diseases, of which dermatomyositis and polymyositis are the most important members, has nonbacterial inflammation of the muscles as a prominent feature. Various suggestions have been made as to the origin of these diseases. They include virus infection and auto-immune disorders (allergic reaction to one's own tissues). Many types of polymyositis have been differentiated. At least two of the types affect women in disproportionate numbers, and one type is found in children. Another is associated with cancer. Rash and inflammation of the joints may be present. Treatment with hormones from the adrenal cortex (the outer substance of the adrenal gland) or with synthetic substances for these hormones is beneficial. Formerly acute polymyositis was generally fatal, from effects on the respiratory system, the heart, or the kidneys. Some affected persons now recover, although muscular weakness usually persists.

Nonbacterial inflammation

Endocrine and metabolic disorders. In thyrotoxicosis, a systemic disease caused by overactivity of the thyroid gland, muscular weakness and atrophy are frequently present and may be the first manifestations of the disease. The protrusion of the eyes that is often a feature of thyrotoxicosis is due to the collection of abnormal quantities of fluid in the soft tissues of the eye socket; among these tissues are the extraocular muscles, which control the movements of the eyes.

In myxedema, a conditon resulting from insufficient secretion by the thyroid gland, movements are often slow and the muscles weak and easily fatigued.

When the parathyroid glands are overactive for any reason, there may be substantial muscular weakness, occasionally with muscle cramps and pains. The same set of manifestations occurs when there are abnormally high levels of calcium in the blood for other reasons, which may include the existence of cancer.

Diseases of the pituitary and adrenal glands may cause muscular weakness and wasting. On the other hand, increased muscle bulk and strength may be found early in the course of acromegaly, a disease marked by enlargement of both bony and soft parts of the hands, feet, and face. Acromegaly results from oversecretion of the growth hormone by the anterior lobe of the pituitary. As the condition progresses, the increased muscular strength gives way to weakness.

Cushing's syndrome and Addison's disease

Cushing's syndrome involves weakness of the muscles, particularly of the shoulders, pelvis, and trunk. This condition, the effects of which include fattening of the face, neck, and trunk, impotence, hairiness in females, and high blood pressure, results from abnormally high secretion of adrenal hormones.

Generalized weakness and fatigability of the muscles are also symptoms of Addison's disease, in which reduced secretion of adrenal hormone results in reduced blood levels of sodium and chloride and raised potassium levels, conditions that are probably directly responsible for the effects on the muscles. Limpness and weakness of muscles may occur in any situation in which the blood levels of potassium are abnormally low, as in certain forms of kidney disease and periodic paralysis, a rare hereditary disease.

The carbohydrate glycogen is an important source of muscle energy, and muscle weakness is found in a rare but important group of hereditary diseases, the so-called glycogen-storage diseases, in which various enzyme defects prevent the release of energy by the normal break-

down of glycogen in muscles. As a result abnormal amounts of glycogen are stored in the muscles and other organs. The best known type clinically affecting muscles is McArdle's disease, in which the muscles are unable to degrade glycogen to lactic acid on exertion because of the absence of the enzyme phosphorylase. With a microscope, abnormal accumulations of glycogen can be seen distributed within muscle cells. Features of the condition include pain, stiffness, and weakness in the muscles on exertion. McArdle's disease usually begins in childhood. No specific treatment is available, and persons affected are usually required to restrict exertion to tolerable limits. The condition does not appear to become steadily worse, but there are threats to continuing activity and to life when the muscle protein myoglobin is excreted in the urine or when cardiac problems develop.

Symptoms similar to those of McArdle's disease occur when there is a deficiency of the enzyme phosphofructokinase.

Another glycogen-storage disease results from deficiency of acid maltase, an enzyme normally distributed widely in the body. In this disease both heart and voluntary muscle is affected, and death occurs within a year or so of birth as a rule.

Nutritional disorders and toxic conditions. In man, muscular weakness associated with deficiencies of specific nutrients usually derives from damage to the nervous system. Muscle disorder can derive from vitamin D deficiency. In cystic fibrosis of the pancreas, a disease in which abnormal spaces and accumulations of fibrous tissue are found in the pancreas and in which other abnormalities occur involving the lungs, the liver, and the exocrine glands (glands that have ducts), the muscle weakness that occurs is thought to be due to lack of vitamin E.

Cystic fibrosis of the pancreas

Muscle disorders can be caused by poisoning with any of a number of substances. There are two types of muscle disease in chronic alcoholics: an acute form is characterized by muscular pain, tenderness, and swelling, with excretion of the muscle protein myoglobin in severe cases; in a less acute form, weakness with little wasting develops over the course of a few weeks. Recovery occurs in both varieties if alcohol is not consumed. Carbon monoxide poisoning can cause damage and death of muscle tissue. Chloroquine, an antimalarial drug also used in the treatment of other diseases, may have a toxic effect upon muscle tissue, as may vincristine sulfate, a preparation employed in the treatment of malignant disease.

Malignant disease. The direct invasion of muscles by primary cancer is more common than the involvement of muscles by metastases (*i.e.*, secondary cancers that develop at a distance from the primary site). Muscle disorders are encountered that result from metabolic diseases associated with cancer. Characteristic weakness and other abnormalities of the shoulder and pelvic muscles and muscles of the trunk have been reported in persons with lung cancer. A linkage between dermatomyositis and cancer has been reported for many years. Indeed, the presence of certain types of muscle disorder, such as those mentioned of the shoulder and pelvic muscles, may be the first indication of cancer.

Aging. Muscle bulk and strength both diminish with the coming of old age, although the loss of strength does not equal that of size. Pathologic changes found in muscles of the aged include reduction in the number of fibres with irregular reduction in their bulk; appearance of pigment; and increase in fat cells and connective tissue.

Primary muscle diseases and disorders. *Myoglobinuria.* Myoglobinuria, the excretion of the muscle protein myoglobin in the urine, is a warning of the serious danger of kidney disease. It occurs as the result of crush injuries to the muscles or of death of muscle tissue because of interference with the blood supply. It also occurs from metabolic disease, polymyositis, and poisoning of the muscle tissues, as from eating contaminated fish. Myoglobinuria can also occur in young men after heavy exertion and in young children experiencing acute infections.

The muscular dystrophies. The muscular dystrophies are a group of hereditary disorders characterized by progressive muscular atrophy and weakness. In most varieties the muscles of the limb girdles—the pelvic and shoulder muscles—are involved.

In contrast to the several varieties of muscular dystrophy that are relatively benign, the Duchenne type, which commonly affects boys in early childhood, steadily grows worse and usually ends fatally in adolescence. The first symptoms include difficulty in walking, rising from the floor, or climbing stairs. Certain muscles, usually of the calves, buttocks, and thighs, become enlarged but weaker (a condition called pseudohypertrophy) in the early stages. Death is from heart failure or respiratory complications.

Duchenne type of muscular dystrophy

The myotonias are a subgroup of the muscular dystrophies. In dystrophia myotonica, a form of muscular dystrophy that affects both sexes, muscular weakness and wasting are accompanied by myotonia, the continuation of muscular contraction after the cessation of voluntary movement. After the eyelids have been firmly shut, for example, by a person affected by this disorder, he may be unable to reopen his eyes instantly. Other features of dystrophia myotonica include premature baldness, emaciation, atrophy of the sex glands, cataracts (opacity of the lenses of the eyes), abnormalities of the heart, diabetes, and mental changes. Disability can be evident in childhood but usually is first noticed in adolescence or early adulthood. The condition becomes progressively worse, although some affected persons survive into extreme old age.

Two other types of myotonia are myotonia congenita, or Thomsen's disease, and paramyotonia congenita. The former is a familial disorder causing widespread muscular hypertrophy and stiffness that is relieved by exercise and made worse by rest and by cold. It becomes evident in early life. The tendency toward a delay in relaxation of muscles after they have undergone a vigorous voluntary contraction tends to diminish in later years. Paramyotonia congenita is characterized by myotonia on exposure to cold and by attacks of generalized muscular weakness.

Although muscular dystrophy cannot be cured, remedial devices and measures, including physical therapy, if not carried to the point of fatigue, spinal supports, and splints for the limbs, may be useful. Prevention of obesity is considered important, especially in the Duchenne type of dystrophy, and infections are promptly eliminated. The identification of carriers of the trait and genetic counselling represent the best hope of reducing the incidence of this group of diseases.

Myasthenia gravis. Myasthenia gravis, which results from a failure in the transmission of nerve impulses to the muscles, is characterized by persistent muscular weakness and a tendency of muscles to be easily fatigued. Estimates of the prevalence of the disease range from one in 10,000 to one in 50,000. It is considerably more common in women than in men.

In milder cases weakness may be limited to a few muscles—for example, those that move the eyes or are concerned with swallowing and with speech; in severe cases fatigability and paralysis may be widespread.

The most active stage of the disease is during the first five years, during which there may be a threat to life; thereafter the risk to life is less, but there may be serious continuing disability. The diagnosis is made from the response of the muscular weakness to cholinergic drugs. These are drugs that liberate acetylcholine, a substance involved in the transmission of nerve impulses. The cholinergic drugs are also used in treatment of the disorder. Removal of the thymus is also resorted to, especially in young women.

Diagnosis and treatment of myasthenia gravis

Specific inflammatory diseases. Bacterial myositis, inflammation of muscle tissues as the result of infection with bacteria, is commonly localized and is the aftermath of injury. Staphylococcus organisms are usually responsible. General indications of infection, such as fever and increased numbers of white blood cells, are accompanied by local signs of inflammation such as reddening, swelling, and warmth. Abscess formation is rare, except in the tropics. In general, bacterial myositis responds to treatment with antibiotics and with minor surgery. Gas gan-

grene, myositis due to infection with *Clostridium welchii*, is a serious wound infection that poses special problems. Treatment involves combatting the infection, removal of dead tissue, and dealing with shock.

Viral myositis may also occur. The best example of this is Bornholm disease, also called epidemic myalgia, devil's grip, and epidemic pleurodynia, caused by the Coxsackie virus. Affected persons recover completely after a brief period of intense muscular pain and fever.

In addition to bacteria and viruses, the muscles may be invaded by animal organisms, including protozoa, the simplest animal forms, and helminths, or worms. Among the better known examples of worm infestation is trichinosis, infection with the roundworm *Trichinella spiralis* as a result of eating infested pork that has not been thoroughly cooked. Reproduction of the worm takes place in the intestines. Larvae migrate from the intestinal walls and bury themselves in muscle tissue. Fever, muscular pains, and sometimes weakness result from the invasion of the muscles. The clinical course of trichinosis is about two months; most affected persons recover, but death can result from invasion of the heart muscle.

Trichinosis

Effects of injury, exercise, and cold. Muscle can be damaged by direct injury, as by a blow, or from tearing of muscle fibres during heavy work or athletic events. Rupture of muscle fibres is accompanied by bleeding and swelling. Traumatic myositis ossificans may result, in which case there is proliferation of connective tissue in the damaged area of the muscle, followed by the appearance of cartilage and bone. The most familiar example is the development of bony tissue in the muscles on the inside of the thighs in horse riders. Crush injuries of muscle can lead to myoglobinuria and kidney damage. In Volkman's ischemic contracture, which occurs shortly after limb injuries, especially fractures in the region of the elbow joint, the fingers and wrist take on a persistently bent position.

The phenomenon of shin splints, known technically as the anterior tibial syndrome, is well known to athletes. In the early stages of training, the pretibial muscles at the front of the lower legs become swollen and painful. The anterior tibial syndrome is a more profound reaction in the same muscle groups after unusual exertion, especially in the untrained. The muscles are swollen, painful and weakened. In extreme cases permanent muscle damage may result, although local surgery may prevent this.

Muscle damage can follow exposure to cold—for example, when the legs are immersed in cold water for long periods ("immersion" or "trench" foot). The muscle changes are due to impaired circulation and also the direct effects of cold on nerves and muscles.

BIBLIOGRAPHY. A comprehensive modern review of the myopathies will be found in J.N. WALTON (ed.), *Disorders of Voluntary Muscle*, 2nd ed. (1969); this work also contains up-to-date chapters on the anatomy and physiology of muscle. Detailed information on all aspects of the myopathies is contained in R.D. ADAMS, D. DENNY-BROWN, and C.M. PEARSON, *Diseases of Muscle: A study in Pathology*, 2nd ed. (1962). For a comprehensive and authoritative review of the anatomy and physiology of muscle, see G.H. BOURNE (ed.), *The Structure and Function of Muscle*, 3 vol. (1960).

(R.A.He.)

Muscle System, Human

The human muscle system comprises the skeletal muscles —those muscles, sometimes called the striated muscles, that are under voluntary control and are concerned with movement, posture, and balance; the muscles, called smooth muscles, that are under involuntary control and are found in the walls of blood vessels and of such structures as the urinary bladder, the intestines, and the stomach; and the heart muscle, or myocardium. An extensive discussion of these various types of muscle is contained in the articles MUSCLE SYSTEMS and MUSCLE CONTRACTION. The present article is primarily concerned with those changes that occurred in man's muscular system as a result of the long evolutionary process that involved his assumption of an upright posture. Because of the interdependence of the skeletal muscle and bones, evolutionary changes in the latter are also touched upon.

Trend of human specialization. The human muscular system inherits the pattern of vertebrates in general and of mammals in particular. Its most distinctive features reflect varying degrees of success in adapting a quadrupedal type of musculature to the task of controlling an upright biped. Human civilization has been created by emancipated hands, liberated from supportive function by profound redesign of the musculoskeletal complex.

Man has added standing to sitting as a posture of composure. The muscular fatigue that prolonged standing entails has been mitigated by drastic changes in the skeleton. The hip joint and the knee joint have been stabilized in an extended position so that man can stand with hip, knee, and ankle axes in vertical array. This ability minimizes the effect of gravity on the limb segments. Modification of the ankle has allowed the heel to touch the ground and thus has increased the area of foot support. The trunk has been brought into erect position over the hip joints by the development of the lumbar curvature, the curvature in that part of the spine that extends from the level of the chest to that of the pelvis.

Skeletal changes

Even during active locomotion man's hands are free for other employment, since the lower limbs have assumed responsibility for transportation of the rest of the body, the head, arms, and trunk.

Muscles of the lower limb. In addition to the increase in amount of muscle in the lower limb that has brought the mass of each leg to one-sixth that of the body, there have been changes in the relative development of particular muscles. Antigravity muscles, such as the soleus of the calf, the vasti of the thigh, and the gluteus maximus of the buttock, have increased in size. Restriction of the base of support to two feet, as during the process of walking, and frequently to one, has magnified the problem of lateral stability. A tendency to fall sideways is avoided by a change in the arrangement of the ilium, the flat upper part of the hipbone, so that the gluteus medius and minimus muscles are attached to it at the side of the body rather than at the back. The gluteus maximus is enlarged for a different reason: so that it can stretch sufficiently in an oblique direction to decrease the rotation of the trunk and diminish its tendency to jackknife forward when the foot on the opposite side makes contact with the ground. Continued strong development of the adductors, the muscles on the inside of the thighs that bring the thighs toward the midline of the body, is caused by their bending and extending the legs.

The human ankle and foot have undergone drastic redesign. The toes have diminished in size and importance. With prehension (grasping) no longer a function of the first metatarsal (the bone directly in back of the big toe) and the peroneus longus muscle, the muscle remains important as a deterrent of splayed foot. The use of the ball of the foot as a major contact area during propulsion has augmented the importance of the muscles that produce inversion about the subtalar axis (*i.e.*, that lift the heel and lower the ball). Included in this group of muscles is the triceps surae, a calf muscle, the unnatural shortness of which can cause disability. The tibialis anterior, a muscle at the front of the lower leg, prevents the front part of the foot from slapping down after the heel has touched the ground and also lifts the toe when the foot is being swung forward during walking. The plantaris has lost its function of tensing the plantar fascia, a dense, fibrous membrane in the sole of the foot, and is noted in man for the pain that follows rupture of its long tendon.

Muscles of the upper limb. The upper limb of man has retained an overall generalized structure with its details adapted to upright existence. Among the primitive features that persist are the clavicle, or collarbone, that still functions as part of the shoulder; ability to twist one of the forearm bones, the radius, around the other, the ulna, so that the palm is turned forward or backward—a process called pronation-supination—and a full complement of five digits in the hand.

Surviving primitive features

In quadrupedal life the thorax, or chest, is suspended between the shoulder blades by a muscular hammock formed by the serratus anterior muscle. In upright sitting and standing, the shoulder girdle is suspended from the

trunk. The scapula, or shoulder blade, floats over the thoracic surface by reason of the arrangement of the fibres of the serratus anterior muscle and the support against gravity provided by the trapezius, rhomboideus and levator scapulae muscles. When the arms are required to push forward against an object at shoulder level, their action is reminiscent of quadrupedal support.

The change in the shape of the chest to emphasize breadth rather than depth altered the relation of muscles in the shoulder region, with increase in size of the latissimus dorsi muscle and the pectoralis major muscle. The human pectoralis minor muscle has forsaken its attachment to the humerus, the long bone of the upper arm, and presumably derives some stability from attaching to the coracoid process, a projection from the scapula, instead of gliding over it.

The retention of a full-fledged thumb by man distinguishes him from his relatives, who swing through the trees. The muscles that can place the thumb in apposition to the other digits are responsible for human grasp, a manoeuvre that precedes both constructive effort and intelligent inspection.

Muscles of the head and neck. The human head is still evolving. Use of the hands for feeding facilitated a reduction in the size of the jaw, and advances in food processing affect dental function and muscles of mastication. Development of coordinated binocular vision occasioned change in the position of the eye sockets and intricate correlation of the ocular muscles. Control of the apertures of the face by muscles derived from the hyoid arch (the second of the cartilaginous or bony arches that encircle the upper end of the digestive tract in more primitive organisms) has been expanded into the use of facial expression for communication. Increase in the size of the brain has contributed to the positioning of the occipital condyles (rounded projections on the occipital bone, at the base and back of the skull) so that the head balances on the vertebral column with minimum muscular effort.

Muscles of the trunk. The organs of the human trunk must function adequately with the trunk horizontal but must be prepared for maximum activity when it is upright. The weight of the abdominal viscera aids inhalation when the trunk is upright but becomes an expiratory force when the body rests on its back. The muscles of the abdominal wall act against gravity when they contract and force the abdominal viscera upward during upright exhalation. Concurrent changes have taken place in the diaphragm and the muscles between the ribs.

The abdominal pressure that acts on the thorax is even more effective at the pelvic outlet. Its muscular closure involves the perineal muscles (which include the sphincters that close the anus and the urethra), derivatives of the sphincter cloacae (the muscle closing the common passage for urine and feces in lower animals) and the muscles of the pelvic floor. The latter originate from the pelvis and insert on the tail. In man the tail has become vestigial, and the pelvic floor muscles have been renamed the levator ani and the coccygeus.

Muscular compression of the abdomen and thorax aids the vertebral column in supporting the body and providing a firm base for upper limb action. Anteroposterior (fore and aft) stability of the trunk is achieved by balancing the flexing action of gravity against back muscles that act to extend the spine. Lateral stability is enhanced by the augmented lever arms provided the spinal muscles by the broadening of the chest.

BIBLIOGRAPHY. A. KEITH, "Man's Posture: Its Evolution and Disorders," *Brit. Med. J.*, 1:451 (1923), a classic summary of adaptation of musculature to upright posture; H. ELFTMAN, "Evolution of the Pelvic Floor of Primates," *Amer. J. Anat.*, 51:307–346 (1932), and E. HUBER, *Evolution of Facial Musculature and Facial Expression* (1931), detailed presentations of the evolution of two important groups of muscles; H. ELFTMAN, "Biomechanics of Muscle, with Particular Application to Studies of Gait," *J. Bone Joint Surg.*, 48A:363–377 (1966), a summary of physiological adaptation of human muscles to upright movement; P.E. KLOPSTEG and P.D. WILSON (eds.), *Human Limbs and Their Substitutes* (1954), a comprehensive review of normal human limbs and their artificial replacements.

(H.El.)

Muscle Systems

All animals must either move to find food or, if they are sedentary, have means of bringing food to themselves. Many invertebrates and all vertebrates depend upon contractile tissue called muscle to move themselves about. Muscles in turn are grouped into coordinated systems for greater efficiency.

General features

DEFINITIONS AND DISTINCTIONS

Irritability, the capacity to respond to stimulation, is a property of all living matter. Irritability in which a cell develops an internal tension that usually causes it, or some part of it, to shorten, bend, or otherwise change its shape is called biocontraction and is a widespread, possibly universal, phenomenon in all living cells. In all cases of biocontraction, chemical energy stored in the cell must be transformed into mechanical energy; but how this is done at the biochemical level and the nature of the cellular structures (organelles) involved are only beginning to be understood (see MUSCLE CONTRACTION).

Highly organized fibrillar or tubular elements occur in the cytoplasm of cells with well-developed contractile capability. The various kinds of muscle systems are only one category of specialized contractile system. Other, more widespread ones include the spindle, a temporary organelle that appears in the cytoplasm of a dividing cell (see CELL AND CELL DIVISION), and cilia and flagella that occur on many plant and animal cells. The rapid and dramatic responses of certain plants, such as the sensitive plant and the Venus's-flytrap, are only remotely analogous to the performance of animal muscle (see STEREOTYPED RESPONSE).

Muscle systems can be defined as biocontractile systems in which the cell, or part of it, is elongated and specialized so that it has the ability to develop tension along the axis of its elongation. Often the tension is manifested by a shortening of the cell against a constant force (isotonic contraction), but the ends of the cell may be attached to structures held firmly apart so that an increasing tension is developed without shortening (isometric contraction). The contractile organelles of the cell are minute threads called myofilaments, which extend parallel to the cell's long axis (Figure 1). These myofilaments are so small (50 Å [Ångstroms] to 100 Å) that they can be detected only with an electron microscope, but often they are grouped into larger fibrillar bundles called myonemes and myofibrils, which can be seen with a light microscope. Although the larger fibrillar elements are not always present in muscular systems, the myofilaments are. Myofilaments are composed of protein complexes, the most common one being actomyosin, although tropomyosin comprises the major part of the myofilaments in certain invertebrates.

SIGNIFICANCE OF MUSCLE TISSUE

The importance of muscle systems to animals has been indicated. Most invertebrates and all vertebrates depend upon muscle action to move about in search of food, mates, and a favourable environment. Larger animals with skeletons also depend upon muscle action to help maintain the proper relationships of skeletal elements needed to protect or to support the animal.

The action of muscles

Muscle action propels blood through animal circulatory systems, and in many species it also controls the diameter of peripheral vessels and hence the distribution of blood to the tissues. Muscle action also moves food into and along the digestive tract, water or air across respiratory surfaces, and discharges excretory and reproductive products. It may play the sole role in these processes, as in most vertebrates, or may be assisted by ciliary action, as in many invertebrates.

In addition to these widespread functions, a number of specialized activities are performed by muscles in certain species. Muscles attached to pigment-bearing cells (chromatophores) are responsible for the rapid colour changes of a squid or octopus; they surround certain glands and thus aid in the discharge of secretions; they move the hair of mammals, thereby playing a role in temperature regulation; they may move sense organs, or they may form an

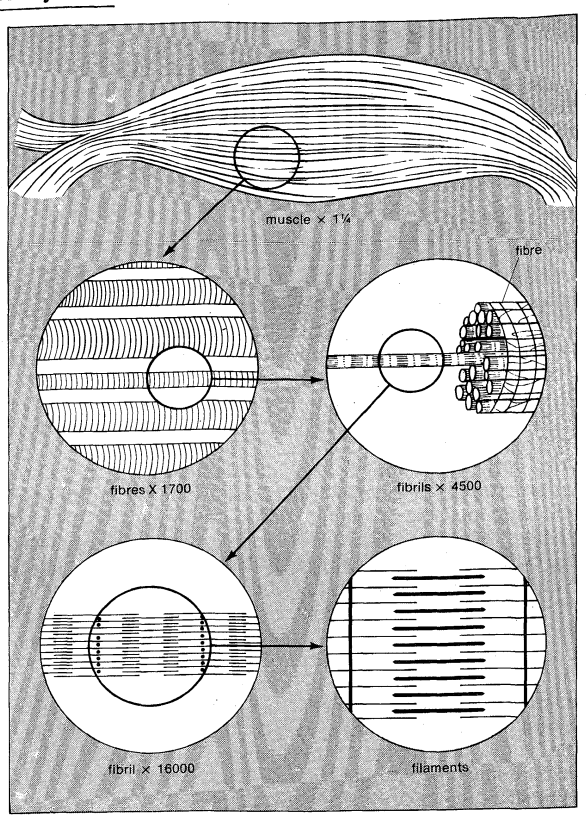


Figure 1: Striated muscle structure in increasingly magnified drawings.
From *Biological Science: Molecules to Man*, Houghton Mifflin Co., Boston, Massachusetts, by permission of the Biological Sciences Curriculum Study

integral part of sense organs, as do the ciliary muscles responsible for accommodation in the eyes of vertebrates and some mollusks.

Muscle contraction is accompanied by electrical phenomena and by heat production, and organisms in fact sometimes make use of these attributes of muscle activity. The electric organs of certain fishes are modified muscular tissue in which the electrical phenomena are accentuated. Body heat derives principally from muscle action. During the isotonic contraction of a typical muscle, only about 40 percent of the energy utilized is converted to work; the rest becomes heat (Figure 2A). In isometric contraction, in which no work other than stretching connective tissues within the muscle is done, most of the energy is released by heat. This heat is simply dissipated by most animals but in the higher vertebrates, particularly birds and mammals, it sustains the high level of metabolism characteristic of these warm-blooded animals. When body temperature falls too low, the rapid-fire involuntary muscle contraction and relaxation called shivering helps to produce heat and restore body temperature.

ARRANGEMENT AND GROSS FUNCTION OF MUSCLE TISSUE

The muscle fibre. Although a few invertebrates have simple muscle systems that involve the specialization of only part of a cell or a few isolated cells, most animals have muscle systems made up of groups of elongated, specialized cells. Each cell is referred to as a muscle fibre. The fibres are bound together by some connective tissues through which blood vessels and nerves pass. In vertebrate muscle, individual fibres have a protein-polysaccharide coating and a network of fine threads that transmit the force of muscle fibre contraction to surrounding tissues. This layer around individual fibres constitutes the endomysium. Groups of fibres may be surrounded by a connective tissue, the perimysium, and arrayed in bundles of various sizes. Finally, the entire muscle, often composed of many bundles, may be surrounded by another connective tissue layer known as the epimysium. Although a number of invertebrate muscles are similarly organized, many are very small and consist of only a few muscle fibres with little connective tissue between them. Collagen, a major component of the connective tissue of vertebrates, is absent in arthropods.

Functional kinds of muscle. Muscles are divided by physiologists into two broad groups: phasic and tonic. Phasic muscles are composed of fibres that are capable of rapid activation and relaxation. They are well adapted to bring about rapid movement of brief duration. The fibres of tonic muscles contract and relax more slowly; they induce slow and sustained movements or hold parts together for a long period of time. While it is often useful to think of muscles as phasic or tonic, it is difficult to characterize a muscle clearly as one or the other, for contraction and relaxation times of different muscles vary widely; furthermore, some muscles may contract phasically at one time and tonically at another.

Phasic and tonic muscles

A typical phasic muscle is usually attached to a skeletal element that can move when the muscle contracts and shortens. In arthropods the muscle fibres attach directly to the body covering, but in vertebrates attachment of muscle to a bone is by connective tissues. The muscle cells may be very close to the point of attachment, in which case the muscle is said to have a fleshy attachment; or the connective tissues may form cordlike tendons or broad, thin sheets known as aponeuroses that pass to the attachment. The end of a muscle attached to the movable element is said to be its insertion; the opposite end, attached to a fixed element, is its origin (Figure 2B). Insertion and origin are convenient descriptive terms, but it must be emphasized that the tension at either end is the same.

Physics of muscle. The skeletal elements to which muscles attach are often lever systems, and muscles exert a force on them that enables work to be done. Most of these lever systems are classified as third order levers (the fulcrum and weight being at extreme opposite ends) because the muscles attach close to the joints (fulcra), whereas the work is done at the farthest end of the lever arms (Figure 2B). Muscles work at a mechanical disadvantage because the length of the force arm (the perpendicular distance between the fulcrum and line of muscle action) is less than the length of the work arm (the perpendicular distance between fulcrum and point of application of the work). The compensatory advantages of this arrangement, however, are compactness of the body and speed of movement. A short contraction of a muscle near the fulcrum induces a quick and more extensive movement at the distal end of the lever. Shifts in the ratio between force and work arms can alter the power and speed relations of the system. A fast-moving animal, such as a horse, has long legs (long work arms) and muscles acting very close to the joints between legs and girdles (short force arms). In contrast, a burrowing animal, such as a mole, has short limbs and muscles that act farther away from the

From (A) R.C. York, and (B) C.A. Villee, W.F. Walker, and F.E. Smith, *General Zoology* (1963), W.B. Saunders Company

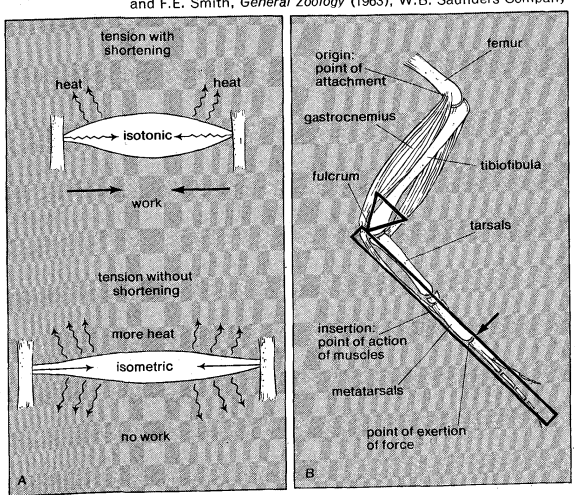


Figure 2: *The work of muscle.*
(A) Comparison of isotonic and isometric contraction. (B) Lever action of muscles of a frog's right hindfoot. Gastrocnemius moves foot in direction of solid arrow.

fulcra; its limbs move more slowly but can develop relatively more power.

Arrangement of the fibres within a muscle is also related to the extent and power of muscle contraction. Seldom are muscle fibres as long as a muscle, but many muscles, such as the biceps in the human arm, are composed of relatively long fibres lying nearly parallel to each other (Figure 3A). Since muscle fibres can contract about one-

From (A) P. Rasch and R. Burke, *Kinesiology and Applied Anatomy*, 4th ed. (1971), Lea & Febiger, and (B) C.A. Villee, W.F. Walter, and F.E. Smith, *General Zoology* (1963), W.B. Saunders Company

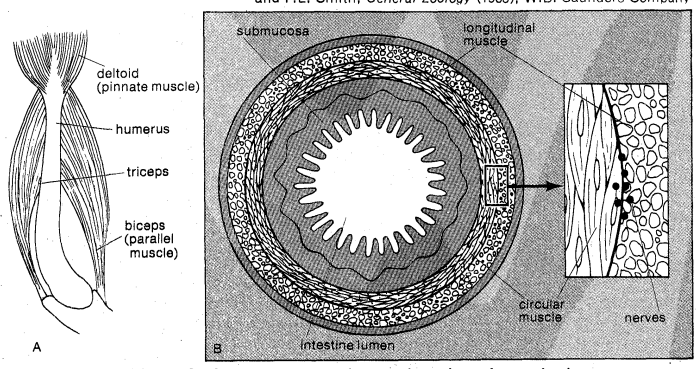


Figure 3: *Arrangement and gross function of muscle tissues.* (A) Lateral view of antagonistic set of phasic muscles of human upper arm. (B) Tonic muscles in the intestinal wall of a mammal.

Pinnate muscle

third of their resting length, this arrangement is suitable to an extensive and quick movement. The deltoid muscle in the human shoulder is said to be pinnate: relatively short fibres attach diagonally onto a tendon that penetrates far into the muscle. A pinnate muscle contains many more and shorter fibres than a parallel muscle of equal mass. This arrangement clearly reduces the extent of muscle contraction but increases the force that it can exert. It must be recognized, however, that not all of the force is effective because the fibres pull at an angle to the direction of tendon pull so that some force is lost. Pinnate muscles also have the advantage of bulging less when the fibres shorten, for much of the shortening is taken up by a change in their angle of pull. Pinnate muscles often occur in places where much bulging would be disadvantageous. Many arthropod muscles, which lie inside a rigid exoskeleton, are pinnate.

Since muscles can only do work by contracting, antagonistic forces must restore the original state. A few muscles work against a nonmuscular force, and so their relaxation permits the opposite action. The phasic adductor muscle of a clam can quickly pull the two halves of the shell together against the force of an elastic spring ligament in the hinge; when the phasic adductor and other muscles relax, the shell opens. Usually, however, muscles are arranged into antagonistic groups so that one will pull in one direction and another in the opposite direction.

Tonic muscle

Fibres in a tonic muscle also may be attached to skeletal elements. In addition to the phasic adductor referred to above, many clams have a tonic adductor that can hold the two halves of the shell together for a long period of time. Usually, however, tonic muscle fibres are arranged in an interlacing mass, as in the foot of a snail, or form layers in the walls of blood vessels, intestines, bladders, and other hollow organs (Figure 3B). The fibres do not have strict origins and insertions; rather, they pull upon each other and surrounding soft tissues so that they change the shape and size of the organs of which they are a part. Often these muscle fibres work against the force of the contents of the organ. When fibres in the wall of a urinary bladder contract, the bladder evacuates its contents. The gradual refilling of the bladder stretches the fibres out again. In the gut, where peristaltic and churning movements are more complex, the muscle fibres are arranged in antagonistic layers. An inner, circular layer in the wall constricts the cavity (lumen) and increases the force on the gut contents. This action, in turn, causes the segment of the gut involved to lengthen, for the force developing on the gut contents can only be released in this direction. The circular layer is opposed by an outer longitudinal muscle layer that shortens the gut segment and widens the lumen.

Muscle contractile systems and contraction

SIMPLE CONTRACTILE SYSTEMS

Contraction in primitive multicellular animals is accomplished by very simple systems that involve either a part of a single cell or, at most, a few cells. Many sponges have simple contractile cells that originate in the epithelial, or surface, layer, among them the myocytes that form a sphincter around the opening through which water is discharged from the central cavity. The epitheliomuscle cells in the polyps of coelenterates constitute another example of a simple contractile system. The main part of each cell is cuboidal and performs a variety of functions, but the base of the cell forms elongated processes that contain contractile elements organized into several relatively large contractile threads called myonemes.

During the course of evolution, animals and their movements have become more complex, and simple contractile systems have been replaced by well-developed muscular tissue. Yet simple contractile systems persist in certain higher animals in which a few isolated contractile cells perform the needed function. Myoepithelial cells of epithelial origin are found in the sweat, mammary, and salivary glands of mammals. They usually are elongate, star-shaped cells 30 to 90 micrometres long and contain myofilaments. They lie directly under the secretory cells of the gland, especially around the terminal part of the gland and around the duct cells. Presumably, they aid in the discharge of glandular secretions.

Persistence of contractile systems in higher animals

COMPLEX CONTRACTILE TISSUES

Complex contractile systems occur in jellyfishes, which swim about and are more active than polyps, and in all groups of animals above coelenterates in the evolutionary scale. These systems take the form of well-developed muscle tissue; that is, integrated groups of elongated cells (muscle fibres) each of which contains a great many longitudinally arranged myofilaments in its cytoplasm.

Vertebrate muscle tissue, which was studied before invertebrate tissue, is classified as striated or smooth according to the presence or absence of cross striations in the fibres when they are viewed with a light microscope (Figure 4). When the spectrum of invertebrate muscle is

From C.A. Villee, W.F. Walker, and F.E. Smith, *General Zoology* (1963), W.B. Saunders Company

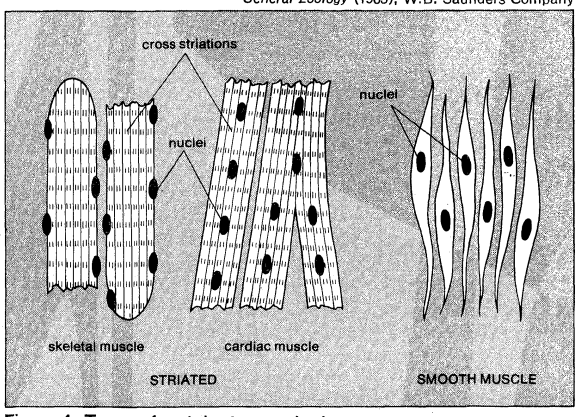


Figure 4: Types of vertebrate muscle tissue.

also included, it is clear that this classification is an oversimplification. "Striated" and "smooth" are simply broad morphological categories of convenience. Smooth muscle in particular shows a wide divergence of morphological and physiological attributes.

Striated muscle. In striated muscle the myofilaments are usually grouped into a number of larger myofibrils, the filaments aligned so that dark and light staining bands alternate along the length of each fibril. Since a cell contains many myofibrils and the dark and light bands of adjacent fibrils are usually lined up, the entire fibre has a cross-striated appearance (Figure 4). In general, striated muscle is present in both invertebrates and vertebrates in

which phasic contractions occur. Bands of striated muscle are found in the more rapid swimming jellyfishes, and striated muscle occurs sporadically in most of the higher invertebrate phyla. Nearly all of the musculature of arthropods is striated. Among vertebrates, striated muscles are found wherever rapid or continuous motion is usual; *e.g.*, in structures used in locomotion, breathing, and pumping blood. Muscle tissue in the heart wall of vertebrates differs somewhat from most striated muscle and is referred to as cardiac muscle. All other vertebrate striated muscles are known as skeletal muscles; they are closely associated with the skeleton or evolved from muscles that were. Invertebrate striated muscles that have been carefully studied resemble the skeletal muscles of vertebrates in their essential structure.

Types of striated muscle

Skeletal muscle is composed of exceedingly long, cylindrical fibres the ends of which are blunt and rounded. Their length ranges from about 1 millimetre to the length of the muscle itself. Most have a diameter ranging from 0.01 to 0.1 millimetre, although fibres in the flight muscle of some insects have diameters of up to 2 millimetres.

The fibres usually contain many elongated nuclei lying just beneath the outer plasma membrane, or sarcolemma. The cytoplasm of the fibre, called sarcoplasm, is largely filled with myofilaments grouped into myofibrils the banded segments of which are designated with letter names (*e.g.*, A, I, H, etc.). A staggered arrangement of thin actin myofilaments and thicker myosin ones forms the basis for the banded pattern. Contraction involves a sliding of these filaments past one another (see MUSCLE CONTRACTION).

In addition to the myofibrils, the sarcoplasm of striated and other types of muscle includes the components characteristic of most cells, although certain ones are reduced and others more obvious. As would be expected, cell organelles called mitochondria are numerous, for they contain the enzymes for oxidative metabolism and therefore play a vital role in energy release. They are particularly abundant in very active muscle and comprise up to 40 percent of muscle weight in the flight muscle of certain insects. Energy reserves in the form of glycogen granules are also abundant.

In most animals the needed oxygen is supplied by capillaries or vascular spaces that are close to the muscle fibres, but in insects oxygen is brought directly to the cells by tubules. Myoglobin, which can store oxygen, is abundant in red fibres, which in general contract and relax slower than white fibres that contain less myoglobin. A muscle may be predominantly one or the other. Certain of man's hip muscles, such as the gluteus maximus, which undergo sustained contraction in the maintenance of posture, are predominantly red muscle; whereas certain finger muscles, which contract less frequently and faster, are predominantly white muscle.

Red and white fibres

Cardiac muscle, the specialized striated muscle of the vertebrate heart, is organized into distinct, elongated, cellular units, each with its own nucleus located near the centre of the cell (Figure 4). The cells are united firmly with each other end for end by specialized junctions known as intercalated disks. Many of the cells branch and unite with other cells so that the whole heart musculature forms a network. Thick and thin myofilaments are interrelated with each other in the same way as in skeletal muscle, so that the banding pattern is identical. The myofilaments, however, are not organized into well-defined myofibrils but rather tend to form a continuous mass of parallel filaments broken up by deep furrows and clefts in which the other sarcoplasmic constituents are located. Cardiac muscle has more sarcoplasm than skeletal muscle; more mitochondria and glycogen are present, providing repeated contractions without an oxygen debt.

The ventricles of the vertebrate heart contain a network of conducting tracts that enable a wave of excitation to spread quickly to all parts. These tracts are composed of modified cardiac muscle known as Purkinje fibres, named for the pioneer experimental physiologist. Purkinje fibres are much larger than typical cardiac muscle cells, lacking intercalated disks, and having fewer myofilaments, which are grouped into myofibrils.

Classic smooth muscle. In all smooth or nonstriated muscle tissue, myofilaments are not arrayed in such a way as to give a cross-banded appearance. In general, smooth muscle is found both in invertebrates and vertebrates where tonic contraction occurs, but smooth muscle is also used by many invertebrates in certain types of phasic activity.

The type of smooth muscle found in the viscera of vertebrates, which was the first to be studied, may be referred to as classic smooth muscle. It is found in the walls of visceral organs, ducts, and blood vessels, in the iris and ciliary body of the eye, and attached to the hair follicles in the skin of mammals. A very similar type of smooth muscle is also found in many invertebrate groups. Occasionally, smooth muscle fibres are grouped into small muscles; but usually they are arranged into complex, interlacing networks, as in the feet of mollusks, or in distinct layers, as in the gut wall of vertebrates.

The cells of classic smooth muscle are spindle-shaped fibres, wide at the centre where the single nucleus is located and tapering to a point at each end (Figure 4). Although the fibres in the pharynx retractor of the garden slug can be as long as 20 millimetres, most classic smooth muscle fibres are much shorter than striated muscle fibres —on the order of 0.2 millimetre in the gut of vertebrates and as short as 0.01 to 0.02 millimetre in the walls of the blood vessels and in the muscles of many invertebrates. The fibres are staggered with respect to each other so that the thick part of one is adjacent to the thinner parts of another.

Spindle-shaped cells

Very thin, parallel myofilaments fill much of the cytoplasm. In invertebrate smooth muscle the filaments are not grouped into myofibrils, but myofibrils can be detected in vertebrate smooth muscle by using special fixing and staining methods. Smooth muscle in the pharynx retractor of the slug contains myofilaments of two sizes and in this respect resembles striated muscle. A similar situation has been reported in several other invertebrates, but most smooth muscle appears to have only very thin filaments with, in some cases, a few widely scattered thick filaments. While actomyosin has been extracted from smooth muscle filaments, actin and myosin have not been localized in any particular filament or part of a filament; their arrangement must be different from that in striated muscle.

Helical smooth muscle. Helical smooth muscle is a specialized type of smooth muscle in which the myofilaments are grouped into myofibrils that have a helical course through the peripheral part of the muscle fibre. This type of muscle is limited to certain invertebrate tissue, notably the body wall of annelid worms, appendages of some sea slugs, most musculature of cephalopods, the mobile arms of certain starfishes and sea lilies, and the heart of some mollusks. In most annelids the cells appear rather flat in cross section, but in other cases they are circular. Two sizes of myofilaments have been found in cephalopods. An actomyosin-like protein complex has been extracted, but it is not known whether the proteins are found in different filaments.

The significance of the helical arrangement of the myofibrils is uncertain. Theoretically, such a pattern should enable the muscle fibre to shorten to a greater extent than it could if its myofibrils were parallel, because a shortening of the myofibrils would also impart a steeper pitch to them. Organs having this type of muscle appear to be capable of quicker and more extensive contraction than organs with classic smooth muscle.

Paramyosin smooth muscle. Paramyosin smooth muscle is the most unusual and least understood of all types. It is limited to mollusks and has been studied most intensively in clams. The adductor muscle or muscles of a clam usually consists of two parts: a tonic part that holds the shell halves together against the tension of the springy hinge ligament and a less tonic, or phasic, part that closes an open shell. The tonic part is always paramyosin muscle, and the less tonic part is often paramyosin muscle, also.

The cytoplasm of paramyosin muscle contains long filaments with a relatively large diameter of 0.1 micrometre equal to that of small myofibrils. Each filament is com-

posed of many longitudinal, ribbon-shaped paramyosin elements tightly packed together. Paramyosin is composed largely of a unique, water insoluble tropomyosin complex. Actomyosin is present, but in small amounts. Paramyosin has an exceptional capacity to remain for long periods in the contracted state.

MUSCLE CONTRACTION

Muscle contraction is discussed in detail elsewhere (see MUSCLE CONTRACTION), but a summary of certain aspects is necessary to understand the differences encountered in the diverse muscle systems of animals. The membrane of a resting muscle fibre has a resting electrical potential that derives from the differential distribution of positive and negative ions upon its inner and outer surfaces (see also BIOELECTRICITY). Excitation involves an increase in the permeability of the membrane to certain ions so that they move quickly through the membrane and reduce the resting potential. A sufficient reduction leads to the initiation of the active state, by which is meant the totality of chemical events within the fibre that leads to the development of tension in the contractile elements. It is important to recognize that two different sets of phenomena are involved: (1) the excitation of the fibre membrane, which is measured by electrical changes, and (2) the development of the active state, which is measured by the mechanical tension.

Excitation and tension

In living systems, activation typically results from stretching, action of certain hormones, and, most importantly, nerve impulses. A single nerve impulse is often not enough to trigger a fibre contraction, but a rapid succession of two or more impulses may influence each other in such a way as to initiate the active state. The first impulse may cause changes that enable a second impulse that quickly follows it to have a greater effect (facilitation), and/or the effects of succeeding impulses may simply be added to those of earlier impulses if they arrive before the effects of the earlier ones have worn off (summation). Facilitation and summation can occur both in the electrical phenomena at the membrane and in the development of tension. Certain fibres contract maximally only if they are stimulated repeatedly and while the active state induced by earlier stimuli is still high so that mechanical facilitation and summation can occur. The active state and contraction eventually reach a plateau in which the fibre is in a state of maximum contraction known as tetanus. Since maximum tension develops gradually and on repetitive stimulation, this method of fibre activation is referred to as a "slow system."

All-or-none response

Other fibres do not respond at all until an action potential is generated, and then the response is maximal (all-or-none phenomenon). Since the active state quickly reaches its maximum without mechanical facilitation and summation, this type of fibre activation is described as the "twitch" or "fast system." The fibre membrane has a brief refractory period during which another action potential cannot be induced.

Muscle tissues differ greatly in their physiological properties and pattern of innervation. Most vertebrate striated muscle belongs to the fast system, for each fibre receives only a single neuron terminal; but there are cases, primarily among amphibians, in which fibres receive the multiterminal innervation characteristic of the slow system. In many invertebrates the same muscle fibre may receive terminals from neurons belonging to both the fast and slow systems (polyneuronal innervation).

Cardiac and classic smooth muscle resemble the fast system in being activated by nerve-induced action potentials, but each fibre does not receive a neuron terminal. Neurons terminate on only a few fibres in the pacemaker region of hearts, but terminals are more widespread in smooth muscle. Excitation spreads from active fibres to adjacent ones through areas of close membrane contact. Spread may be through typical muscle tissue, or it may be assisted, as in the vertebrate heart, by modified muscle tissue. Smooth muscle displays a gradation of tension development, but individual twitches cannot summate in cardiac muscle because the active state elicited by one stimulus has nearly decayed before the muscle can be reactivated (for further information on the heart muscles of invertebrates and vertebrates see MUSCLE CONTRACTION).

Invertebrate muscle systems

EVOLUTION OF INVERTEBRATE MUSCLE SYSTEMS

A bewildering array of structural and functional diversity is found in the contractile mechanisms of invertebrates, yet the main trends in the evolution of the muscular systems can be traced through major representative stages. Morphologically simple contractile mechanisms make their first appearance in single-celled organisms such as Protozoa. *Vorticella*, for example, has a contractile thread (myoneme) in its stalk that enables it to pull away quickly from irritating stimuli (Figure 5A). Other simple mechanisms are found in the sponges. Flattened pinacocytes, which cover much of the sponge surface, can round up slightly and compress the body if the skeleton is not too rigid.

Simple contractile devices

More highly organized contractile tissue is found in the coelenterates, such as jellyfish and sea anemones, in which many surface, or epithelial, cells are modified as epitheliomuscle cells and have elongated bases containing two or more myonemes (Figure 5B). The processes of the outer layer of cells parallel the longitudinal axis of the polyp, whereas those of the antagonistic inner layer lie tangential to this axis. In slow-moving polyps, such as *Hydra*, the processes resemble classic smooth muscle, but in some of the faster moving jellyfishes the processes form bands of striated muscle. Sea anemones have well-developed retractor muscles in the folds that divide the gastrovascular cavity. Other muscle processes mediate the movements of the tentacles and mouth.

Mesodermal muscle tissue, completely independent of epithelial cells, is first found in acoelomate (lacking body

From (A) Parker and Haswell, *A Textbook of Zoology*, St. Martin's Press, Inc., Macmillan & Co., Ltd., (B) W. Andrew, *Textbook of Comparative Histology* (1958), and (C,D,E) *Invertebrate Zoology* by Paul Meglitsch. Copyright © 1967 by Oxford University Press, Inc. Reprinted by permission

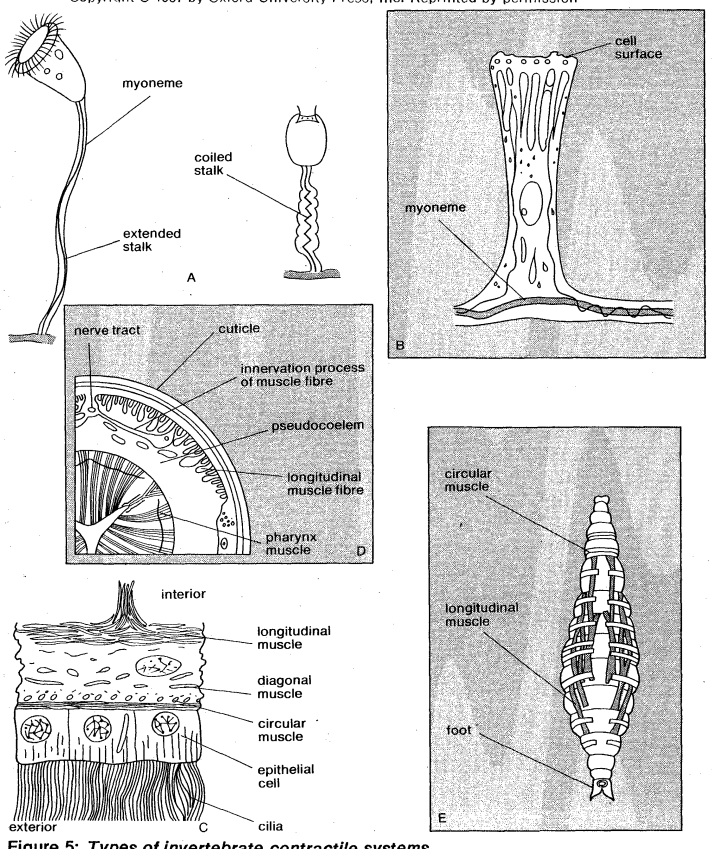


Figure 5: *Types of invertebrate contractile systems.* (A) Contractile stalk in protozoan *Vorticella*. (B) Epitheliomuscle cell of coelenterate *Hydra*. (C) Longitudinal section through ventral portion of body wall of freshwater flatworm. (D) Quadrant of cross section through a nematode *(Ascaris)* at level of pharynx. (E) Muscle bands in body wall of rotifer.

Muscle layers

cavity) worms. Thin layers of elongate muscle fibres lie beneath the epidermis of the free-living flatworms and especially in those species that are too large to rely solely on cilia for locomotion. Often an outer circular layer of muscle is underlain by a diagonal layer and a still deeper longitudinal layer (Figure 5C). Other muscle fibres penetrate deeper tissue, and still others mediate the movements of the proboscis and copulatory organs. Most of the fibres are smooth, but ones with simple striation patterns are occasionally found. Muscles are particularly well developed in the parasitic flatworms, in which they form thick layers in the body wall, and are also associated with the hooks and suckers by which the organism attaches into the tissue of its host. The tail of the cercarian (tadpole-shaped) larva of flukes is the most intricate muscular locomotor organ found in this phylum. A circular layer of smooth muscle fibres lies beneath the cuticle, and a diagonal layer of striated fibres beneath the circular layer. The tissue is very rich in mitochondria, suggesting an efficient energy transfer system.

Pseudocoelomates have muscles in the body wall and muscles associated with certain bristles and with copulatory organs. A muscle layer also begins to become associated with the gut, which is nonmuscular in lower forms. Gut muscles are particularly well developed anteriorly in association with complex mouthparts and a sucking pharynx. Nematodes have only one layer of longitudinal muscle fibres in their body wall (Figure 5D). Fibres contracting alternately on opposite sides of the body produce the characteristic lashing movements of these worms. The nonstriated contractile fibrils of these cells are concentrated in the periphery of the cells beneath the epidermis. These cells are of a unique type in that nerve extensions do not go to them, rather the cells themselves have long cytoplasmic innervation extensions through the pseudocoel to one of four epidermal nerve tracts. Related animals called rotifers do not have a continuous subepidermal muscle layer but rather have discrete bands of fibres (Figure 5E). Contraction of circular bands increases pseudocoelomic fluid pressure and lengthens the body; contraction of longitudinal bands, which extend across one or more cuticular rings, shortens the body or retracts the head and foot. Subepidermal muscle bands are particularly well developed in the primitive creeping species in which one finds an amazing complexity of little muscles, each composed of only a few fibres, both smooth and striated.

Body wall and gut musculature

Higher invertebrates have a coelom (body cavity), except where it has been secondarily reduced, and muscular tissue is more clearly separable than in lower species into parts associated with the body wall and gut. Body wall musculature is particularly well developed. Most coelomates also have circulatory systems in which fluid is propelled by the muscular pulsation of blood vessels or by one or more "hearts." Pulsating blood vessels are rare in lower invertebrate groups.

On the basis of the fate of an embryonic opening called the blastopore, coelomates can be divided into two groups: deuterostomes, in which the blastopore becomes the anus; and protostomes, in which the blastopore becomes the mouth. Most invertebrate members of the deuterostomous evolutionary line are sessile or slow-moving animals in which the musculature is usually not very elaborate. Circular and longitudinal muscles are found in the body wall of echinoderms. Many extend between the calcareous plates and move the arms. These muscles are not well developed in a sluggish starfish, but they are important in the highly mobile arms of serpent stars. In the latter case the muscle is of the helical smooth type, but usually it is classical smooth muscle. Other muscles are associated with the small, nipping pedicellariae (small footlike organs) that keep the surface clean; with the spines of sea urchins; with the ampullae (pouches) and tube feet of the ambulacral system; and with the digestive tract. Sea urchins feed by scraping material from rocks with a device known as Aristotle's lantern, consisting of an intricate framework of 40 bony elements supporting a series of "teeth" and an equally complex array of small muscles.

Most protostomous invertebrates, which include the mollusks, annelids, and arthropods, are very active animals with complex muscle systems. Muscles associated with locomotion and with the skeleton are particularly interesting. In mollusks the body wall is responsible for locomotion; and parts of it, such as the foot or mantle, are very muscular. Snails move by slow rhythmic waves of the foot; bivalves thrust their foot into the substrate, the distal end swells as an anchor, and then the rest of the body is pulled toward the end of the foot. The complexity of foot action requires a complex network of interlacing muscle fibres and not simple layers. Muscle action is responsible for the rhythmic waves and for foot shortening, but the extension of a clam's foot is brought about largely by the swelling of erectile-like vascular tissue. In bivalves an elastic ligament in the hinge opens the shell, but the two halves can be quickly drawn together and held together for long periods of time by the action of one or more adductor muscles. The adductors must be capable of both phasic and tonic action, and an unusual type of muscle tissue is present. Most is paramyosin smooth muscle, and this appears to be responsible for both phasic and tonic action in some muscles. In other cases the phasic adductor, or phasic part of a single adductor, is a different type of tissue. The phasic adductor of the scallop, which swims by rapidly clamping its shells together, is striated muscle. Squids, the most active of mollusks, can dart quickly by the contraction of longitudinal and circular layers of muscle in their mantle. Their action alternately takes water into the mantle cavity and expels it as a jet through a funnel. These muscles are of the helical smooth type.

Complex muscle systems

Most annelid body wall muscles are also of the helical smooth type and have a ribbon shape. Earthworms make an extensive use of a hydraulic coelom in locomotion. Contraction of an outer layer of circular muscle fibres in the body wall compresses the coelomic fluid and extends the worm; contraction of a deeper longitudinal layer of fibres shortens the worm. Many marine worms creep and swim by lateral trunk undulations and finlike parapodial movements. Longitudinal muscle bands are particularly prominent in these worms, and parapodial muscles are complex. Striated muscles occur in part of the gut (the gizzard) in certain species.

Arthropods have an armoured body composed of many segments and jointed appendages. As in vertebrates, muscles act across each of the joints so that the muscles are very numerous. Individual muscles may consist of only a few fibres. Arthropod muscles, even gut muscles, are nearly always striated. There is little connective tissue between the fibres that attach directly to the epidermis or to the overlying exoskeleton. Often the attachment area and mechanical advantage of a muscle is increased by its insertion onto a lever-like inward extension of the exoskeleton known as an apodeme (Figure 6).

The basic arrangement of arthropod muscles can be seen in one of the legs. The base of the leg articulates with the body by a surface layer, or flexible cuticle, allowing a wide range of movements. Many muscles attach to this

From C.A. Villee, W.F. Walker, and F.E. Smith, *General Zoology* (1963), W.B. Saunders Company

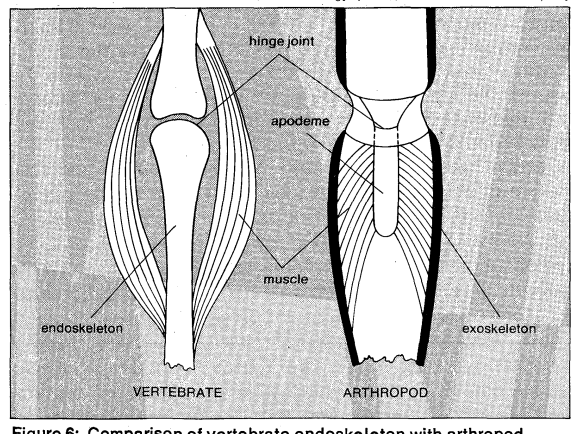


Figure 6: Comparison of vertebrate endoskeleton with arthropod exoskeleton, showing arrangement of muscles and skeleton at a joint.

part of the leg. More distal limb segments, however, are usually simple hinge joints capable of moving only in a single plane; they are operated by a single pair of antagonistic muscles. Muscles also attach directly upon the bases of the wings and move them up and down in those insects having slow wing movements (*e.g.*, butterflies, cockroaches, and dragonflies). The flight muscles of mosquitoes and other insects with very rapid wing movements attach onto parts of the thoracic skeleton and move the wings indirectly by rapidly altering the shape of the thorax.

FUNCTION AND REGULATION OF MUSCLE ACTION

To be of value to the organism, muscles must be capable of making a response appropriate to the changes in the internal and external environmental conditions. This capability, of course, requires that the action of different muscles be integrated so that they interact in an adaptive way. It also requires that single muscles be capable of giving varied responses with respect to speed of contraction and relaxation and the degree of tension development. Conditions may call for a quick contraction of short duration to escape a predator or to catch prey, or, for a slow sustained contraction to maintain posture; they may call for a weak contraction or for a very powerful one. Gradation of muscle action with respect to these variables may be a function of the muscle tissue itself, of the way it is stimulated, or of both.

Integration of muscle action

The contractile elements of sponges are independent effectors capable of responding directly to environmental stimuli. There is a great deal of self-regulation in such effector cells, for their action is appropriate to the nature and intensity of the stimulus. As animals and their responses become more complex during the course of evolution, regulation of muscle activity comes increasingly under the control of the nervous system. Some independent muscle activity is retained, however, in certain higher organisms. Coelenterate contractile tissue can give a local response to a local stimulus. Muscles in insect spiracles that control the passage of air in and out of the tracheal system are particularly sensitive to carbon dioxide and relax when the gas reaches certain levels; gut and heart muscles often have an inherent rhythm and are only partly under the control of the nervous system.

The first stage in the evolution of nervous control of muscle action is seen in the coelenterates. Neurosensory cells, some receptor cells, and a primitive nervous system are present. Impulses can travel from a point of stimulation in all directions and elicit a diffuse and general response. Certain locomotor and feeding activities, however, require an ordered sequence of specific and graded actions, so it is clear that some integration and regulation also occur. Most invertebrates above the coelenterate level have well-developed nervous systems that make possible a high degree of integration and control of muscle activity.

Coelenterate muscle. Despite the apparent simplicity of the neuromuscular apparatus of coelenterates, certain muscles of these animals are capable of making both slow and fast responses and responses of different tensions. Sometimes these differences involve a different pattern of innervation. Some polyps and jellyfishes have two nerve nets, one responsible for rapid and specific conduction and contraction and one for a slower and more general contraction. It is possible, however, in some cases that both types of responses result from a different pattern of stimulation through the same nerve net. A single stimulus does not induce a contraction in the muscle that retracts the mesentery (a tissue partition) of a sea anemone but causes a facilitation in the muscle, enabling a second stimulus following at a one-second interval to elicit a contraction. Succeeding stimuli at one-second intervals induce further contractions that rapidly summate to a maximum. In addition, stimulating the same muscle with stimuli of the same intensity but at a rate of one every three to five seconds induces a slower contraction the tension of which develops gradually over a period of minutes. A similar situation is found in the sphincter of the sea anemone. Although there is a fast and slow system physiologically in these muscles, there is no clear morphological evidence for the dual pattern of innervation usually found where there are both types of response. Some anemone muscles show only the slow pattern. In these cases the muscle fibres are supplied by relatively few neurons, and it is probable that the slowness of the response reflects some conduction of excitation from one muscle fibre to another. Muscles showing fast and slow responses are supplied by a much richer nerve net containing more neurons. It has been proposed that low frequency stimulation and the slow response involve some fibre-to-fibre conduction and a gradual recruitment of more contractile elements, whereas more rapid stimulation activates more fibres directly.

Stimulation via nerve nets

Arthropod postural muscles. Slow, fast, and intermediate rates of contraction are developed to a high degree in arthropod postural muscles. Limb muscles of decapod crustaceans, such as shrimps, lobsters, and crabs, have been studied most intensively. Only a few neurons need supply a muscle because the neurons branch extensively. One neuron may supply muscle fibres in several different muscles, and within a single muscle each neuron supplies many muscle fibres with multiterminal endings. A wide range of responses is possible because certain neurons have different effects and most muscle fibres are supplied by more than one type; *e.g.*, in muscle from the claw of the rock lobster (*Panulirus*) 38 percent of the fibres receive branches from a single excitatory neuron; 26 percent from two excitatory neurons; 29 percent from three excitatory neurons; and 4 percent from four different excitatory neurons.

In addition to receiving one or more excitatory neurons, each muscle fibre in crustacean muscle also receives one or two inhibitory neurons. Stimulation of these slows down, reduces, or may completely prevent muscle contraction.

Final responses in decapod limb muscles

The final response of a decapod limb muscle depends on the rate of stimulation of different types of neurons and the interactions between them. A continuum of contraction is possible from fast to slow, short to long, weak to strong. This sytem is a remarkable adaptation, enabling a few neurons and a relatively small number of muscles to give a range of responses comparable to that seen in the far larger systems of vertebrates.

Insect postural muscles have a pattern of innervation somewhat similar to that of crustaceans except for the absence of inhibitory neurons. The muscle fibres in the jumping muscles of the hindleg of a grasshopper are adapted to give a quick, maximal response to a single stimulus; but muscle fibres in many other muscles are supplied by a fast and a slow neuron, which may terminate in the same junctions. The number of neurons is low, and cases are known where only two neurons supply an entire muscle. As in crustaceans, fewer impulses are needed to activate the muscle through the fast system than through the slow, and tension development is more rapid. In some cases a third nerve fibre to the muscle fibres appears to increase the tension elicited by the stimulation of the other neurons.

Insect resonant flight and tymbal muscles. The indirect or resonant flight muscles of flies and of bees and wasps and the sound-producing, or tymbal, muscles of cicadas are of special interest because they contract at frequencies far greater than those in any other muscles. These muscles move the wings indirectly by distorting the body wall. High-speed frequencies can be determined by the sound pitch generated. The whine of a mosquito indicates a wing beat of over 600 per second, and tymbal membranes vibrate at maximum frequencies of nearly 7,000 cycles per second. Such frequencies far exceed the rate at which muscle contraction can be caused by nerve-induced action potentials. At best, several milliseconds are required for each action potential and its attendant refractory period. Moreover, rapid stimulation through nerves usually leads to a fusion of twitches into a tetanic contraction. Contraction of bumblebee flight muscles tetanize at a nerve impulse rate of only 40 per second. Clearly some other mechanism must be involved in the very high contraction rates of indirect flight and tymbal muscles.

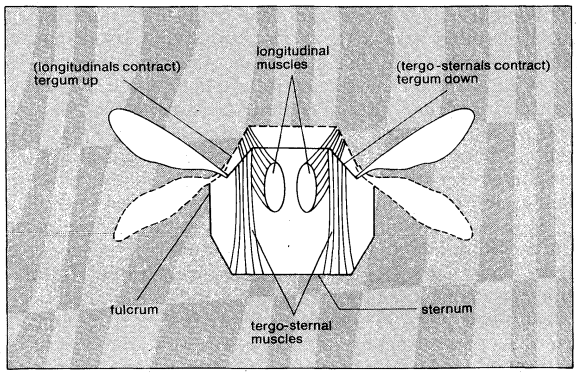


Figure 7: Muscles involved in insect flight (see text).
From C.A. Villee, W.F. Walker, and F.E. Smith, *General Zoology* (1963), W.B. Saunders Company

Mechanics of flight muscles

Mechanical factors play a very important role in the activity of these muscles. The major indirect flight muscles are a pair of vertical ones that extend between the dorsal plate (tergum) of the wing-bearing segment and the sternum (Figure 7). These muscles contract nearly isometrically because an increase in tension must develop in them before the tergum moves. When a certain tension is reached, the elastic tergum "buckles in" and pushes the wing base to which it is attached ventrally. Since the wing is hinged on a lateral section of the body wall (lateral pleural plate), a downward push on its base raises its distal end. The sudden buckling of the tergum releases the tension in the vertical muscles, permitting the tergum to snap back to its original position, lowering the wings and stretching the muscles as it does so. This stretching, rather than a nerve impulse, reactivates the muscle and elicits another contraction. Recovery is assisted by an antagonistic set of longitudinal muscles that attach onto elastic tergal processes extending inward at the anterior and posterior ends of this thoracic segment. These muscles act in a similar way; they are stretched out and activated by the contraction of the vertical muscles, and their contraction assists in bowing out the tergum and stretching the vertical muscles. In the tymbal organ, a tymbal muscle pulls a chitinous membrane inward beyond a point of stability. It snaps inward, releasing the tension in the muscle, and then snaps outward through its own elasticity and restretches the muscle.

Contractions induced by mechanical stretching of the muscle continue as long as the muscle remains in an active state and can follow each other quickly because the active state, unlike the action potential, has no refractory period. Nerve impulses and action potentials are not needed for every contraction but are only needed to maintain the contractile elements in an active state. There may be one action potential for every 20 wing beats, or, in insects having slow wing movements, one action potential for each wing beat.

Mollusk muscle. Separate fast and slow systems exist in the muscles of many other invertebrate groups, but the basis for the two systems is not well understood. Some of the complexities involved can be appreciated by considering the muscles of certain mollusks.

The fast system of the squid

Squids, which rely on rapid movements to catch food or to escape prey, have a fast system in their mantle muscles. These muscles are of the helical smooth type and are innervated by ten giant nerve axons. Their stimulation leads to a rapid and synchronous contraction of the mantle. A single stimulus through one of the axons elicits both a maximal electric response and twitch in the muscle fibre it supplies. There is no electrical or mechanical facilitation and summation, so that a more rapid rate of stimulation does not increase the tension, although it can lead to a brief tetanic contraction. In addition to this fast system mediated by the giant axons, smaller neurons supply mantle fibres and induce slow, graded contractions. One stimulus induces a small electrical difference at a junction of nerve and muscle and a contraction that facilitates subsequent stimuli that build upon it. A maximum tension is reached, however, at a stimulus rate of only a few per second; the degree of gradation, therefore, is not nearly as extensive as in arthropods. It is not clear whether the fast and slow systems work through the same or different muscle fibres. Inhibitory fibres have not been found in mantle muscles, but they have been demonstrated in the muscles of pigment cells of cephalopods.

Fast (phasic) and slow (tonic) systems are also present in the adductor muscle of bivalves, and here different muscle fibres, often with a different tissue structure, are sometimes involved. The phasic portion of the adductor may be striated as in the scallop, which swims by rapidly clamping its shells together. Fast, synchronized electric responses have been recorded from striated fibres of the scallop. The tonic portion of the adductor is paramyosin smooth muscle and is remarkable for its ability to remain contracted when only one stimulus every few seconds is delivered.

In other cases both phasic and tonic activity have been observed in the same paramyosin muscle. An example is the byssus retractor muscle of the mussel, which pulls on the byssus threads, which the animal secretes to cling to its substrate. Stimulation with alternating current leads to a brief phasic contraction, but stimulation with direct current leads to a sustained tonic contraction. Tonic contraction can be quickly terminated by stimulating inhibitory or relaxing nerves or by applying certain drugs.

The maintenance of tonic contraction for a long period has posed a particularly puzzling problem. In life, the byssus contractor can remain contracted for weeks and relaxes only when the animal changes its position. This prolonged contraction may reside in unusual elastic properties of the muscle.

Or it may be that paramyosin changes its state when tonically stimulated and that the increased viscosity characteristic of contraction can "set" and hold its changed state. Paramyosin muscle clearly has unusual elastic properties and relaxing capabilities, but a complete answer to its capacity for sustained contraction is not yet known.

The vertebrate muscular system

EMBRYONIC DEVELOPMENT AND DIVISIONS OF THE MUSCULAR SYSTEM

Except for the ectodermally derived muscles of the iris of the eye and the myoepithelial cell of certain superficial glands, vertebrate musculature originates from the middle layer, or mesoderm, of the embryo. In areas where muscle is destined to develop, mesodermal cells, many of which are at first loosely associated star-shaped cells from loose tissue termed mesenchyme, begin to elongate and myofibrils begin to appear within them. These early muscle-forming cells, known as myoblasts, differentiate into smooth, cardiac, and striated muscle fibres. There is evidence that multinucleate striated fibres derive from a fusion of myoblasts, but nuclear division without cytoplasmic division may also occur.

Mesodermal origin of somatic and visceral muscles

Several mesodermal regions can be recognized in an embryo, and these give rise in lower vertebrates such as fishes to the two major groups of muscles recognized in comparative studies: somatic and visceral muscles. The most dorsal mesodermal region lying along the back on either side of the developing neural tube gives rise to the somites. These take the form of a series of paired segments that extend from the region of the eye into the tail, with a gap in the series adjacent to the developing ear. Most of each somite develops into a block of premuscular tissue known as a myotome, and most of the myotomes extend laterally and ventrally beneath the skin toward the midventral line of the embryo.

Musculature derived from the somites is called somatic musculature. It lies in the body wall and appendages and is innervated by somatic motor neurons that travel in the dorsal roots of spinal nerves and in those cranial nerves believed to be serially homologous to the dorsal roots. All somatic musculature becomes striated, and most of it is associated with the skeleton. The mesodermal region just below the somites constitutes the kidney-forming mesoderm; most of it is destined to form urogenital structures. Lateral plate mesoderm lies ventral to the nephrogenic mesoderm. It is not segmented but is cleaved by the body

cavity, or coelom, into an outer, or somatic, layer and an inner, or splanchnic, layer. Although the somatic layer forms only the outer lining of the coelomic cavity in lower vertebrates, it contributes in higher vertebrates to the cells that give rise to the somatic musculature in the lateroventral part of the body wall and appendages. The splanchnic layer gives rise to the visceral muscles. These lie in the wall of the gut and derived organs, in blood vessels, and in the heart. Muscles in certain urogenital ducts, which come from nephrogenic mesoderm, are also considered to be visceral. Visceral musculature is innervated by visceral motor neurons that travel in specific cranial nerves and in the autonomic parts of spinal nerves.

The somatic musculature is often subdivided into axial musculature, along the axis of the body, and appendicular musculature, located more peripherally in the appendages.

Axial musculature

The most anterior axial musculature, derived primitively from the first three body segments, or somites, becomes associated with the developing eye and forms the extrinsic ocular muscles. Myotomes in the occipital region of the head and in the embryo's neck cannot extend directly ventrally because of the presence of gill pouches. They grow posteriorly and ventrally behind the gill pouches and then extend forward in the floor of the gill region to the lower jaw. In this way the ventral portions of these myotomes give rise to the hypobranchial musculature of the tongue and ventral surface of the neck. The rest of the axial musculature can be thought of as trunk musculature; most lies in the trunk, although some extends onto the dorsal surface of the neck and into the tail. Except in the most primitive fishes, the trunk myotomes separate during development into dorsal parts that give rise to the epaxial (above the axis) muscles of the back and underparts that give rise to the hypaxial (beneath the axis) muscles on the sides and ventral surface of the trunk.

Appendicular musculature

Appendicular muscles are defined as those somatic muscles that develop from mesenchyme within the embryonic limb buds. In fishes this premuscular mesenchyme separates from the myotomes very early during development, but in higher vertebrates some of it migrates into the limb buds from the somatic layer of the lateral plate mesoderm. As the appendicular skeleton develops, the premuscular tissue in a limb bud becomes separated into dorsal and ventral masses from which the individual muscles later differentiate. Consequently, dorsal (abductor and extensor) and ventral (adductor and flexor) limb muscles can be recognized. Most of them extend from the girdle into the limb or lie within the limb; but some grow from the limb base medially to the trunk—*e.g.*, latissimus dorsi. Such muscles should not be confused with muscles that start their development elsewhere and secondarily extend to the girdles; *e.g.*, certain trunk muscles and visceral muscles.

Most of the visceral musculature forms part of the wall of the digestive tract, blood vessels, and other hollow organs. Visceral musculature in the anterior part of the gut tube, however, becomes associated with the gill, or visceral, arches and constitutes a special subgroup known as branchiomeric musculature.

Branchiomeric musculature

Branchiomeric musculature becomes striated, but most of the remaining visceral musculature is smooth or, in the heart, cardiac. Only the branchiomeric musculature forms discrete enough entities to be of interest to gross and comparative anatomists. These muscles are grouped according to the visceral arch with which they are associated: mandibular muscles with the first arch, hyoid muscles with the second arch, and branchial muscles with the remaining arches.

Ancestral vertebrates evolved morphogenetic processes that led to the development of an adult musculature adapted to the animal's aquatic mode of life. The muscles that develop from the primordia referred to above are not complex, and most stay close to their sites of origin. As trunk muscles differentiate from the myotomes they remain segmented, for a segmented musculature is very effective in inducing the lateral body undulations used in swimming; the dorsal and ventral premuscular masses often differentiate into a single dorsal and ventral fin muscle, for fin movements are not complex; hyoid muscles remain attached to the hyoid arch, and so forth.

As vertebrates adapted to terrestrial life, fundamental changes occurred in methods of support, locomotion, feeding, and breathing. These changes were reflected in alterations in the musculature, but morphogenetic processes leading to these alterations evolved as modifications primarily in the later stages of an inherited pattern of development. The early embryonic muscle primordia of mammals, therefore, are very similar to those of fish; and myotomes, for example, are still segmented even though there will be little segmentation in the adult musculature. The alterations that occur in later development result chiefly from three processes. (1) Parts of adjacent muscle primordia may fuse to form continuous muscles extending over many body segments. The lateroventral parts of the myotomes in the abdominal region unite to form continuous sheets of muscle such as the external oblique. (2) A single muscle primordium may become subdivided. The dorsal and ventral premuscular masses in the appendages give rise to a great many appendicular muscles used for support and locomotion on land. (3) A muscle primordium, or muscles that develop from it, may migrate some distance from the site of origin. In an early mammalian embryo the developing heart lies far anteriorly in the future neck region, and the diaphragm begins to form directly posterior to it. Presmuscular tissue that enters the diaphragm is, therefore, derived from the posterior cervical myotomes. Only later in development do the heart and diaphragm migrate posteriorly.

Muscle homologies

These sorts of changes often make it difficult to recognize the group to which some muscles of adult higher vertebrates belong and to establish their homologies with the muscles of lower vertebrates. Relative position in the body, which is usually a good clue to homology, may be different. Fortunately, nerves enter the developing muscles early and tend to accompany the muscles during their later developmental movements. Innervation is conservative and is a good guide to muscle identity, but changes in innervation sometimes occur. Recourse must then be made to developmental studies to establish muscle homology.

EVOLUTION OF THE VERTEBRATE MUSCULAR SYSTEM

The major changes that have occurred in the muscles of vertebrates during their evolution from fishes to mammals can be most easily followed by discussing them according to the groups to which they belong.

Ocular muscles. All vertebrates with well-developed eyes have a group of strap-shaped extrinsic ocular muscles that insert onto the eyeball and move it. The basic pattern of the group is established in fishes, in which four rectus muscles originate from the posterior corner of the orbit and two oblique muscles from the anterior corner. All develop from the first three somites and hence are supplied by the cranial nerves of these somites.

Eye movements are similar in all vertebrates, and few evolutionary changes have occurred. Terrestrial vertebrates retain the six muscles found in fishes. Many species, especially those with relatively large eyeballs, have a retractor bulbi that ensheathes the optic nerve and helps to hold the eyeball in place. Tetrapods (the land vertebrates) have moveable eyelids that protect the eye. A levator palpebrae superioris raises the upper eyelid in mammals; and an orbicularis oculi, derived from the hyoid group of branchiomeric muscles, closes the eyelid.

Hypobranchial muscles. Fishes have a simple hypobranchial musculature consisting of several longitudinal muscles that extend from the pectoral girdle forward in the floor of the gill region to the bases of the branchial arches, hyoid arch, and lower jaw (Figure 8A). The coraco-mandibular and rectus cervicis, which attach, respectively, on the lower jaw and hyoid arch, open the jaw; the deeper coracobranchials attach onto the gill arches and help to expand the pharynx and gill pouches to suck in water for breathing. These muscles develop embryonically from the ventral parts of the occipital and anterior trunk myotomes.

Gills are lost in terrestrial vertebrates and with them the respiratory functions of these muscles. Manipulating food

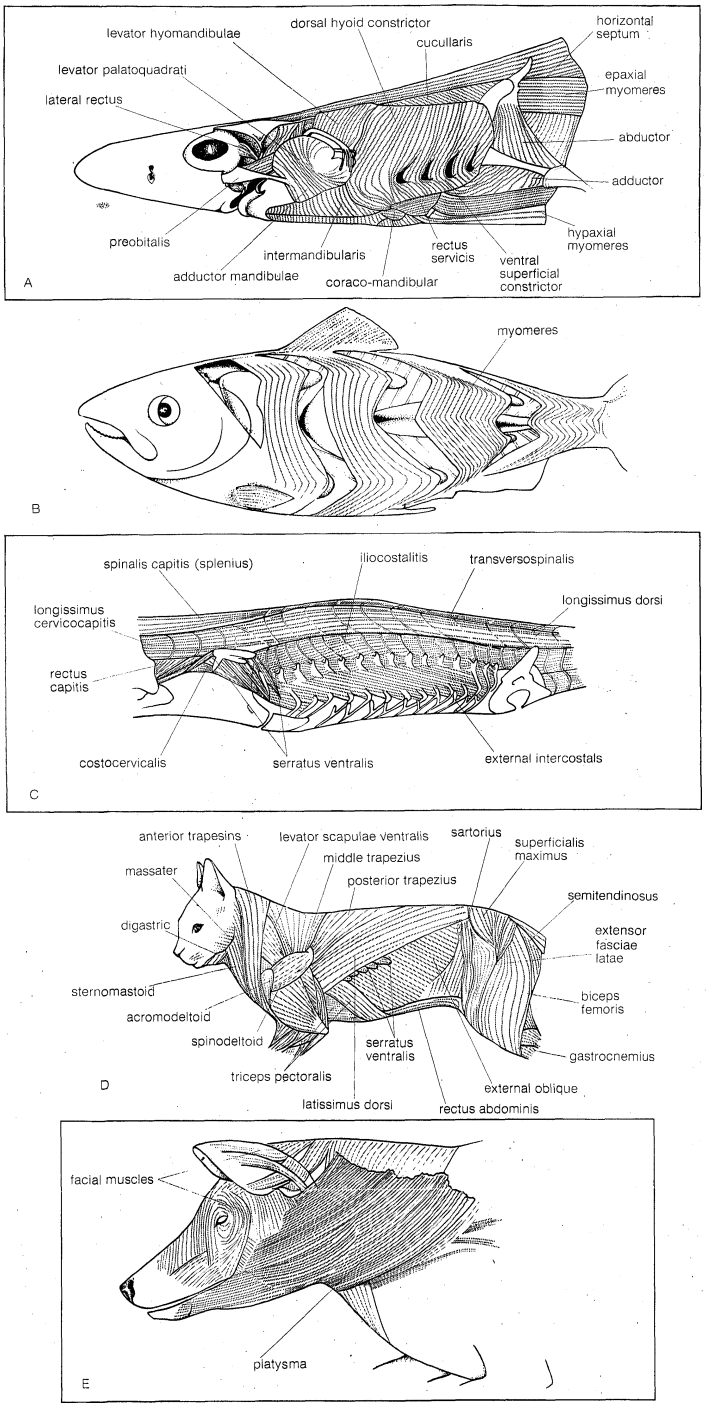


Figure 8: *Vertebrate trunk musculature.*
(A) Lateroventral view of anterior muscles of dogfish. (B) Lateral view of trunk muscles of salmon. (C) Axial muscles in lizard. (D) Lateral view of trunk muscles of a cat. (E) Lateral view of facial muscles of a dog.

From (A) W. Walker, *Vertebrate Dissection*, 4th ed. (1970); W.B. Saunders & Co., and (B,C,E) A. Romer, *The Vertebrate Body* (1964); W.B. Saunders & Co., and (D) *Chordate Anatomy* by H.V. Neal and H.W. Rank. Copyright © 1944 by McGraw-Hill Book Co. Used with permission of McGraw-Hill Book Company

in the mouth and swallowing it, however, becomes more difficult because there is not a steady flow of water through the mouth and pharynx. Terrestrial vertebrates evolved a muscular tongue that plays an important role in food manipulation and swallowing. Mammals have a lingualis muscle within the tongue and others that extend into the tongue from the chin (genioglossus), hyoid (hyoglossus), and skull base (styloglossus). All have evolved from the anterior part of the hypobranchial musculature. Swallowing also entails a forward and backward movement of the hyoid bone and larynx, which is accomplished by longitudinal muscles derived from both the anterior and posterior parts of the hypobranchial musculature. Their names indicate their attachments: geniohyoid (chin hyoid); sternohyoid (sternum hyoid); omohyoid (pectoral girdle hyoid); sternothyroid (sternum thyroid [the thyroid is a laryngeal cartilage]); and thyrohyoid (thyroid hyoid).

Trunk muscles. Trunk and tail muscles constitute the bulk of the musculature of a fish. They consist of a series of segmented myomeres derived from the embryonic myotomes. Except in cyclostomes, such as the hagfishes and lampreys, the myomeres are divided by a longitudinal and horizontal sheet of connective tissue into dorsal epaxial and ventral hypaxial portions (Figure 8A). The epaxial segments extend from the back of the skull nearly to the end of the tail; the hypaxial ones do not extend farther anteriorly than the pectoral girdle. Waves of contraction, which become more pronounced posteriorly, pass alternately down the myomeres of opposite sides of the body so that the posterior part of the trunk and tail undulate from side to side and propel the animal through the water.

Locomotory musculature

The myomeres are folded in a complex zigzag fashion. The apex of each "zig" and "zag" is roughly cone shaped so that it extends farther forward or backward than appears on the body surface (Figure 8B). Although most muscle fibres in the myomeres extend longitudinally, some are diagonal. These complexities appear to enable short fibres to exert their effects over several body segments and also cause the trunk and tail to twist during their undulations in such a way that they exert a thrust downward as well as laterally. The effect is somewhat like that of a sculling oar.

Terrestrial vertebrates, since they are not buoyed up by the water, must have a strong vertebral beam braced by strong back muscles and strong appendages that propel them by thrusting against the ground (Figure 8C). Their trunk muscles play a role different from those of fishes. The powerful epaxial musculature, which braces the back, can be organized into three columns, as seen in the mammals: (1) A median transversospinalis group lies next to the spinous processes of the vertebrae. The deeper parts of the muscle group remain segmented and include short muscles, such as the interspinales and intertransversarii that interlace successive vertebrae. The superficial parts fuse to form more extensive longitudinal muscles, the multifidus and spinalis that act across many body segments. (2) A longissimus dorsi lies lateral to the transversospinalis group, and (3) a long iliocostalis lies lateral to it. The anterior parts of these columns break up into smaller muscles that support and help to move the head and neck. The very long muscles, spinalis, longissimus, and iliocostalis, are often closely associated to form a functional complex called the erector spinae.

Hypaxial musculature in primitive terrestrial vertebrates

The hypaxial musculature remains extensive and segmented in primitive terrestrial vertebrates, such as salamanders (body squirming still plays some role in locomotion); but it becomes much reduced and loses most of its segmentation in higher species. A column of hypaxial musculature lies ventral and lateral to the vertebral column and helps to support and move the back.

Most of the rest of the hypaxial musculature helps to form the thoracic and abdominal walls, and most of this has an important role in breathing in higher vertebrates that ventilate their lungs by changing intrathoracic pressure. The muscular diaphragm of mammals is considered to be hypaxial because it develops from the ventral parts of cervical myotomes. It forms a dome-shaped partition between the thoracic and abdominal cavities, which moves posteriorly when the muscle contracts and reduces intrathoracic pressure. The elastic recoil of the abdominal viscera, assisted by the contraction of muscles in the abdominal wall, restores the diaphragm to its resting position. The abdominal wall contains a longitudinal rectus abdominis muscle on either side of the central line and, laterally, three broad, thin sheets of muscle: external oblique, internal oblique, and transversus abdominis. Comparable layers occur in the thoracic wall: external and internal intercostals lie between the ribs; a reduced transversus thoracis lies deep to the sternum, and a reduced rectus thoracis occurs near the front of the sternum. In addition, a serratus dorsalis inserts near the verte-

bral attachments of the ribs, and several scalenus muscles extend from the neck to the ventral portions of the anterior ribs. All these muscles assist in changing thoracic size and intrathoracic pressure by moving the ribs forward during inspiration and backward during expiration.

A final group of hypaxial muscles is associated with the pectoral girdle in tetrapods. Posteriorly, there is an articulation between the pelvic girdle and sacral vertebrae, but none exists anteriorly. Body weight is transferred from the trunk to the pectoral girdle and appendages by a muscular sling. A major component of this complex is the serratus ventralis (Figure 8D), which extends from the vertebral border of the scapula to the ventral ends of the ribs. Other hypaxial muscles, belonging to the rhomboideus group, extend between the scapula and spinous processes of the vertebrae. A levator scapulae ventralis extends between the skull and scapula.

Appendicular muscles. The paired fins of a typical fish do not have a role in body support; rather, they are used as stabilizing keels and rudders for turning. Appendicular musculature is correspondingly simple and relatively weak. Many fishes have only a dorsal abductor or extensor muscle that pulls the fin dorsally, and a ventral adductor or flexor that pulls it ventrally. Certain fibres of these muscles spread far enough fore and aft so that they can also pull the fin forward or backward slightly. The appendicular musculature is derived primitively from the mesenchyme that originates in the ventral portion of the myotomes.

The tetrapod limb

The limbs of terrestrial vertebrates become the primary supporting and locomotor organs, and the appendicular muscles become correspondingly complex. An early stage in the evolution of tetrapod limb muscles is seen in salamanders. The proximal segment of each limb lies and moves primarily in the horizontal plane, and the distal segment extends perpendicularly from the elbow or knee to the ground. There are many appendicular muscles, but they can be grouped into a dorsal and a ventral series that have evolved, respectively, from the abductor and adductor series of fishes. Dorsal muscles abduct, or lift away, the proximal segment of the limb and extend the lower segment and foot, thereby raising the limb from the ground during the recovery phase of a stride. Ventral muscles are more active during the propulsive phase; for they adduct, or pull in, the limb and flex its distal segments, thereby placing the foot firmly on the ground. Both dorsal and ventral muscles are involved in moving the limb forward (protraction) and backward (retraction). Although limb muscles are numerous in salamanders, they are not very powerful, and lateral undulations of the trunk help in limb placement and movement. When the anterior portion of the trunk bends to the left, for example, the right pectoral appendage is protracted to some extent because it lies on the convex side of the trunk fold. As a wave of contraction spreads down the right side of the body, the right appendage, now on the concave side of a fold, exerts a posterior thrust.

By the mammalian stage of evolution, the appendages have rotated so that they extend ventrally perpendicular to the trunk. This arrangement provides better support and makes for a long fore and aft swing of the legs. Lateral trunk undulations are no longer important in locomotion, although there is some extension and flexion of the back in certain fast-running mammals such as carnivores. Mammalian limb muscles as a group are very powerful, and although there has been a further subdivision and increase in their number, dorsal and ventral groups can still be recognized. Since the body is supported by vertical limbs, the ventral group is relatively less powerful than in primitive tetrapods that need to raise the body from the ground. Some ventral muscles shift during embryonic development from a ventral to a dorsal position, where they can participate in the swing of the legs. The supraspinatus and infraspinatus, which lie on the lateral surface of the scapular bone, are examples.

Dermal muscles

Fibres from some of the appendicular muscles fan out beneath the skin and form integumentary or dermal muscles that can pull the skin. Largest and most common of these is the panniculus carnosus, which originates from the pectoralis and latissimus dorsi in the shoulder region and spreads out beneath the skin of the trunk. Integumentary muscles beneath the skin of the neck and head are branchiomeric in origin.

Branchiomeric muscles. The bronchiomeric group of visceral muscles is very well developed in fishes, in which the visceral arches are prominent structures. All of the arches support gills in primitive jawless fishes, and the posterior ones still do in jawed fishes. The musculature of one of the posterior arches in a shark probably comes close to the primitive condition of the musculature for all of them at an early stage in their evolution. Most of the musculature extends as a sheet laterally and posteriorly from a gill arch and spreads over the top and bottom of a gill pouch, forming the dorsal and ventral superficial constrictors (Figure 8A). Deeper parts of the musculature extend between segments of an arch. All these muscles compress the gill pouches and drive water out of the external gill slits during expiration. Primitively, there probably was a dorsal levator for each arch that helped to expand the gill pouches during inspiration; but the levators of the last four or five arches have aggregated to form a triangular-shaped cucullaris, most of whose fibres act on the pectoral girdle. The gill pouches are expanded by the elastic recoil of the gill arches aided by the coracobranchial muscles of the hypobranchial group.

The visceral arches in sharks

The first visceral arch of sharks forms the jaws, and hence it is referred to as the mandibular arch. The second, or hyoid arch, helps to suspend the jaws; its upper segment, known as the hyomandibular cartilage, extends as a prop from the ear region of the skull to the posterior end of the upper jaw, or palatoquadrate cartilage. The ventral part of the hyoid arch extends into the floor of the anterior pharynx and mouth cavity. Muscles of these arches are naturally somewhat modified. The levator of the hyoid arch supports the hyomandibular, and most of the rest of the hyoid musculature forms constrictors that compress the gill pouch lying posterior to the hyoid arch. The levator muscle of the mandibular arch supports the palatoquadrate, and most of the rest of the mandibular musculature forms a large adductor mandibulae that closes the jaw. The most ventral part of this musculature forms the intermandibularis lying between the two halves of the lower jaw. Certain of its fibres incline posteriorly enough to help certain hypobranchial muscles open the jaw. With few exceptions, jaw-opening muscles are always relatively weak in comparison with jaw-closing muscles. Muscles of the various visceral arches are innervated by specific cranial nerves, and the pattern established in fishes is maintained throughout the vertebrate series.

During the evolution of terrestrial vertebrates, branchial respiration is replaced by pulmonary respiration; the jaws are formed of more superficial dermal bones that develop around the mandibular arch, the upper jaw is self-supporting and firmly united with the skull proper, and some changes occur in methods of feeding. Profound changes take place in the visceral arches and their muscles. Most of the mandibular musculature continues to act on the jaws, and the bulk of it forms the jaw-closing muscles, which in mammals are the deeply situated pterygoid muscles and superficial temporalis and masseter muscles (Figure 8D). The jaws are opened primarily by the digastric muscle, which extends from the lower jaw to the hyoid bone and thence to the base of the skull. The front half of it has evolved from mandibular muscles; the back half from hyoid muscles. Each part retains its characteristic innervation. A mylohyoid compresses the floor of the mouth, a tensor veli palatini lies in the soft palate, and a tensor tympani attaches to the malleus bone in the middle ear and regulates tension of the eardrum. It may seem odd that a mandibular muscle should attach to an auditory bone, but the malleus is a derivative of the mandibular arch.

Fate of the jaw muscles in mammals

The hyoid bone of mammals is moved primarily by the hypobranchial muscles. Only two small hyoid muscles remain associated with this arch; a stylohyoid extends to the hyoid bone from the skull base, and a stapedius muscle attaches onto the stapes bone in the middle ear cavity.

Most of the hyoid musculature has shifted superficially and spread out beneath the skin of the neck and face as the platysma and facial muscles of expression (Figure 8E).

Muscles of the third arch contribute to the musculature of the pharynx wall. The last arch is lost in mammals, but arches four to six form the laryngeal cartilages. Intrinsic laryngeal muscles, including those of the vocal cords, have evolved from the muscles of these arches. The primitive cucullaris expands and forms a group of muscles that act on the pectoral girdle; *i.e.*, the trapezius and sternocleidomastoid complexes.

FUNCTION AND REGULATION OF MUSCLE ACTION

Postural muscles, fast and slow. Less variation is found in the physiology and regulation of vertebrate muscle than in that of invertebrates, yet several distinct conditions are found. Most vertebrate postural muscles, which are the striated somatic muscles of the skeleton, are composed of muscle fibres belonging to the fast system described earlier. Each fibre receives one neuron terminal, and the fibre is activated only when the junction potential reaches a threshold that sets off an all-or-none action potential. Although any given fibre has a single neuron terminal, a single neuron branches and supplies a number of fibres. Groups of fibres activated simultaneously by one neuron constitute a motor unit, and there are many such units in a single muscle.

Coarse and fine control of muscle

Additional control of tension development derives from the number of motor units that are involved. Only a few motor units are contracting during tonic activity, but as more force is needed during phasic activity, more motor units are recruited. Muscles in which only coarse control over tension development is needed have few motor units, each is made up of many fibres, often 100 or more. Many leg muscles belong in this category. If finer control is needed, as in the extrinsic ocular muscles, the muscle contains many motor units; each is made up of only a few fibres, sometimes no more than ten.

All vertebrates have muscles containing fast-acting fibres in which a maximum tension can be quickly developed. These fibres are supplied by relatively large neurons having a diameter of 12 μm (micrometres). Some muscles of amphibians are known to contain fibres supplied by smaller neurons only 5 μm (micrometres) in diameter; these fibres belong to the slow system. They are supplied by many nerve terminals, and graded junction potentials can directly elicit a gradual contraction. Although some muscles are composed almost entirely of slow fibres, slow and fast fibres are intermingled in others; *e.g.*, the iliofibularis in the leg of the frog. Stimulation of both the large and small neurons can cause fast, all-or-none-type twitches to be superimposed upon the graded contractions of the slow system.

Visceral smooth muscle. When smooth muscle is active, there is a fall in its membrane potential and an increase in the frequency of small, spinelike action potentials. This is accompanied by an increase in tension development. Both electrical and mechanical events are graded. The lower the membrane potential, the greater is the frequency of action potentials and the greater is the tension.

Spontaneous muscle action

Unlike postural muscles, smooth muscle displays spontaneous and rhythmic activity that is independent of nerve action. This activity can be observed in nerve-free tissues, such as the amnion of a chick embryo or when nerve action is blocked in other tissues. Excitability spreads from fibre to fibre, in part, by the pull exerted by contracting fibres on their neighbours and, in part, by current flow from the membrane of an active cell to those of other cells through areas of close membrane contact where resistance to flow is low. Most conduction is along the long axes of the fibres because the fibres have an extensive and intimate overlap in this direction. Conduction is slow, often on the order of five centimetres per second.

Smooth muscle in different vertebrates, or in different organs of the same animal, varies greatly in its response to stretch, nerve stimulation, chemical mediators, and hormones. Parasympathetic (vagus) stimulation in mammals for example, is excitatory to the stomach and sympathetic (splanchnic) stimulation is inhibitory, but both are excitatory to the stomach of elasmobranch fishes, such as sharks, rays, and chimeras. In most mammals adrenalin, which is similar to the neurotransmitter of the sympathetic nervous system, inhibits and relaxes the entire intestinal region, but in adult guinea pigs, it excites and contracts the lower ileum (small intestine). Adrenalin relaxes the uterus of a rat; it also relaxes the uterus of a nonpregnant cat, but contracts that of a pregnant one. The hormone oxytocin specifically contracts smooth muscle.

Multi-unit nonstriated muscle. Muscles that erect the hairs, adjust the cornea and the iris of the eye, move the eyelid, and control the size of many arteries are morphologically smooth. They do not, however, show any of the intrinsic activity characteristic of visceral smooth muscle, and they are capable of rapid and direct nervous stimulation. Some can also be activated by hormones. They are often referred to as multi-unit nonstriated muscles.

ELECTRIC ORGANS

Electric organs are an unusual modification of muscle tissue that has evolved independently in several groups of fishes. The muscles involved may be extrinsic ocular muscles as in the marine stargazer (*Astroscopus*); branchiomeric muscles as in the marine electric rays (*e.g.*, *Torpedo*); or trunk muscles as in the electric eel of South America (*Electrophorus*) and the mormyrid knife fishes of Africa (*e.g.*, *Gymnarchus*). The electric catfish of Africa (*Malapterurus*) has an extensive electric organ beneath the trunk skin, but its derivation from muscle is not certain.

In the electric organ in *Torpedo*, each muscle fibre, or its motor end plate, is modified as a disk-shaped structure called an electroplax. During development the incipient muscle cell, or myoblast, flattens and loses its contractile apparatus. The surface receiving the nerve supply usually remains rather flat, but the opposite surface usually develops many protruding papillae. The motor neurons branch extensively and have many points of contact with the electroplaxes (see BIOELECTRICITY).

BIBLIOGRAPHY. W. ANDREW, *Textbook of Comparative Histology* (1958), college-level textbook containing information on invertebrate muscles; J.R. BENDALL, *Muscles, Molecules and Movement* (1969), a review for the specialist on activity of vertebrate striated muscle, with emphasis on physiology and biochemistry; W. BLOOM and D.W. FAWCETT, *Textbook of Histology*, 9th ed. (1968), a college and medical school textbook with a chapter on muscular tissue; G.H. BOURNE (ed.), *Structure and Function of Muscle*, 3 vol. (1960), an advanced treatise; J. BRACHET and A.E. MIRSKY (eds.), *The Cell*, 6 vol. (1959–64), a treatise for the expert with information on muscles included in vol. 4; E. FLOREY, *An Introduction to General and Comparative Animal Physiology* (1966), with chapters on the structure and function of muscle and on the external control of muscular activity; G. HOYLE, *Comparative Physiology of the Nervous Control of Muscular Contraction* (1957), a review of the literature for the general biological reader; P.A. MEGLITSCH, *Invertebrate Zoology* (1967), a college-level survey of the invertebrate phyla, with more information on the structure of invertebrate muscular systems than most comparable books; O.E. NELSEN, *Comparative Embryology of the Vertebrates* (1953), college textbook with a chapter on the muscular system; C.F.A. PANTIN, "A Problem in Muscular Excitation in the Anthozoa," in J.W.S. PRINGLE (ed.), *Essays on Physiological Evolution* (1965), a discussion of the problems of muscle control in organisms having a very primitive nervous system; C.L. PROSSER and F.A. BROWN, *Comparative Animal Physiology*, 2nd ed. (1961), college textbook with a chapter on the physiology of invertebrate and vertebrate muscle; J.A. RAMSAY, *Physiological Approach to the Lower Animals*, 2nd ed. (1968), an excellent general account of comparative physiology with a chapter on muscles; A.S. ROMER, *The Vertebrate Body*, 4th ed. (1970), a textbook of comparative anatomy that describes the embryonic development and evolution of the vertebrate muscular system; W.F. WALKER, *Vertebrate Dissection*, 4th ed. (1970), a laboratory manual for comparative anatomy that gives a thorough description of the muscular systems of the dogfish, mudpuppy, cat, and rabbit; J.Z. YOUNG, *Life of Mammals* (1957), a college-level textbook carefully integrating the structure and physiology of mammalian organ systems.

(W.F.W.)

Museum

As institutions developed by modern society to stave off for as long as possible the deterioration and loss of objects treasured for their cultural value, museums do for such objects what libraries do for books and what archives do for official documents. Society does not save these things just to hoard them but to use them. Museums accordingly are organized to use their treasures for cultural purposes.

There are many kinds of museums. They all collect, preserve, study, and exhibit objects, but, within the common framework of these basic functions, they differ widely in what they do and why they do it. Though the museum as it is known in the 20th century is of relatively modern conception, its name was used for a different kind of institution, though of not dissimilar character, in ancient classical times: the word museum, derived, through Latin, from the Greek *mouseion*, "seat of the muses," was defined by Guillaume Budé in his *Lexicon-Graeco-Latinum* (1554) as "a place dedicated to the muses and to study, where one engages oneself in noble disciplines." Thus, the museum at Alexandria, Egypt, created at the beginning of the 3rd century BC in the Ptolemaic era and later destroyed, was not a museum in the modern sense but an institution of research, a library, and an academy. (Nearer to modern notions of museum is the Greek *pinakothēkē* [French *pinacothèque*, German *Pinakothek*], which Budé defined as the place where painted panels, objects of jewelry, standards, and other ornaments are kept.)

Notion of a museum

The notion of a museum springs from the passion for collecting, which is deeply rooted in human nature. All civilizations, from the most primitive to the most advanced, share the desire to accumulate objects that are beautiful, costly, rare, or merely curious. Religious communities, rulers, and magnates set up the first collections, open to the public or at least to certain members of it; examples are seen in the treasury of the Athenians at Delphi and later in the cathedral treasuries of the Middle Ages, as well as in the acquisitions of the Roman emperor Hadrian and, before him, of the Roman provincial governor Gaius Verres; similar examples are also found in the primitive civilizations of Africa and Oceania, at a stage of development comparable to that of the Neolithic Period or the European Bronze Age. From the basic satisfaction of the collector's instinct, one proceeds in time to the making use of assemblages of objects to suit the needs of study. Since society became organized, museums have opened their doors ever more widely to those sections of the population that have had access to knowledge through education. More recently, with the spread of university education, museums have begun to assimilate and make use of techniques of mass communication.

Today the museum is an institution that assembles, studies, and conserves objects representative of nature and man in order to set them before the public for the sake of information, education, and enjoyment. By this definition, the term museum includes not only those institutions known as such but also art galleries (noncommercial), picture galleries, secular and ecclesiastical treasuries, certain historical monuments, permanent open-air exhibitions, botanical and zoological gardens, aquariums, and libraries and archives in so far as they are open to the public.

This article is divided into the following sections:

I. Museum operations
  Functions of museums
    Social role
    Service to the community
    Educational and cultural roles
  Museum procedure
    Legal and financial arrangements
    Personnel
    Collection
    Exhibition
    Other operations
    Conflicts of interest in museum operations
II. Types of museums
  Art museums
  Historical museums
  Science museums
  Specialized museums
    Open-air museums
    Regional museums
    Other institutions functioning as museums
III. Museum architecture
  Temple or palace museums
  Monuments transformed into museums
  Recently constructed museums
  Theory of museum architecture
  Aesthetic and functional architectures
IV. History of museums
  Through the 17th century
  The 18th century
  1790–1850
  1850–1950
  1950–70
  Prospective

I. Museum operations

FUNCTIONS OF MUSEUMS

The museum as a reflection of cultural and social environment

Social role. A museum is above all a reflection of man and his activity, of his natural, cultural, and social environment. It speaks a language that is specific, that of the object, the "real thing." In this way it exerts its influence on the illiterate visitor as well as on the visitor steeped in visual impressions from two-dimensional objects (pictorial images, reproductions, texts). Because of this, the museum has always played a social role; reflecting the might of church or state, offering a feast of research to the university and to the lone scholar, mirroring the varied interests of a local learned society, symbolizing the wealth and social status of a collector, and standing as a cultural monument to the glory of a city. In the contemporary world, a museum tends to be both an expression of the community and an instrument to serve it. Nonetheless, it often happens, particularly in the countries of Europe where it was born in the 18th and 19th centuries, that the museum remains a treasure-house frequented more by foreign tourists than by its immediate community. This mummified kind of museum, however, must disappear in the face of the needs of education and cultural activity in contemporary society. Indeed, through its collections, a museum above all presents the past, in everything that concerns natural history, art, archaeology, ethnology, or anthropology. That is to say, it presents the values of a vanished or vanishing society or, even further, focusses attention on permanent natural values endangered by modern progress. It marks out these values for the public as a "heritage" because of their importance for the continuity of culture. The values exemplified are moral, religious, aesthetic, historical, and biological, and interested visitors and research workers alike are reminded of forgotten ideas and are made aware of a starting point for an enhanced life. A window opens on the world, whether the infinitely small one of the electron microscope or the infinitely large one of the astronomical telescope, whether that of a neighbouring people or that of the inhabitants of the most faraway countries.

Preservation of cultural values. Through the values that it embodies and expresses in visual terms, a museum conveys an objective view of the three-dimensional world to its visitors, allowing them to exercise their critical spirit over their own heritage and that of other peoples, shaping their taste and arousing their curiosity. In the aristocratic world extending from Renaissance times to the end of the 18th century, collections of curiosities made by universities and wealthy individuals were the source of much scientific development. In the 19th century, young painters quite often received their training in the galleries of paintings of royal palaces. Twentieth-century civilization, insofar as it turns to the past for cultural values, is perhaps more than ever based upon the museum.

Role as an instigator of new cultural trends

The museum has an even more important role: it can stimulate artistic ability and intellectual purpose, awakening the creative spirit latent in each of us. As examples of spectacular success in this respect may be cited the Polytechnic Museum of Moscow, the Birla Industrial and Technological Museum, Calcutta, the East Harlem Art Studio (Neighbourhood Museum), the A.R.C. section

(Animation, Research, Confrontation) of the Musée d'Art Moderne de la Ville de Paris. Research undertaken by sociologists at Toronto, Tel Aviv, and Munich confirms this as does the success achieved by children's museums. A guarantor of the continuity of culture, the museum thus plays its part in the permanent creation of new cultural habits.

Service to the community. The museum serves primarily its community. The latter may be the nation as a whole (as in the case of the national central museums—British Museum, Louvre, Museum of History and Technology, Washington, D.C.) or be provincial or local. A community is perhaps also metropolitan; major cities have museums, which, like the Metropolitan Museum of Art of New York, have a principal responsibility to serve the population of the huge area surrounding them. The first duty of a museum is, therefore, to define its community, to understand its needs, as a whole and in its diverse parts —social classes, minorities, and so on. On this basis a museum's program, including the size and scope of its collections, the spread of its activities, and the material means to give them effect, will be established. Furthermore, the program must take into account the existence of the community's other cultural centres, the needs of teaching syllabi, from primary school to university, and lastly the reaction of the community itself, studied and evaluated at each level of museum activity.

Exhibition. Naturally, a museum's principal activity is the use of the "language" peculiar to itself—that is, of presentation—to display the objects belonging to its collections or occasionally lent to it for temporary exhibition. The aim of presentation is to provoke and facilitate direct contact between the individual and the object, whether the individual be a child member of a school group or an adult alone, whether the object be a work of art, a specimen belonging to natural science, or a working model in a gallery of technology. The science of the museologist and the art of the museographer (museology meaning the philosophy and theory of museums, while museography is the description of practical museum work) will convey the message simultaneously to diverse categories of persons. The object will be presented in the light of scientific, aesthetic, and psychological criteria and will be displayed in association with other objects in an explanatory and documentary display incorporating a coherently related group of things and ideas. This group of objects will be set in the larger context of the natural or cultural heritage. In the case of a work of art, the message of which is above all aesthetic and therefore individual, contemplation and assimilation will be made easier by a layout whereby a significant number of persons may without mutual hindrance enter into intimate contact with the work and benefit from what have been called the "tactile values" by the art historian Bernard Berenson. Recollection of the content will thus be more pliant and more allusive. The diversity of the public obviously gives rise to the greatest difficulty in the choice of display. After the static and quantitatively exhaustive displays favoured in the 19th century and in some "elite" museums of the 20th (crowded objects and fading labels unchanged for years in their showcases), the time has at last arrived for a division into permanent displays in great public galleries that are more or less educational, study galleries for the well-informed art lovers, and reserved material accessible to qualified researchers only. At the same time, a formula of flexible architecture permits frequent changes of display and the multiplication and rapid succession of temporary exhibitions, thus making it possible to serve all sections of the community simultaneously or successively, as is done, for example, in the Museum of Modern Art, New York, the Victoria and Albert Museum, London, the Musée National des Arts et Traditions Populaires, Paris, and the National Museum of Victoria, Melbourne.

Museology and museography

Collection. The problem of collection has likewise changed, since a museum now serves the community and no longer the cultivated social group that brought it into being. Most museums, with the exception of recently created institutions and university museums, have acquired their collections a little haphazardly, depending upon the taste of the donors and the particular scientific interest of the staff. For some years, it has been stated that a museum should have an acquisition policy defined by two criteria: the state of scientific knowledge and the needs of the community—thus a work of art should be acquired not for its rarity or price but to fill a vacuum in the existing collection's representativeness. For the same reason, a field expedition or an expedition for purchase will be a profitable enterprise for a museum of archaeology or ethnology. Museums often create new departments in order to respond better to public demand; for example, the Department of Architecture of the Museum of Modern Art, New York, the Department of Biology of the Moravian Museum at Brno, Czechoslovakia, and departments of city planning in a number of city historical museums.

Acquisition policies

Educational and cultural roles. Even if a museum has its specific form of address, or "language," and its particular effect on the public, it is not the only institution of education and culture at the heart of the community; it should cooperate closely with others, especially to become some sort of teaching institution itself. The École du Louvre, centre of the teaching of the history of art, set up at the end of the 19th century in the same building as the Louvre, is undoubtedly the foremost example of a popular university attached to a museum. Since then, a number of museums, particularly in the U.S., have initiated series of lectures and have organized a permanent cooperation with schools and universities. From the organization of guided tours, still widespread in France and the U.S.S.R., has arisen the creation of many-sided educational services, which have tended to become departments of education. The advent of audiovisual methods has compelled museums to present, in special projection rooms, programs of films complementing the displays of exhibits. In many central museums and in certain regional museums, teaching chairs at university level have been founded, principally in art history, anthropology, natural sciences, and museology, notably at Paris (Musée de l'Homme), at New York (Metropolitan Museum of Art), at Brno (Moravian Museum), and at Accra, Ghana (the Ghana National Museum). In an increasing number of countries, schoolteaching is closely linked to a museum. In Mexico, a visit to the museums of Mexico City is required annually for all secondary school pupils in classes of history, archaeology, and natural science. In Great Britain and in New Zealand, future teachers or headmasters of primary schools receive a special introduction to museum teaching. East Germany has produced school textbooks that continually refer to museum collections and provide a guide to a systematic tour of them; science museums have been most active in this field. The Museo Nazionale della Scienza e della Tecnica "Leonardo da Vinci" in Milan puts its laboratories at the disposal of senior high school classes.

The extension of direct educational activities of the museum and the growth of its social role to the benefit of the community have been such in the second half of the 20th century that it has appeared necessary to make museums extend outside their walls. In practice, they do not lend themselves to indefinite expansion; too great a flood of visitors prejudices the effectiveness of the instruction provided, by making it difficult to achieve that essential intimate contact between the visitor and the object on display. From this there have developed, during the 1950s, loan services to schools and travelling educational exhibitions to be succeeded by mobile museums (Richmond's Art Mobile, the science exhibitions on "museobuses" organized by the Birla Industrial and Technological Museum, Calcutta, the United Nations Educational, Scientific, and Cultural Organization [UNESCO] pilot museobus in Nigeria). Likewise, since the 19th century, and particularly in France, the great central museums have made use of their important collections in order to effect long-term loans and deposits in provincial museums to enable them to establish collections representative of all periods. More recently, particularly in the United States, opinions have been voiced calling for the decentralization

Mobile museums and museobuses

of traditional museums, established in city centres, for the benefit of suburbs and city neighbourhoods deprived on the cultural plane. The Smithsonian Institution has shown the way by establishing the first neighbourhood museum, at Anacostia, in Washington, D.C. The first colloquium of experts on the subject was held in 1969 in Brooklyn, New York, in another neighbourhood museum (Muse, located in the Bedford-Stuyvesant area), set up by the Brooklyn Children's Museum. In Sweden, in 1965, the government founded Riksutställningar, a pilot scheme for travelling exhibitions, to serve schools, urban and rural cultural centres, and provincial museums by using the collections of museums of the large cities. In countries such as France, where multipurpose cultural centres with up-to-date equipment have been created, museums have undertaken to cooperate with these centres by providing them with all types of exhibition.

Conservation and restoration. In order to be in a position to play its social, educational, and cultural role to full advantage, a museum should not only collect but should also cherish that part of the heritage of which it is in possession and for which it assumes responsibility. That is where the function of conservation, one of the more ancient of a museum, is so important. Museums and picture galleries have always had restorers attached to their staff to maintain their works of art in a good state of preservation, while science museums have at their disposal technicians able to mount specimens carefully. With advancement of scientific knowledge, more and more thorough research has been undertaken in laboratories, dealing with the conditions of conservation required for different kinds of exhibit. There has come to be a distinction between conservation and restoration. Conservation concerns the care given to collections as a whole, to protect them against three causes of alteration: environmental (temperature, humidity, atmospheric pollution, lighting); internal (biological, physical, chemical, or mechanical action within the object itself); and human (handling, transportation, etc.).

Distinction between conservation and restoration

Restoration is the work of repairing and reconditioning deteriorated objects; this requires manual skill and considerable scientific knowledge. The specialized services undertaking conservation and restoration can be classified under three categories: scientific laboratories employing research personnel with elaborate equipment (Institut Royal du Patrimoine Artistique, Brussels; Conservation Center, New York; Centraal Laboratorium voor Onderzoek Van Voorwerpen van Kunst en Wetenschap, Amsterdam; Istituto Centrale del Restauro, Rome); technical laboratories employing personnel of average qualifications under the direction of a scientist, using simpler equipment; and workshops devoted entirely to restoration work, in which the personnel consist of expert workmen. Such services can handle all materials comprising museum collections or deal with one specific material (*e.g.*, iron, wood, or paper). They are national or attached to a particular museum. The first great institutes go back to the 1930s; in the meantime, restoration workshops, which have always existed, have equipped themselves with apparatus for examining early works of art—X-ray, spectroscope, and so on. Training schools for specialists in conservation and restoration have been set up in Rome, Warsaw, New York, Brussels, Mexico, New Delhi, and Tokyo. Furthermore, the problems of conservation are now written into the programs of all courses of museology and museography. An international professional organization, the International Institute for Conservation of Historic and Artistic Works (IIC), with its offices in London, was founded in 1950. In 1958 UNESCO established the International Centre for the Study of the Preservation and the Restoration of Cultural Property, with headquarters in Rome. This organization, which is intergovernmental, was composed of 53 member states in 1971.

Research and publications. Research workers must be counted as a particularly important group of the public served by a museum; for a museum must obviously base its activities on thorough studies of its collections. Research, therefore, forms an integral part of a museum's life, whether it is research undertaken by the museum's own scientific personnel or by outside researchers making use of the museum's documentary resources. These resources ought, therefore, to be developed in a manner to respond to these needs. A first requirement is that inventories and catalogs of the collections should be as complete as possible. Unfortunately, few of them have been able to proceed beyond the level of accession registry or rudimentary card indexing. The reason lies in the shortage of personnel and also in the ceaseless curiosity of conservators, who will often give new acquisitions priority over old ones. The more advanced museums have catalogs or card indexes that include, besides essential identificatory information, complete, scientific documentation of the object. Some have even considered the utilization of computers for the systematic treatment of information concerning collections (for example, the University of Oklahoma's project for cataloging the state's ethnographic collections; Museum Computer Network, consolidating more than 20 museums in the U.S.; a Swedish project in the process of realization around the Livrustkammaren and Skokloster). Few museums, however, are likely to be able to assume the cost of such a system, and its introduction presupposes a good deal of preliminary groundwork yet to be done. Lastly, it appears difficult to envisage computers answering all the questions raised in the domain of the humanities, especially those about aesthetics.

Inventories and catalogs of collections

Besides documenting their collections, museums should afford to their own personnel and to outside researchers complementary facilities such as study rooms and accessible reserves, libraries, archives (particularly for museums of modern art) and occasionally even, as in the case of developing countries, make provision to house foreign visitors. Museums of archaeology or site museums in such countries as India, Pakistan, Peru, and Senegal usually put rest houses at the disposal of foreign researchers. Lastly, reserves and documentation of collections clearly ought to be readily accessible to all researchers, and this involves demands on such things as classification, furniture, and service personnel. Museums should be more and more considered as the normal places for the results of field expeditions—written, graphic, visual, and audiovisual objects and documentation—to be deposited. Thus the museums of natural science of Mexico (Museo de Historia Natural) and of San Francisco (California Academy of Sciences) have received in trust nearly all of the natural science collections from neighbouring universities, so that these shall be taken care of. To conclude the subject, one must mention particular university museums that have a more acknowledged research vocation than others, and the most important of these is without doubt the University Museum of the University of Pennsylvania at Philadelphia; most of the universities of the world, and the greatest of them, such as Oxford, Cambridge, Moscow, Ibadan, Harvard, and Yale, created such museums long ago.

Museum publications do not all fall under the heading of research. Many are of an educational and cultural character or are popular; however, the scientific resources of a museum obviously ought to be made available not only to research workers who are able personally to visit the collections but also to specialists all over the world. That is why scientific publications, periodical or other, are important. If the first are usually devoted to the research work of museum personnel, the others generally comprise exhaustive publication of collections, prepared with care and illustrated. The cost of such productions makes demands on a museum's limited budget, the more so because they cannot be presented as spectacular undertakings to the authorities or those otherwise responsible for finance.

Museum publications

MUSEUM PROCEDURE

Legal and financial arrangements. *Statutes.* Concerning their statutes, museums may be classified in two major groups: public museums and private museums. If they are state museums, the first group generally comes under the ministry of national education or the ministry

of culture, more rarely under the "technical ministries" (*e.g.*, museums of communications under the ministry of communications, army museums under the defense ministry, etc.) or under the ministry of tourism. If they are provincial, they come under district or municipal administration. This is the case of the great majority of museums of Europe and of the developed countries of Asia, Africa, and the Americas. Their personnel have the status of civil servants, and their management is ultimately controlled by the national administration. In certain cases, most notably in Great Britain, a committee appointed over the museum administers its funds and guides its policies. The collections belong to the national patrimony and are normally inalienable. In certain very centralized countries, such as France, even the municipal museums are placed under the technical control of a central administration that selects the scientific personnel and proposes nominations to the municipal administration. Private museums are totally or partly independent of the central administration. They are generally under the authority of a board of trustees or a council that is renewed by co-optation. The director, chosen by the council or the board, is responsible to it for the management of the institution and selects the senior members of the staff with its approval. This is the usual situation in the United States, where only the dependent museums of the Smithsonian Institution and the National Park Service have national statutes. The same is true of Canada, in such cities as Montreal and Toronto. In other countries, only a few isolated private museums come under national regulation, and it is the same with company museums.

Organization of private museums

Foundation. Historically, museums have been founded by individuals, by organizations, and by governments, but generally without special planning. A museum is established to house a collection that has been bequeathed or confiscated, in memory of a celebrated person, to protect archaeological discoveries at a site, or to raise a monument to the glory of a regime or a town. This procedure still applies except for a few great undertakings, such as those of the National Museums of Canada, the National Park Service of the United States, the open-air museums of Romania, and the national and provincial museums of Mexico. This has generally led to undesirable proliferation: in 1964 the rate of creation of new museums in the United States was estimated to be one every 3.3 days. For this reason many poor countries, with the support of UNESCO, now aim at long-term planning, while some European countries are preparing to consolidate their museums according to discipline or to province. It seems desirable not to discourage creative enthusiasm, while maintaining the principle of a minimum of planning, taking account of the needs of communities, and imposing on them a multidisciplinary program, indispensable in the present state of knowledge in order to avoid the old errors that had led to the division (sometimes even administratively) of museums into museums of science, art, history, and technology.

Resources and funding

Finance. Museums' resources can be ranked under four headings. Allocations and subsidies received from public funds constitute the main element of the budget of public museums regulated by statute and often form a not inconsiderable part of that of private museums. The drawback of such resources is a lack of flexibility in the application of their funds and the effect of fluctuation in the political order, through which museums come to receive a low priority in terms of budgetary considerations. A museum's own revenues, which only apply in the case of private or at least largely autonomous museums (notably in the United States), are sometimes considerable, but they seldom suffice to cover the expense of current running and only in exceptional circumstances allow augmentation of activities and acquisitions. Subsidies, gifts, and private legacies are often the sole means whereby a museum can pursue a policy of expansion, acquisition, and development; unfortunately, these resources are somewhat unstable, being often accompanied by conditions incompatible with the museum's freedom of action and susceptible to the effects of economic circumstances. A museum's receipts of a commercial nature are not of considerable importance except in the case of large institutions, and then only when they are statutorily permitted to retain what they receive for their own use, which is not the case with many national museums.

Until latterly, public museums had only public resources at their disposal (gifts could be used only for acquisitions), while the essential element in the budgets of private museums had been set up through private funds, thanks primarily to favourable tax legislation. Meanwhile, it has become apparent that museums, if they are to expand, must be able to benefit at the same time from both public and private funds in order to permit an administrative flexibility that can match the increase in educational activities, research, and conservation.

Entrance fee revenue

There is a tendency to believe that entrance fees to a museum constitute an important source of income. They are in fact negligible for a number of reasons. For one thing, in many countries, the sum of admission charges reverts to the public treasury; for another, a museum worthy of its name should keep admission charges as low as possible if it wishes its collections to be accessible to all and wishes to admit certain classes of visitors free of charge: children, groups, military personnel, students, and friends of the museum. This explains why many museums have given up charging entrance fees, after having established that their receipts hardly covered the salaries of staff employed to sell tickets and control them. They have turned rather to admission fees to temporary exhibitions, at a much higher charge, and to the proceeds of the sale of catalogs to these exhibitions, which are generally used to meet part of the cost of organizing the exhibitions. The British government decided in 1971 to charge admission to the national museums of England and Scotland. This move has been strongly criticized in political and professional circles because public collections are usually donated under the condition that museums and art galleries be accessible to all free of charge.

Personnel. Museum personnel may be classified as scientific, technical, and service. The scientific personnel—director, curators (keepers, assistants), specialists in education, conservation, and restoration—have had university training adapted to the museum's needs and, if possible, a complementary professional training in museology, teaching, conservation, and restoration. According to the international doctrine established at the ICOM seventh general conference, New York, 1965, such personnel should be recognized as having the status and advantages equivalent to those of members of the teaching profession with similar qualifications and responsibilities. This principle is, unfortunately, still far from being universally applied, which explains the difficulties in recruiting encountered by museums. Technical staff—draftsmen, museographers, documentalists, librarians, security specialists, preparers or restorers of the middle range—should possess the professional qualifications of their respective crafts. Museographers should have received, in principle, special training in the techniques of presentation, cataloging, and so forth. Service personnel include attendants, maintenance workers, and secretaries.

Specialized training of personnel

Specialized training of scientific personnel is a recent development. Since the day of the earliest museums, directors and curators have been recruited from among amateurs or from researchers in the disciplines represented in the museum. In the 19th century, some patrons were even put in charge of directing great institutions. In the course of the 20th century, it has become the practice to call on specialists in the scientific disciplines the museum is concerned with: a museum of art will be headed by an art historian, a department of archaeology by an archaeologist, and so forth. Unfortunately, the unfavourable status of museum staff and, above all, the salaries—inferior to those of universities—sometimes prevent the engagement of persons of the highest qualifications and in any case tempt such a person to leave after a few years to find a better situation elsewhere. On the other hand, this system is not suitable for museums of average importance, which are generally provincial and the collections of which are dependent upon a number of disciplines. That is why, under the impetus first of na-

tional professional organizations, then of international organizations, specialized training has been created; a hundred courses existed in 1970 throughout the world, of which half were in the United States. Many of these courses are of interest only to museums of art and history, few covered the generality of museums, and only some offered an advanced degree or diploma. Among the latter, one may mention the centres or courses at Brno, Leicester, Baroda (India), New York, Boulder (Colorado), and Rio de Janeiro. The École du Louvre at Paris is above all a teaching institution for the history of art. Some national museum associations have their own courses, of which the best known and most outstanding is the museum diploma awarded by the British Museums Association. An international committee of ICOM is endeavouring to establish norms for programs and diplomas. On the average, such training is postgraduate, lasts one or two years, and is directed toward university graduates in various disciplines.

Courses also exist for museum technicians, notably at Jos, Nigeria, and Santiago, Chile. These courses specialize in museographic techniques and are directed toward high school students.

Collection. Three questions can be asked in connection with museum collections; they concern program, acquisition, and management. The program is essential: established scientifically and in accordance with the aims of the museum and the needs of the community, it determines the type of objects that ought to be acquired to serve the objectives of research and exhibition. Acquisition itself depends above all on the available means (special budgets, field expeditions, and gifts and legacies); it can be made in two ways, directly and indirectly. Direct acquisition is essentially the collection of objects from the field by means of excavation and ethnographic or natural-science expeditions. The objects or the specimens recovered are scientifically documented and immediately available for research, education, or exhibition. Indirect acquisitions are brought about with the aid of one or more intermediaries: collectors and art and antique dealers. This is generally the only method for the acquisition of works of art, but its disadvantage is absence of documentation and the risk of illegality. In fact, since World War II, illegal traffic in cultural assets has become of great magnitude—theft, clandestine excavation, illicit export from developing countries, and pillage of poorly guarded monuments. Furthermore, an industry of fakes and copies has developed. Another means of acquisition is exchange with another museum or the deposit on loan of pieces belonging to a collector or scientific institution. UNESCO has successfully adopted a variety of international instruments on these subjects: recommendation on archaeological excavations (1956), recommendation on the illicit export and import of cultural property (1964), and a convention on the same subject (1970). Furthermore, a number of conferences have been held, notably in 1967 at the Metropolitan Museum of Art, New York, on fakes, and in Paris (ICOM) in 1970, on ethical rules of acquisition. No satisfactory solution has yet been found to facilitate the enrichment of the museums of Europe and North America while protecting the cultural patrimony of countries of the developing Third World.

Direct and indirect acquisition

The management of collections begins in the course of field expeditions (summary identification) or with their entrance into the museum. It includes primarily the registration on the museum inventory register. An inventory number will thus identify the object and will be marked on the object itself or attached to it. This number usually follows a code that includes, as a minimum, an indication of the museum or department, the year of accession, and the accession serial number during that year. The next step is the cataloging procedure of the collections, which usually consists of two cards: a museographical card carrying essential information on origin, composition, author, use, title, location in the museum, type of acquisition, price, state of conservation, etc.; the other a scientific card carrying all the information that may be useful to investigators. Depending upon the museum, these cards are written, typed, perforated, or computerized.

Documentation of collections

Museum collections are made up not only of original objects but also of documents referring to them, original audiovisual materials (ethnographic films, recordings, photographs), specially constructed models (above all for scientific and historical museums), and generally all documents referring to the museum's program that are of scientific or didactic interest.

Exhibition. From the church treasures of the Middle Ages to the museums of various categories of the end of the 19th century, exhibition or presentation hardly varied: as much as possible of the collection was shown to the public. The latter, belonging to the cultivated classes, was expected to be capable of finding what it was looking for among a medley of incongruous objects that came to be classified after a system inspired by the science of the epoch. During the first half of the 20th century, the notion of aesthetic presentation prevailed even in museums not dedicated to art. Since the 1950s, a rapid revolution has made itself felt, toward a presentation of items in relationship to one another, rather than in isolation, for everything that is not strictly artistic. Notwithstanding this, all the various schools still coexist: documentary presentation (eastern Europe), aesthetic (Italy and France), ecological (Mexico), and others. At the risk of arbitrary schematization, one can, however, distinguish three types, corresponding to three principal categories of museum. The first, aesthetic presentation, applies to art collections and seeks to isolate a work so that it may be totally perceived; it places the accent on backgrounds, lighting, space, and volume, while the documentary and didactic apparatus is either abandoned or placed as an introduction to the display or on an optional circuit parallel to the display. The complex historical presentation finds application in museums of history, archaeology, anthropology, and ethnology; it seeks to bring to life a cultural ensemble display by a combination of visual and audiovisual means with documentation prominently integrated; it generally brings into conjunction display galleries for a large public and study galleries; it integrates elements belonging to a number of disciplines—geography, economics, sociology, art, religion, and urban-development studies. The ecological presentation is applied to collections of natural sciences and recreates the environment of each species, of the kingdoms of plants, animals, and man. The latter tends to influence the historical presentation, associating man to his environment.

Types of presentation

Exhibition techniques have likewise evolved considerably in the past 100 years. Under the influence of technical progress, the static exhibitions of the past, tied to monumental architecture, have yielded place to much more varied display arrangements. To take only a few examples, in the field of modern art light and sound complete the object; in scientific and technical museums the animated model that can be manipulated by the visitor is being further perfected; the documentary apparatus, strongly influenced by the graphism of the Bauhaus (the name given to the group of modern architects and designers working first at Weimar and then at Dessau, Germany, in the 1920s), becomes an element of the aesthetic framework; temporary and travelling exhibitions have also become increasingly complex. Theatre staging has influenced the conception of exhibition space, and modern museums such as the National Gallery of Victoria, Melbourne, are only empty volumes in which the covering on floors, walls, and ceilings can be modified in a few hours. Complex lighting effects can be achieved by the modulated and controlled use of natural and artificial light sources. Guided instruction, formerly undertaken by guides conducting homogeneous groups in parties, is now becoming incorporated in the form of record machines as part of the display framework in support of the objects. Research and literature on display methods and techniques for all types of exhibition have proliferated, and during the 1960s the first private firms specializing in museography appeared, while the most important museums greatly strengthened their workshops in matters of exhibition design and mounting. Similarly, research on the museum public has led to the constant adaptation of display to meet the needs of the visitor.

Exhibition techniques

Travelling exhibitions

The improvement of transport facilities and the development of the museums' cultural role, with the diversification of their activities, have led to a multiplication of temporary and travelling exhibitions on a national and international level. The latter can be prestige enterprises, organized, say, on the occasion of world fairs (New York, Montreal, Osaka), or sponsored on a bilateral basis by two states. Most of them, however, have a purely scientific and cultural character, or they may be educational. In this last case, the program of the exhibition is drawn up by an expert or experts, who also take charge of the display as a whole, while the details are taken care of by the museums taking part. The principal problems are insurance, the increasingly prohibitive cost of which (due to the influence of the art market on the estimated value of the works to be insured) is a heavy burden for the museums concerned and gives government-organized exhibitions an advantage; and also security and transport, particularly with regard to the packing and unpacking of the crates.

Other operations. *Public relations.* In the past the public was simply a security problem, and little else, to a museum, because of the risk of vandalism and theft. These risks still exist, of course, and have multiplied with the growing popularity of museums and the increase in the value of their collections, above all those of art museums. But growing consciousness of their social role has led museums to create new services: publicity, public relations, information, documentation. The increasing flow of visitors has begun, especially in the tourist towns of Europe and North America, to pose problems of space and surveillance, on the one hand, and of air conditioning of galleries, on the other. Specialists have begun to realize that these problems are not very different from those affecting other institutions dealing with a mass public, such as theatres, operas, cinemas, and supermarkets. It is also the case that transport and visitor parking assume increasing importance in the site plan of a museum, while up-to-date electronic systems of surveillance have been widely adopted.

Assistance to research. For obvious reasons of conservation and management, a museum cannot lend its collections out for people to consult, as a lending library does. The problem of research, one of growing importance in a museum's everyday life, has already been mentioned. Every museum is also a documentation centre where objects, photographs, films, recordings, texts, and index cards ought to be readily available and be looked after by qualified staff. The library, formerly reserved for the exclusive use of the museum's own scientific personnel, has become a general library covering the range of disciplines represented at the museum. The film, photograph, and phonograph libraries fill a threefold role: as archives, as museographic reserves for use in display to the public and for cultural activity, and as centres for consultation. It is not unusual, especially in developing countries and some provinces without research institutions, for a museum to be a headquarters for archaeological and ethnographic expeditions, for the inspection of historic monuments, for local learned societies, and sometimes even for teaching at the university level.

Conflicts of interest in museum operations. It stands to reason that the development of a museum's social and scientific missions involves internal tensions in the heart of the institution and its personnel. Whereas in the past a museum was a temple, a depository of objects and works for the enjoyment of a minority, it must now satisfy conflicting needs: to educate, to entertain, and to contribute to the development of knowledge. Too many museums still consider that research is their essential job and attach little importance to their cultural and social role. These museums appear static, shut-in upon themselves for the benefit only of their scientific personnel who can devote themselves to basic research without sufficient justification, vis-à-vis the community. Other museums, in increasing numbers, have devoted all their means to education and culture, becoming many-sided cultural centres without the necessary scientific basis. In fact, like a university, a museum should reconcile the requirements of science, which cannot proceed beyond research, and of culture, which should be oriented toward as large a public as possible. Various writers, moreover, see the museum's future as an institution situated at the city centre and housing its collections and its research and conservation activities there but decentralizing itself over the entire urban conglomerate or rural area in carrying out its cultural and educational activities.

II. Types of museums

The origin of museums is found either in the royal or ecclesiastical treasures of the Middle Ages or in the cabinets (little rooms) of curiosities and specimens first kept from the 16th to the 18th centuries. In a general way, art museums may be said to have sprung from the treasures, including royal or princely collections, while science museums have inherited the cabinets of curiosities. It was at the end of the 18th century, under the influence of the French philosophes and the Enlightenment, and at the beginning of the 19th under that of scholars that the classification of museums into categories, according to the terminology of the disciplines to which their collections belonged, was born. This first typology, which remains in force in a number of countries and which is still reflected in the names of the specialized committees of ICOM, distinguished between the museums of fine arts, applied arts, archaeology, history, ethnography, natural science, science and technology, regional and local museums, and specialized museums. Meanwhile, the evolution of museums since World War II has made the barriers between the disciplines and types of collections disappear progressively. It seems more justifiable now to divide museums into three categories: artistic or aesthetic, historical, and scientific.

ART MUSEUMS

Kinds of art museums

Art museums are those whose collections are conceived and displayed essentially for their aesthetic value, even if the objects they enclose are not all works of art in the intention of their creator. In this way, in addition to the museums of painting, sculpture, decorative arts, applied and industrial arts, one may include most museums of antiquities, of folklore, and of primitive arts. Some museums, generally those of more recent creation, demonstrate such arts as those of film, music, and theatre. As has been said above, the work of art is displayed for its own sake in order that a direct and intimate contact between the visitor and the object may be established. The state of preservation of the work, the quality of restoration, the environment it is given, notably the background and lighting, assume special importance. The didactic character of the display has disappeared almost entirely, since enjoyment carries the day over acquisition of knowledge. Concerning the fine arts, the masterpiece is the rule, and here temporary exhibitions allow the public to become familiar with collections that the museum could not acquire because of the rarity and high price of masterpieces. For the other arts, it is the objects as a whole that are important: series of tapestries, period rooms, services of silver plate, and so on. Regarding primitive art or folklore material, this art tends to link itself with the fine arts, to the extent that the modern socialization of culture no longer allows the art of the elite, in the European sense of the term, to be considered the only valid art. In all circumstances, the significance of the work of art, its functional or religious intention, is hardly important: a 15th-century virgin and child are not presented as a cult object any more than is an African mask. The danger of this separation of beauty from function is that the display risks becoming theatrical, all the more as museography attaches as much importance to the aesthetic beauty of the framework as to that of the object. In Italy, for example, in the museums of antiquities, stands and showcases are often works of art in themselves, but they therefore hamper contact with the objects they contain. Artificial light is usually considered preferable for paintings, natural light for sculptures. Precautions need to be taken to avoid alterations resulting from the lighting (*e.g.*, the level of

lighting, infrared and ultraviolet radiation). Museums of modern art occupy a separate place among art museums. They are, in effect, given up to art in the process of development and to the aesthetic environment of contemporary life. So it is natural that their activities are of a character that is above all temporary, reflecting the diversity of aesthetic movements and the problems of the moment. The museums of modern art of Stockholm and New York, the Musée d'Art Moderne de la Ville de Paris, the Stedelijk Museum, Amsterdam, among others, are also places of experimental action, open to all forms of art, comprising the theatre and music. The open-air sculpture museum is a recent form of exhibiting contemporary art, as for example, the Billy Rose Art Garden of Jerusalem and the Hakone Open Air Museum of Japan.

Museums of modern art

The public of a museum of art is essentially a public of amateurs and the curious, and this type of museum is of particular importance in the framework of leisure and tourism. The question of education raises specific problems: when the system of primary, secondary, and higher education gives art a large place in it, the art museum becomes an integrated part of such a school system, bringing it the help of its visual resources, as is the case in Great Britain; if a school does not teach art or does not teach it well enough, the museum should alleviate the deficiency and find new methods—whence the creation of children's museums and of rooms of aesthetic initiation.

HISTORICAL MUSEUMS

All museums where collections are conceived and presented in a historical perspective should be classed as historical museums, their object being essentially to document a chronological sequence or an ensemble representative of a moment in an evolving pattern. The Musée de l'Histoire de France, created by King Louis-Philippe (1773–1850) at Versailles, displays, with the aid of pictures, the outstanding events and persons in the country's history during more than 1,000 years. Few museums, however, are devoted entirely to the history of a country, although that formula is being used by some countries that have recently become independent in order to arouse national consciousness and a sense of cultural–historical unity. In the countries of Europe and America, history museums of provinces and towns are found frequently: the Musée de Bretagne (Brittany) at Rennes (France), city museums at Warsaw, Dresden, Stockholm, Amsterdam, Rotterdam, London, New York, and the Heimatmuseum in Germany. The great national historical museums are the Museo Nacional de Historia of Mexico City, the Museum of History and Technology of Washington, D.C., the Musée National de Versailles et des Trianons, the Muzium Negara of Kuala Lumpur, Malaysia, the National Museum of Budapest, and the national history department of the Rijksmuseum, Amsterdam, which was inaugurated in October 1971. The most modern of these museums display the whole history of a country, region, or town from its origin (including natural history and geography) to the present age, without neglecting the perspectives of actual development, such as land acquisition and urban growth. They integrate science and art, call upon all kinds of audiovisual media, and give a large place to written graphic documents, to ecological reconstructions and models, and to maps. The exhibitions are often housed in an ancient building, itself a historic symbol.

National historical museums

Other kinds of history museums include those at archaeological sites, museums installed in a historic monument or on a battlefield, and personal memorial museums. Examples of these include Mount Vernon (Virginia, for George Washington), Archaeological Museum (Taxila, Pakistan), Lenin Museum (Ulyanovsk), and Museo Arqueológico de Teotihuacán (Mexico). These museums are often didactic and comprise reproductions and models; sometimes, and especially in the case of memorials, they appeal to the feelings of the visitor by respecting the actual setting of a person's life and the simplest objects that belonged to him there: this is the case with museum homes, such as that of the writer Ernest Hemingway (1899–1961) in Cuba.

Finally, there is the historical museum, dedicated to a period in the history of a country or region: such are ethnographical museums, better named in Scandinavia as museums of cultural history, whether they are presented in traditional fashion in a building or in the open air. These museums—for example, the Skansen of Stockholm, Colonial Williamsburg of Virginia, the Musée National des Arts et Traditions Populaires of Paris, the Village Museum of Bucharest, Romania, and the Museum of English Rural Life of the University of Reading, Berkshire—retrace the preindustrial or protoindustrial epoch. Attractive to the public, especially in open-air form, these museums contribute notably to the understanding of and respect for the human environment and fast-vanishing cultural traditions. Developing countries attach importance to them in their efforts to reconcile the survival of traditional values with industrialization. One of the most successful examples is undoubtedly the National Museum of Niamey (Niger), which combines a zoological garden, an open-air museum, and an ethnographic exhibition.

Museums of cultural history

The problems of display that arise for museums of the historical type are specific, in that objects are shown for their functional and historical significance, and the written or recorded commentary, the relationship between groups of objects, the scientific rigour of the program, and the flexibility of display take on a determining importance. In the same way, reserves and study galleries take up a great deal of room in these buildings because the breadth of the subjects treated is such that knowledge about them makes continual progress. Frequent temporary exhibitions, modifications of permanent displays, and scientific activity at the museum and on field expeditions are vital for the museum, except where very specialized museums (*e.g.*, memorial museums) are concerned.

SCIENCE MUSEUMS

Museums of natural science, of exact or applied science, and technical museums (except for museums of the history of science and technology, which are counted as historical) are classified as science museums. The task of science museums is to communicate in three-dimensional form a scientific spirit and mentality, to arouse natural inclination for science, to give information on research and progress, to give each person a sense of sharing in technological development, to make people wish to understand, appreciate, and conserve nature and man's natural environment in an ecological and historical perspective, so as to demonstrate the evolution of nature and man. The museums of natural science that were created first were the Muséum National d'Histoire Naturelle of Paris, at the end of the 18th century, and the British Museum (Natural History), of London. They multiplied considerably during the 19th century with the development of natural science, and, in the United States and Latin America, they have absorbed physical anthropology and then social anthropology. If their evolution has been slow during the first half of the 20th century, they have been transformed quickly enough since 1950: witness the renovated sections (since 1960) of the American Museum of Natural History, New York, and the Museo de Historia Natural, Mexico City. In Europe, their rise has been less, although mention may be made of "pilot" museums such as the Moravian Museum at Brno, which combines biology, genetics, and other sciences, and the Museum of Zoology of the University of Copenhagen.

The functions of science museums

Museums of the applied sciences and technical museums, of which the most celebrated example is the Deutsches (German) Museum of Munich, are more recent. Some of them are devoted solely to the sciences, such as the Palais de la Découverte (Palace of Discovery) of Paris. Others are essentially technical, such as the Museum of Science and Industry of Chicago. Developing countries are anxious to establish these museums for the sake of popular education; and India has succeeded in establishing three, at Calcutta, Pilani, and Bangalore. Large industrial companies see in these museums a means of publicizing their production and applied research.

Use of models in science museums

All of these museums associate the real object (natural specimen, machine) with the model (diorama, dummy, or animated model) and with demonstration (physical experiments, planetarium exhibitions, field trips). Their aim is essentially didactic, and their educational activities are very important and decentralized. Some museums organize travelling dismountable exhibitions to serve distant regions and those least developed technologically; others use mobile museums, as at Calcutta; still others offer various facilities to schools that do not possess laboratories and workrooms, as at Milan. More recently, the tendency has been to renounce the name of museum and adopt instead the expression centre of science and technology, which is the case with institutions sponsored by UNESCO and the new Ontario Science Centre.

Science museums are probably the best attended and the most active of museums. They have made the most vital progress in museology and in the techniques of presentation. Within the framework of positive efforts to conserve the environment, museums of natural science have a particular role to play in informing and persuading the public. On a scientific level, these museums have also supplied researchers with documentation (millions of specimens) permitting the advancement of knowledge in a multidisciplinary perspective.

SPECIALIZED MUSEUMS

Open-air museums. Mention should be made of museums that have not been established in a classic exhibition building but that have been created instead within the bounds of a garden or park, the exhibited objects being set up in the open air. These museums are dependent upon all the scientific disciplines but raise special museological and museographic problems. One finds, thus, ethnographic museums (Skansen at Stockholm, National Museum of Niamey, Rijksmuseum voor Volkskunde "Het Nederlands Openduchtmuseum" at Arnhem), museums of natural science (zoological or botanical gardens), archaeological museums (Prehistoric Museum at Moesgård, Denmark, archaeological sites considered to be museums), museums of art (sculpture gardens), and museums of the history of technology. Though the administration, services, and reserves are lodged in spacious buildings, special problems of conservation and of the circulation of the public remain. Because of their character as parks, these museums are very attractive and are often described by visitors as "more lively" than traditional museums. Furthermore, their dimensions permit the reception of a greater number of visitors than any other kind of museum. It is probable that their future is unlimited, above all in developing countries, where the land is not dear, even in the centre of cities. From the point of view of presentation, one finds here the means of artificially creating a more satisfactory environment than the ecological reconstructions of indoor museums. There is, however, a danger that exists in the temptation to "fake," especially in open-air ethnographic museums. These museums should in fact obey extremely strict rules, on the scientific level, for collection and presentation. Finally, conservation is made difficult by climatic conditions, for the collections are left outdoors under the open sky and are subjected to variations in temperature and humidity, without the possibility of control or correction.

Though they can all be fitted into the categories already described, their specialization itself poses particular problems for specialized museums. Some among them are of great importance and have contributed much to the development of museology: like the regional museums, they are usually obliged to call on a wide range of techniques and to integrate many scientific disciplines. A transport museum, for example (such as those of Budapest, Lucerne, and Dresden), will make reference to geography, economics, history, and sometimes even to art if ancient or modern artists have illustrated aspects of transportation. Thus one finds museums of wine, clock making, fabrics, hygiene, musical instruments, ceramics, fishing, and other subjects. Many of these museums describe limited economic activity (mineral exploration, the shoe industry) and, in this case, are more or less tied to the industry they illustrate; others treat of an artistic activity (theatre, marionettes, music, cinema) and have a threefold aspect—historical, artistic, and technical.

Children's museums

For want of something better, one may classify under this heading museums that appeal only to one section of the public, as do children's museums, when the latter are not simply a section for children in a big museum. The Brooklyn Children's Museum of New York, the Bāl Bhavan of New Delhi, and the Children's Museum of Algiers have their own collections and their own activities, which make appeal only to children. It is the same with museums for the blind and with some university museums that serve only an already very specialized student public. These museums have their systems of presentation and collection and have programs of activities determined in their application by the homogeneity of their public and, except for university museums, not generally undertaking any scientific research.

Regional museums. Regional, or local, museums devoted to all aspects (natural, historical, and artistic) of a region or district, such as are found in western or eastern Europe, India, Mexico, Chile, and Canada, are of interest through their multiplicity of values. At the same time, they serve a community that is culturally homogeneous or is destined to become so. Their first responsibility is therefore to reflect this community and to give value to its traditions and its creative spirit; the second is to open the community to the outside world, particularly in serving as a relay station for the central museums, which can send them exhibits or collections on deposit that illustrate the cultural and natural, national and international, heritage. Their usual handicap is to have been founded by a patron or by a local learned society that imposes obligations that are incompatible with the development of a modern museum. The majority of them still consist of departments or sections representing scientific disciplines: this setup, which is indispensable for reserve collections and for research, should gradually disappear to give place, on the level of public exhibition, to a better integration as a museum of history. Some have already taken this step, such as the Kanagawa Bunko Museum of Yokohama.

Other institutions functioning as museums. Besides museums properly speaking, numerous institutions employ the same language of presentation regarding the object itself, sometimes with a directly cultural aim, sometimes with a view to publicity or the diffusion of various aspects of knowledge. In the first category are institutions such as archives and libraries, whose collections are not intrinsically of the same nature as those of museums but whose role is equally cultural and educational and, indeed, scientific, and also cultural centres and exhibition halls (such as Kunsthalle in Germany) that do not have permanent collections. These last are not recognized by the American Association of Museums, whose rules of accreditation are very strict, but they are recognized by the International Council of Museums (ICOM), which seeks to associate them with museums in order to make them observe the professional regulations established for museums, in the interest of the collections and the public. One must not neglect these institutions, which in many cases constitute useful and natural staging posts for museums, particularly for their travelling exhibitions.

Archives and libraries

Distinct from those above, many enterprises have been created by business and industry, whether they be company museums such as the Evoluon, built by Philips N.V. at Eindhoven, The Netherlands, exhibition fairs, or small specialized displays, not to speak of international fairs and exhibitions the contents of which are partially cultural and partially scientific, technical, and economic. It is difficult to say, in each case, which of them is an activity of museum type and which is simply publicity or a commercial event. Nonetheless, their techniques of display, use of animation, and contact with the public are often worth the attention of the museologist. That is why, in 1969, the museums of The Netherlands organized their country's museums fair, called Musement, in the halls that house the commercial fair of Utrecht; this allowed a number of tests of public reactions and of methods of presentation to be made.

III. Museum architecture

Influence of the past on museum architecture

Just as with the content of a museum, the container itself is largely fashioned by the past. First, a museum's history conditions its architecture; the religious treasures of the Middle Ages were kept in churches and monasteries, and the royal collections were lodged in palaces or castles; art itself was closely linked to the dwellings of the social class that generated its creation. When the relics of feudalism made room for the political and cultural democratization that developed progressively during the 19th century, many palaces, symbols of the old order, were transformed into museums, while, in countries or towns without palaces, the new museums imitated, with greater or less exactness, temples, palaces, and castles. So it is that the first great U.S. museums adopted Neoclassical or Palladian architecture and that the museums established by the British in India reproduced Neo-Gothic or Victorian architecture. In the Soviet Union, after the Revolution of 1917, the royal or aristocratic dwellings were transformed into houses of culture for the people. In central and western Europe, this tendency was prolonged and still continues for other reasons: the great number of historical monuments left to the state or to local administrations, following their abandonment by their traditional owners or their confiscation by a new political regime, obliged those responsible for them to find new uses for them; and their transformation into museums became the customary method to preserve and reanimate them.

Since World War II change in museological theory has caused the theory of museum architecture to evolve, at least in advanced countries, and national and international meetings of experts have attempted to define a doctrine of functional architecture for museums. Successively now will be examined: the traditional temple museum; the monument transformed into a museum; the architectural style of recent museums; the most advanced theories actually put into effect; the problem of aesthetics versus function in modern museums.

TEMPLE OR PALACE MUSEUMS

The British Museum in London, the complex of the "Museum Island" in East Berlin, the Metropolitan Museum of Art in New York, the National Gallery of Art in Washington, D.C., the National Gallery, London, and many others are structures of monumental character, generally imitating Greek temples in their exterior appearance: colonnades, pediments, heavy classicizing decoration. Within are rooms covered with marble or hard stones, with lofty ceilings, vast staircases sometimes occupying a third of the available space, inconvenient circuits for visitors, lighting as high up as possible—all elements designed to emphasize the value of the collections of masterpieces in a solemn and almost religious atmosphere. The visitor has to be impressed, dazzled by the splendour of past centuries. It is a curious thing, for the same phenomenon can be found in museums of natural sciences, such as the American Museum of Natural History in New York or the Muséum National d'Histoire Naturelle in Paris. In all these edifices, the space devoted to reserves and services is insufficient, inadequate, and impractical. This type of museum, born at the end of the 18th century and in the 19th, continued to be built thereafter. In some way, one can think of the Israel Museum in Jerusalem and the Nationalgalerie in West Berlin—this last the work of the German architect Ludwig Mies van der Rohe —which date from the 1960s, as representing the modern conception of the temple museum. Such museums are generally constructed in the centre of old towns, and their adaptation to the needs of the modern world is inconvenient and costly. In other respects, they often become historical monuments themselves, and their extension raises difficult problems, as with the Tate Gallery in London. Educated traditionalists are attached to them, while the "new public" and the contesting groups take exception to a stultifying consecration of the heritage. The solutions adopted for the transformation of these edifices to turn them into modern museums are generally radical: installation of false ceilings or mezzanine floors, suppression of natural light by obscuring windows and glass roofs, subdivision of rooms, suppression of decorative painting and sculpture, and so on. In many cases interior courts have been covered over in order to install reserves there, and extensions have been constructed on adjacent land. Among the most spectacular achievements of modernization may be cited the covered courts at the Rijksmuseum of Amsterdam, the projects being carried out at the Metropolitan Museum of Art of New York, the new installation of Greco-Roman rooms at the British Museum of London, and the internal remodelling of the whole of the National Museum of Budapest.

The modern temple-museum

MONUMENTS TRANSFORMED INTO MUSEUMS

The great majority of museums created between 1800 and 1950 have been housed in existing historical monuments: the Louvre in Paris, the Hermitage in Leningrad, the Uffizi in Florence, the Prado in Madrid, and innumerable provincial museums. Developing nations forgoing construction of new buildings as museums have also followed this trend: the Queen's Palace at Tananarive (Malagasy Republic), the National Museum at Libreville (Gabon), the Fort Jesus Museum at Mombasa (Kenya), the National Museum at Bangkok, the Museu Nacional of Boa Vista at Rio de Janeiro. These museums offer problems of great architectural interest, for the buildings utilized are no easier to transform than the temple-like edifices, and they also impose restrictions in respect of their aesthetic and historical character. Their restoration and equipment for use as a museum are carried out by the same architect, who is seldom chosen by the museum director but is more often imposed by the administration responsible for historical monuments and public works. Italy and France have gathered experience from these problems, for which it is impossible to lay down rules, the building being a work of art in itself (*e.g.*, the case of the Petit Palais at Avignon, France, which is to become a museum of Italian primitive art).

Limitations of historic buildings

The position of this sort of museum is seldom good, it being situated in the historic centre of town, with attendant problems of traffic circulation and access. The humidity of the walls and the internal atmosphere are often damaging to the collections. The natural lighting is either insufficient or too dazzling, in which case one must either add or suppress windows. Space is not flexible enough, and the museographic equipment needs to be adapted to work on inconvenient surfaces. Generally speaking, their intimate character makes these monuments better suited for museums of painting, for memorial museums, for specialized history museums, and even for the installation of private collections. The didactic aspect of contemporary museography is not at ease there, and, with the exception of a few vast buildings like the Louvre and Versailles, these museums cannot receive more than a limited number of visitors at a time. One notes, however, the great success of the Historical Museum of the city of Warsaw, installed in eight bourgeois houses joined together. It is usually necessary, as in the example above, to make the program of the museum coincide closely with the character of the monument and to avoid installing a folklore museum in a royal palace or a museum of secular art in an old church. Another important element to take into consideration is that of the security of the collections, to the extent that historical monuments are, in general, more susceptible to fire or theft than other museums.

Nevertheless, the tendency to make use of this type of building will continue in countries rich in monuments, and it is a question then of a form of museum that permits the safeguarding, for cultural aims, of important evidences of the heritage.

RECENTLY CONSTRUCTED MUSEUMS

It is necessary now to distinguish again between museums of private or semiprivate status (endowed with a board of trustees or a council) and museums of public status. In the first, the influence of the director and the scientific personnel in the choice of architect and those over him is often weighty: those responsible for the museum must establish a program in accordance with which the architect will be chosen and by which he himself will then

have to abide. In the case of the publicly run museum, the personnel do not possess the same liberty. The architect is chosen either from among the official architects of the administration or municipality or by a national or international public competition; it is then very difficult for the museum personnel and for the director to bring influence to bear on the work of a technician who has been more or less imposed upon them. Moreover, the decision to build a publicly regulated museum always has, in one form or another, a political character in the widest sense, and the appearance of the building does not belie it.

Museums designed by modern architects

Among the most modern museums, many are in themselves monuments, heirs to the concept of museum-temple or museum-palace. The great modern architects, such as Le Corbusier of Switzerland and Mies van der Rohe and Frank Lloyd Wright of the U.S., in the museums of Tokyo, Chandīgarh, Berlin, and the Solomon R. Guggenheim Museum, in New York, seem to have intended to show in practice their own ideas of museology and thereby to express enduringly their own artistic genius. This kind of solution has seldom led to museums that are really effective, even if their aesthetic character assures their fame. These buildings are, in fact, conceived from the outside, and their interior space, though occasionally suitable for an artistic exhibition according to the norms of the time of their construction, cannot be adapted to the incessant changes in museological concept. Other museums, less prestigious and less celebrated, seldom situated in capital cities, owe their construction to architects who are less well known but who have been willing to collaborate closely with museological personnel capable themselves of preparing a precise program. Some examples, among many others that have been completed or are in course of construction, are the museums of Edmonton (Canada), of the Abegg-Stiftung at Riggisberg (Switzerland), of Arts et Traditions Populaires at Paris, of Antiquities of Sicily at Syracuse, and, above all, the complex of modern museums of Mexico, at Mexico City and in the provinces. They are characterized by the systematic use of the most modern material and techniques, the spatial division of service areas and zones open to the public, the creation of flexible space in volume for permanent and temporary exhibitions, the extent of the reserves and of research facilities, and by a carefully studied ground plan to work in conjunction with all-round urban planning and with means of transportation and parking (many levels of parking, for example, under the National Gallery of Victoria in Melbourne). If museums of this type are still found situated at the centre of a city, there has been a tendency in favour of museums in the country (Rijksmuseum Kröller-Müller at Otterlo; Louisiana Museum near Copenhagen) or in new towns.

In developing countries, for want of means, new building is rare and the lack of qualified personnel sometimes allows great freedom to the architect, who is usually appointed by the administration (public works department). The result is mediocre more often than not, because a program is lacking or else because it has been prepared by an expert strange to the country and ignorant of the community's requirements. The architecture itself scarcely differs from that of a school or a hospital built by the same architect. In the richer developing countries, the wish to construct museum-monuments linked to prestige urban planning (Abidjan, in the Ivory Coast; São Paulo in Brazil) has begun to make its appearance.

THEORY OF MUSEUM ARCHITECTURE

Collaboration between architect and museum directorate

Several international meetings have studied the problems of architecture, especially the relations between the architect and those responsible for the museum. Two of them, organized by ICOM in 1961 (Milan) and in 1968 (Mexico), have contributed greatly to the laying down of the main lines of an international code linked to the development of museology and museography. The director, assisted by his scientific personnel, is considered as the "master of the work"; by this title he lays down the program and chooses the architect (or gives his approval to the choice). The architect is the "master of the workmanship" and prepares the project in accordance with the program, with which he is provided. Then, the master of the work and the master of the workmanship cooperate in perfecting the program and the project derived from it until the building is completed and the collections and displays are installed. Increasingly, specialists of other disciplines—urbanists, sociologists, climatologists, specialists in conservation—are added to the basic team and intervene at various stages of preparation and completion.

The building should be flexible and extensible: interior and exterior circuits should separate the public from the service personnel and the unexhibited collections as far as possible. Environmental conditions (lighting and climate) should be studied before the adoption of the project and at the same time as the museological program. The interior fittings and displays depend, at least in their main lines, upon the cooperation of the director and the architect, since the choice of technical solutions will be conditioned by the architecture and the program. When the museum is functioning, it should be the subject of an evaluation through a study of the public and the personnel to discover their reactions. This is why it is desirable that the architect responsible for the construction should remain in charge of the building after its inauguration, in order that he may be in a position to correct his work should the need arise.

The program's importance should be emphasized. Its quality will condition that of the building. Its establishment is long and exacting labour, which relies on a clear idea of the aims and functions that it is desired to give the museum, the needs of the community and of research, the nature of the collections, the requirements of display and conservation, and the scientific conditions. The program must be prepared and discussed, not only with the scientists of the disciplines concerned but also with those who will be responsible in the future for conservation and education, representatives of the regular users, and the museographers who will be in charge of displaying the collections.

AESTHETIC AND FUNCTIONAL ARCHITECTURES

Until the 1950s it was more important, as has been noted, that the architecture of museums should reflect the concept of a museum as the sanctuary of beauty and of the past: the building reflected a striving for the monumental to the exclusion of real architectural creation for the sake of striking the imagination and providing a jewel case worthy of the treasures enclosed in it. In the same way, newly constructed museums in several developing nations still go on imitating traditional architecture: the Muzium Negara of Kuala Lumpur and the National Palace Museum of Taipei. Since 1950, influenced by modern architectural trends and the development of museology, two theories have emerged: a museum will either be a cultural monument in itself, whose siting, external appearance, and materials employed will be a symbol of its role at the centre of community and as a key element of the new city; or it will simply be a box concentrated on its interior and conceived strictly in terms of its collections and the intellectual satisfaction of its visitors. The first concept applies to some of the most celebrated museums built since World War II: the Solomon R. Guggenheim Museum, Oakland Museum (California), the Nationalgalerie of West Berlin, the Israel Museum in Jerusalem, and others. Such edifices are in general neither flexible nor capable of extension, even when they are in complete accord with the program designed for them in terms of current museological doctrine. Consequently, in conceding that they may even become classics of 20th-century architecture, it is unlikely that they can be easily adapted to the museum's development needs. The second concept is illustrated by the Musée National des Arts et Traditions Populaires of Paris, the principal new museums of Mexico, the National Gallery of Victoria of Melbourne, and the Louisiana Museum of Copenhagen, as well as others. These museums, which can sometimes also be considered as masterpieces of architecture, are essentially conceived in terms of the collections, and their aesthetic character is secondary, without being altogether excluded.

Modern architectural trends

Though the functional museum is clearly the more de-

sirable, there is no disguising that it will be difficult to escape from the museum-monument, often insisted on for reasons of prestige by the authorities on whom it depends. The real solution lies, as has been mentioned already, in the close cooperation between the museum director and the architect.

Concerning the interior of the museum, which is often entrusted to the architect, and the special museum fittings, the principles are the same. The aesthetic should be a secondary consideration in the decor, in the furniture, and in the lighting; the most important thing is to begin by giving attention to the objects and to facilitate the visitor's contact with them. This is particularly true of the design of showcases and stands, the forms and materials of which still too often have aesthetic pretensions, so that the collections they contain or support can no longer be "seen" for themselves.

IV. History of museums

The history of museums and of the institution "museum" through the ages is linked closely to the history of society and to the evolution of knowledge. In fact, every museum is the work of one or more founders, it addresses itself to a well-defined public, and it is dependent upon one or more sciences. In the paragraphs that follow, the history of museums will be presented from an essentially sociological point of view. The influence of the history of sciences on museums will also be glanced at.

THROUGH THE 17TH CENTURY

Medieval museums

The museum was born, so far as collections are concerned, in the geocultural area defined by the Mediterranean Basin and in the countries situated immediately to the north, in the zone of Greco-Latin civilization. Several phases can be distinguished. First, the Middle Ages, characterized by feudalism and Christianity, saw the first collections appear in three institutions dominated by the clergy as the sole custodians of art and science: the church, the monastery, and the university. Works of art, manuscripts, apparatus, and curious specimens were kept for their religious, scientific, or economic value, seldom for aesthetic reasons. They were certainly not museums as such, for the collections were neither for public display nor for systematic research. The dates of foundation are not known, but the religious treasuries of the Abbey of St. Maurice in Valais, Switzerland, of Aachen, Germany, and of Monza, Italy, comparable, moreover, to the collections of Japanese temples, contain objects extending from the late Roman Empire to the late Middle Ages. The transformation of these treasuries into modern museums is recent and dates practically from the development of international tourism.

From about 1400 to 1700—that is, from the beginning of the Renaissance to the apogee of Classicism—the rediscovery of Greek and Roman antiquity, the fragmentation of Christiantity and the secularization of education, the general emergence of an aristocratic and bourgeois cultivated class, and the great discoveries resulted in a great increase in collections, at first in royal collections (explorers' spoils, diplomatic presents) and then in the private in emulation. The collections of Vienna and of the Escorial, near Madrid, succeeding those of the dukes of Burgundy and the kings of Spain, those of Versailles and of Florence and of the Vatican, gave birth to the greatest museums of our time. Naturally, the collections were open at that time only to members of the cultivated class, if and when they could enter their owners' palaces. There was no desire to collect systematically nor any attempt at conservation or valuation—merely curiosity, a taste for luxury and prestige, but also the birth of a sense of beauty and a consciousness of the scientific value of certain specimens. One finds during this period only one exception, that of the creation of the Ashmolean Museum at Oxford (1683), the first museum deserving of the name and the first university museum.

THE 18TH CENTURY

The 18th century saw a rapid evolution and the birth of the principal public collections at the instance of artists and naturalists; among these were the British Museum, the Louvre, the Muséum National d'Histoire Naturelle, the Vatican Museum, and others. This was due, in the most important countries of Europe of this time, to the growth of the bourgeoisie in importance and numbers, to the broadening of public education, to the rise of research, to the birth of a first form of tourism, to the movement of the Enlightenment and the Encyclopaedic spirit, and to the growing taste for the exotic. The aristocracy began to lose importance as a governing class, the aristocratic, intellectual, and cultural life began to become democratized, and things from abroad and of the past were becoming appreciated. It was at this time that the collections from Siberia and Alaska were set up at St. Petersburg (Leningrad), which created the great collections of natural science in parallel with the first attempts at systematic classification. The opening of salons of painting and of the royal galleries of pictures and sculptures placed past and present art within view of a fairly large public. The most distinct movement toward the creation of museums proper was to be found in Great Britain, in spite of numerous restrictions that still limited access to collections. Meanwhile, a public had been called into existence, authority was given for research on museum objects, and concern for conservation showed itself in the many restorations carried out on works of art, by distinguished artists, not necessarily for the best. Simultaneously, the area of the collections, if not yet themselves museums, broadened with the dissemination of ideas and extended to the Slavonic world and to North America.

The first public collections

1790–1850

The first half of the 19th century (which began in fact with the French Revolution, in the last decades of the 18th) saw the rise, in Europe and in the United States, of the middle classes, the advent of revolutionary ideas, and the development of the sciences, henceforth set free from religious constraint. From the first the French Revolution opened the Louvre to the people. (The Russian people had to wait for their Revolution of 1917 for the same privilege.) During that period France and Great Britain opened the way to the museums: not only were the great museums of today opened, but many provincial museums were created on the initiative of local learned societies of intellectuals wishing to serve their community. The private collections, formed since the beginning of the 18th century, became museums in their full right, either as the result of political events that brought about the nationalization of the property of the former governing classes or by voluntary gift of their owners, won over to new ideas; the architectural type of museum-temple became widely diffused, and the feeling for the past, awakened by the discoveries made by the first archaeologists in Egypt, in Italy, and the Near East as well as by the fashion for Gothic art (this word covering all the art of the Middle Ages), absorbed the public. For the first time, the museums came to play a major role in the creation of styles, such as the Neoclassical and the "troubadour" style. Artists and men of culture went on pilgrimage to Rome, where the Vatican Museum offered a vast range of aesthetic experiences. During and after the Napoleonic era (following probably the infatuation of Napoleon I and his court for works of art) all the great cities of Europe opened museums. Princess Czartoryska even organized in the park of her castle at Pulawy, near Warsaw between 1801 and 1809, the first Polish museum, embellished with the motto "the past for the future," in imitation of such patrons as Alexandre Lenoir, who opened his Musée des Monuments Français in Paris in the middle of the Revolution, seeking to preserve for the future the most remarkable monuments of the past that were in danger of being destroyed by the political and popular upheavals. The museums of this era were essentially of an artistic and historical–archaeological character and were to play a considerable part in developing a consciousness of the intrinsic value of European or Western civilization, at the time of the rise of European nationalism and the imperialism that accompanied it. On the other hand, this

Nationalization of private collections

museological explosion, coinciding with the Industrial Revolution, was to preserve an important part of the cultural values of the ancient world, which otherwise would undoubtedly have been destroyed by technical and economic progress and new aesthetic fashions.

1850–1950

The hundred years from 1850 to 1950 saw a number of phenomena, which, influencing one another, led to the formation of a "world of museums" that carried the germ of the present situation: the continuing development of the Industrial Revolution in Europe and the United States; the institution of democratic regimes (in spite of dictatorial regimes, mostly short-lived); fundamental transformations in the sciences and accelerated technical progress; extension of European–American colonialization, economic or political, to the whole world; a rise of the means of communication; and the democratization of education and culture. On the level of museums, the larger institutions created in the course of the preceding periods were reformed and enlarged; national centralized administrations were organized in Belgium, France, Italy, and the Soviet Union; they ran the central museums directly and controlled the provincial ones. These last were enriched notably by deposits from the central museums and improved their management by throwing off the tutelage of local learned societies composed largely of amateurs. In the United States and in Canada, the multiplication of private museums commenced and has not yet ceased: Royal Ontario Museum (Toronto, 1912), Metropolitan Museum of Art (New York, 1870), American Museum of Natural History (New York, 1869), Museum of Fine Arts (Boston, 1870), Museum of Modern Art (New York, 1929), Art Institute of Chicago (1879), and many others. In the colonial countries, under the influence and often on the initiative of the colonial powers and their scientific institutions or their patrons, museums were created after European models: Prince of Wales Museum of Western India, Bombay; Indian Museum, Calcutta; Coryndon Memorial Museum, Nairobi. In Latin America, which was liberated earlier from political colonization but remained closely linked with Europe and the United States intellectually and economically, numerous museums of art appeared, devoted principally to the more classic European art, while local arts, pre-Columbian and colonial, were neglected.

New types of museums

This same period saw, in addition, the birth of new categories of museums—of anthropology and ethnology, of science and technology, of the history of cities, of archaeological sites. It saw the discovery of the educational role of museums and the birth of a professional consciousness, with the creation of national associations of museum professionals (British Museums Association, American Association of Museums, Association Générale des Conservateurs des Collections Publiques de France, Scandinavian Museums Association, Deutscher Museumsbund), the publication of professional journals and books, the creation of the first international organization, the International Museums Office, dependent on the League of Nations, but enjoying a mixed statute, being at the same time both governmental and professional. The first researches on what is now called museology and museography were undertaken on the initiative of the great museums or professional organizations: studies on conservation, display, management, international cooperation, and the history of museums.

The evolution of the museum during this epoch reflects perfectly that of the society that gave it birth. Civilization, threatened by technical progress, sought to affirm the permanence of its intellectual values: the Western world, ruined spiritually by world wars and engaged in the "work" of colonization, tried to justify and maintain its fundamental unity. In the face of the difficulty of creating new forms of life and culture, a growing importance was attached to the preservation of the past; new ruling classes born without much past, especially in North America, adopted the past of classes that had disappeared and yet remained sentimentally attached to forms of popular cultural expression. Scientific knowledge became stricter and more systematic and was no longer content to speculate about ideas but sought the support of concrete things. Finally, in the artistic domain, a market was created, and of this museums were both origin and end product: the origin because they awoke the taste for collecting and "connoisseurship" and the end product because ultimately collectors give what they have assembled to museums, for economic, social, and prestige reasons.

The three principal facts of the century from 1850 to 1950 may be defined thus from the point of view of the museum concept: the idea that culture is for the people, as expressed by Lenin and by every action of the Soviet government since 1917; the patronage of a wealthy class at the service of the community, as can be seen in the United States in both large and small towns; the development of tourism, which has brought into favour museums of art and archaeology in countries of great cultural riches in the Mediterranean Basin and western Europe.

Effect of World War II on museums

World War II terminated this period, with upheaval in the museum world as elsewhere. In Europe, museums and art treasures were destroyed, and entire collections were transferred from one country to another; political and economic changes involved profound modifications in the conception of culture for some countries, while the United States, less affected on the domestic plane, was able to continue developing its museums and acquiring collections that had become too expensive for war-devastated and impoverished countries to buy or sometimes even to preserve.

1950–70

The two decades that elapsed from 1950 to 1970, once the wounds of war had been attended to, saw a renaissance of museums, which, without breaking with the major trends of the preceding period, required their adaptation to a world that was largely new. Most of the museums of Europe had first to be reconstructed or restored and to reconstitute collections that had disappeared or had been sent away to shelter during the war. The unprecedented development of tourism in the countries of the Mediterranean Basin, led, under governmental pressure, to a reorganization of the museum as an institution, which now no longer concerned itself with the national public, at least in the larger centres, but endeavoured to attract a mobile, sheeplike public of tourists, who "followed the guide." By contrast, in northern Europe and in the Socialist countries, museums turned resolutely toward education, cultural invigoration, and the service of the community: it was a time of great development of regional museums, open-air museums, exhibition centres, "museum forums," and museums of modern art. In the United States, the expansion of museums proceeded at an accelerated rate (as noted earlier, a new museum every 3.3 days in 1964). The price of objects in the art and antique market became such that a constant flood of imports from all over the world found its way to the shops and sale galleries of New York, by the mediation of sales rooms or the traffickers in London, Paris, and Zürich. More recently, between 1965 and 1970, social and racial problems have stimulated the appearance of a new type of museum, the neighbourhood museum, a sociocultural community centre, a virtual relay station for the great museums at the centre of the metropolis, which are increasingly criticized for their elitist and centralizing character.

Museum organizations

In the same period a profession or, rather, group of professions was created around international, national, and sometimes provincial associations and organizations. On the international plane, UNESCO, the International Council of Museums (ICOM) and its specialized international committees, the International Institute for Conservation (IIC), and the International Centre for the Study and the Preservation and Restoration of Cultural Property in Rome (governmental) organize the cooperation of institutions and professionals, diffuse information, and assist in exchanges and contacts. On the national plane, nearly every country equipped itself with a professional organization that sometimes has a permanent structure

and a secretariat, publications, and training activities. In some very large countries possessing numerous museums, provincial, regional, and specialized organizations were founded. In the United States, their proliferation is astonishing; alongside the American Association of Museums, one finds an Association of Art Museum Directors, an Association of Science Museum Directors, and regional conferences. The training of personnel has assumed considerable importance: chairs have been created in the universities of Brno, Leicester, Paris, Colorado (at Boulder), New York, and elsewhere. On the initiative of national associations and of ICOM and UNESCO, centres for museographic training have been created, and attempts are being made to unify and standardize their diplomas and programs.

Museological and museographical research, conservation studies, and sociological surveys carried out on the public have multiplied, expressing both anxiety and desire—anxiety of a museum that feels itself challenged, criticized, and overtaken by the progress of culture and society; desire, on the contrary, to adapt itself and find new methods.

Museums in Asia and Africa

As in Europe, but to a lesser extent, numerous people of Asia and Africa have known the idea of collecting. The Shōsō-in of Nara and the royal collections in Japan and China bear witness to this, going back to the 1st millennium of the Christian Era. Certain African tribes and secret societies have always possessed treasures of religious and genealogical significance. But the concept of museum, properly speaking, only appeared late, usually under Western colonialism or at least under the influence of Western ideas. In the 19th century, important museums were created in Latin America and Asia. European researchers in Africa assembled collections that were sometimes dispatched to museums in European cities and sometimes kept on the spot in embryo museums or, rather, depositories and centres of study. In most of these countries, however, genuine museums only appeared with political independence, the necessary condition of political and cultural autonomy. Only a few countries have succeeded in endowing museums that can be compared with similar European institutions and yet do not imitate them but find instead original solutions to problems of presentation, communication, and animation. Mexico, Japan, India, and Malaysia are excellent examples of success in this respect. In these countries, history, exact and natural sciences, and technology most often constitute the core of the most modern museums, because of the needs of education. Africa, more backward because less culturally independent, owes its only successes in museums to their systematic use as a factor in national unity and cultural identification, as at Niamey (Republic of the Niger), Kampala (Uganda), and Accra (Ghana).

In the less developed nations, problems of museum modernization, training of personnel, research, conservation, and studies on the public pose themselves with far more acuteness than they do in the industrialized countries of Europe and North America. If Japan and Mexico have discovered original methods of presentation, if India possesses three university departments of museology (at the universities of Baroda, Calcutta, and Vārānāsi), they are exceptions. The action of international organizations has been applied to putting at the disposal of these countries the means of undertaking the necessary reconversion of the museum (UNESCO training centres at Jos in Nigeria, Mexico, and Honolulu; of ICOM at Santiago, Chile, and soon at Kuala Lumpur). It is certain, however, that the aim will be reached only through the integration of museums into the struggle toward sociocultural development, which means their largely forsaking the tradition of the great European museums, the historical evolution of which has taken place in an entirely different context.

In these countries, the profession has begun to organize itself: there are powerful museum associations in Brazil and in India, a Museums Association of Tropical Africa, regional activities initiated by ICOM in Asia and in Latin America. But there is much to do, and the national heritage is impoverished continually, for the very reason that swelling collections and museums in rich countries have encouraged an export drain of such items as African sculptures, Maya stelae, Indian bas-reliefs, and Chinese ceramics.

Since World War II the determining factors of museum development in all countries are the countless meetings, conferences, and seminars organized by professional groups or by UNESCO at the national, regional, and international level. Publications, periodical or otherwise, play an equally essential part. Recently, there has been launched the idea of creating institutes of museology that would permit professionals to think together about institutional and professional problems. All these phenomena, all these movements tend to unify the world of museums and eliminate discrimination in museological matters between developed and developing countries, between the West and the rest of the world. The innovative and progressive activity of Japan, Mexico, and India is significant in this respect because the achievements of these countries now offer examples and models to those where the museum was once born.

PROSPECTIVE

A cursory balance sheet of the situation of the museum in the world of 1970 may be presented. Every time a museum, in accordance with its vocation and with the support of public authority and influential private sources, has given its public leadership and has integrated itself into society and served it, while at the same time drawing upon intensive and systematic scientific research and attending both to the requirements of conservation and those of social communication, it has given proof of its indispensable character and its value as a fundamental institution in the modern world. When, on the contrary, a museum refuses to modify its routines or to accept the revolution in museum concepts, devoting itself to an elite or a basically tourist public, it cuts itself off from the mass of society and from the young. There are countries where the museum's role does not cease to grow, but there are others where, without the intervention of radical transformation, the museum is condemned to vegetate and perhaps even to disappear.

Continuing problems

Certainly, there are obvious obstacles and limitations. Museums that have the greatest success can no longer handle the flow of visitors easily. To set against this, too many of them, and sometimes the same ones, appear as "centralizers," monopolizing a wealth of exhibits in too confined a space at the centre of a metropolis. Thus, in a period of multidisciplinary research and action, their rigorous specialization, their disdain of reality and their contempt for the reasonable requirements of an increasingly well-read and well-informed public lead one to think that they are hardly capable of further adaptation. In fact, nothing in all this is ultimately of grave danger to the museum, so long as it maintains an active and constant pursuit of the revolution that began during the last few decades. The museum, which alone can speak the language of real things, will more and more become an antidote to the culture of the image and the word. If it did not exist, it would be necessary to invent it. In scientific matters, it is the centre of a three-dimensional documentation of the history of the world and of man, which no publication can replace. As an instrument of education and culture, provided that it adopts an integrated vision of science and humanism and provided also that it decentralizes itself physically and intellectually, the museum is a centre of development for all men through their participation in the creation and continuity of culture. Socially, the museum does not accept distinctions of class nor those of race or age. It invites an opening of the spirit, an enduring insight into the problems of yesterday, today, and tomorrow. It is, finally and above all else, an essential element of the cultural environment of man, present and future.

BIBLIOGRAPHY. Few books have been published on museum matters, but a large number of periodicals, national and international, contain articles of widespread museological or museographical interest. Among them are: *Museum* (quarterly); ICOM *News* (quarterly); *Studies in Conservation* (quarterly); *Art and Archaeology Technical Abstracts* (biannual);

Museum News (monthly); *Curator* (quarterly); and *Museums Journal* (quarterly). The major bibliographical works are: *Bibliographie muséologique internationale/International Museological Bibliography* (annual); STEPHAN F. DE BORHEGYI, ELBA A. DODSON, and IRENE A. HANSON (comps.), *A Bibliography of Museums and Museum Work, 1900–1960* (1960, suppl. 1961); FREDERICK A. RATH and MERRILYN ROGERS O'CONNELL (comps.), *Guide to Historic Preservation, Historical Agencies, and Museum Practices: A Selective Bibliography* (1970).

The most recent history of museums is GERMAIN BAZIN, *Le Temps des Musées* (1967; Eng. trans., *The Museum Age*, 1967). A controversial study of museum history with an attempt to show the future developments of the institution may be found in S. DILLON RIPLEY, *The Sacred Grove* (1969). On general organization of museums, see: UNESCO, *The Organization of Museums: Practical Advice* (1960, reprinted 1967); INTERNATIONAL COUNCIL OF MUSEUMS, *Les Problèmes des musées dans les pays en voie de développement rapide* (1964); and G.H. RIVIERE, *Report of the* UNESCO *Regional Seminar on the Educational Role of Museums* (1960). The most comprehensive book on recent museum architecture is MICHAEL BRAWNE, *The New Museum* (1965). On museum techniques for management administration, registration, and acquisition, see the booklets published by the Smithsonian Institution; the Canadian Museums Association, especially G.B. MACBEATH and S.J. GOODING (eds.), *Basic Museum Management* (1969); the British Museums Association; and DOROTHY DUDLEY and IRMA BEZOLD WILKINSON, *Museum Registration Methods*, rev. ed. (1968). On conservation, although the amount of technical literature is extensive, UNESCO, *The Conservation of Cultural Property* (1968), is the most up-to-date work. An effort has been made recently to make available to all curators basic information on important features such as lighting: ROBERT L. FELLER "Control of Deteriorating Effects of Light upon Museum Objects," *Museum*, 17:57–98 (1964). On research and museum collections, the best theoretical study is J. NEUSTUPNY, *Museums and Research* (1968). The largest amount of museum literature concerns education and communications methods. Among many titles: DUNCAN F. CAMERON (ed.), *Are Art Galleries Obsolete?*, a report of the proceedings of Seminar 2, organized by the Art Gallery of Ontario (1969); and "A Viewpoint: The Museum As a Communications System and Implications for Museum Education," *Curator*, 11:33–40 (1968); ERICK LARRABEE (ed.), *Museums and Education* (1968); HANS L. ZETTERBERG, *Museums and Adult Education* (1968); ALMA S. WITTLIN, *Museums: In Search of a Usable Future* (1970).

Training of qualified personnel has recently become a major issue although little has been published yet, except the ICOM, *Training of Museum Personnel* (1970). All countries with a sizable number of museums have published directories, and a complete list of them cannot be given here. Special mention must be given to AMERICAN ASSOCIATION OF MUSEUMS, *The Official Museum Directory* (1971); and to the BRITISH MUSEUMS ASSOCIATION, *Museums Calendar* (annual), including information on countries other than the U.K.

(H.de V.-B)

Music, Art of

Both the simple folk song and the complex electronic composition belong to the same activity, music. Both are humanly engineered; both are conceptual and auditory, and these factors have been present in music of all styles and in all periods of history, Eastern and Western. What follows is an examination of the accounts of music that have been given by philosophers and musicians from earliest times to the present. The historical role of music as an adjunct to other human activities and its gradual emergence as an activity in its own right will be observed, along with the forms through which it found expression. Attention will then be directed to the situation of music in the world today.

Music is an art that, in one guise or another, permeates every human society. Modern music is heard in a bewildering profusion of styles, many of them contemporary, others engendered in past eras. Music is a protean art; it lends itself easily to alliances with words, as in song, and with physical movement, as in dance. Throughout history, music has been an important adjunct to ritual and drama and has been credited with the capacity to reflect and influence human emotion. Popular culture has consistently exploited these possibilities, most conspicuously today by means of radio, film, television, and the musical theatre. The implications of the uses of music in psychotherapy, geriatrics, and advertising testify to a faith in its power to affect human behaviour. Publications and recordings have effectively internationalized music in its most significant, as well as its most trivial, manifestations. Beyond all this, the teaching of music in primary and secondary schools has now attained virtually worldwide acceptance.

But the prevalence of music is nothing new and its human importance has often been acknowledged. What seems curious is that, despite the universality of the art, no one until recent times has argued for its necessity. The ancient Greek philosopher Democritus explicitly denied any fundamental need for music: "For it was not necessity that separated it off, but it arose from the existing superfluity." The view that music and the other arts are mere graces is still widespread, although the growth of psychological understanding of play and other symbolic activities has begun to weaken this tenacious belief.

SPECULATIONS ON THE NATURE OF MUSIC

Music, then, is everywhere to be heard. But what *is* music? Commentators have spoken of "the relationship of music to the human senses and intellect," thus affirming a world of human discourse as the necessary setting for the art. A definition of music itself will take longer. As Aristotle said, "It is not easy to determine the nature of music or why anyone should have a knowledge of it."

Early in the 20th century, it was regarded as a commonplace that a musical tone was characterized by the regularity of its vibrations; this uniformity gave it a fixed pitch and distinguished its sounds from "noise." Although the view may be supported by traditional music, it would be an unacceptable yardstick today, when "noise" itself may be treated as an element in composition, to say nothing of the random sounds incorporated (without prior knowledge of what they will be) by modern composers, like the American John Cage, and others in works having aleatory (chance) or impromptu features. Tone, moreover, is only one component in music, others being rhythm, timbre (tone colour), and texture. Electronic machinery has enabled some composers to create works in which the traditional role of the interpreter is abolished and to record, directly on tape, sounds that were formerly beyond human ability to produce, if not to imagine.

Non-Western ideas. From historical accounts, it is clear that the power to move men has always been attributed to music; its ecstatic possibilities have been recognized in all cultures and have usually been admitted in practice under particular conditions, sometimes stringent ones. In the civilization of India, music was put into the service of religion from earliest times; Vedic hymns stand at the beginning of the record. As the art developed over many centuries into a music of profound melodic and rhythmic intricacy, the discipline of a religious text or the guideline of a story determined the structure. Even today the narrator is central in most performances of Indian music, and the virtuosity of a skillful singer rivals that of the instrumentalists. There is very little concept of vocal or instrumental idiom in the Western sense. The vertical dimension of chord structure—that is, the effects created by sounding tones simultaneously—has never been developed in South Asian music; the divisions of an octave (intervals) are more numerous than in Western music, and melodic complexity in Oriental music goes far beyond that of Western practice. Moreover, an element of improvisation is retained that is vital to the success of a performance. The spontaneous imitation carried on between an instrumentalist and narrator, against the insistent rhythmic subtleties of the drums, can be a source of the greatest excitement, in large measure because of the faithful adherence to the rigid rules governing the rendition of ragas—the ancient melodic patterns of Indian music.

Indian music in religion

Chinese music, like Indian, has traditionally been an adjunct to ceremony or narrative. Confucius (551–479 BC) assigned an important place to music in the service of a well-ordered moral universe. He saw music and government as reflecting one another and believed that only the superior man who can understand music is equipped to

Confucius and music in China

govern. Music, he thought, reveals character through the six emotions it can portray: sorrow, satisfaction, joy, anger, piety, love. According to Confucius, great music is in harmony with the universe, restoring order to the physical world through that harmony. Music, as a true mirror of character, makes pretense or deception impossible.

Ancient Greek ideas. Although music was important in the life of ancient Greece, it is not now known how that music actually sounded. Only a few notated fragments have survived, and no key exists for restoring even these. The Greeks were given to theoretical speculation about music; they had a system of notation, and they "practiced music," as Socrates himself, in a vision, had been enjoined to do. But the Greek term from which the word music is derived was a generic one, referring to any art or science practiced under the aegis of the Muses. Music, therefore, as distinct from gymnastics, was all encompassing. (Much speculation, however, was clearly directed toward that more restricted meaning with which we are familiar.) Music was virtually a department of mathematics for the philosopher Pythagoras (*c.* 550 BC), who was the first musical numerologist, and who laid the foundations for acoustics. In acoustics, the Greeks discovered the correspondence between the pitch of a note and the length of a string. But they did not progress to a calculation of pitch on the basis of vibrations, though an attempt was made to connect sounds with underlying motions.

Plato (428–348/347 BC), like Confucius, looked on music as a department of ethics. And like Confucius he was anxious to regulate the use of particular modes (*i.e.*, arrangements of notes, like scales) because of their supposed effects on men. Plato was a stern musical disciplinarian; he saw a correspondence between the character of a man and the music that represented him. Straightforward simplicity was best. In the *Laws*, Plato declared that rhythmic and melodic complexities were to be avoided because they led to depression and disorder. Music echoes divine harmony; rhythm and melody imitate the movements of heavenly bodies, thus delineating the music of the spheres and reflecting the moral order of the universe. Earthly music, however, is suspect; Plato distrusted its emotional power. Music must therefore be of the right sort; the sensuous qualities of certain modes are dangerous, and a strong censorship must be imposed. Music and gymnastics in the correct balance would constitute the desirable curriculum in education. Plato admitted and valued music in its ethically approved forms; his concern was primarily with the effects of music, and he therefore regarded it as a psychosociological phenomenon.

Aristotle's views

Yet Plato, in treating earthly music as a shadow of the Ideal, saw a symbolic significance in the art. Aristotle carried forward the concept of the art as imitation, but music could express the universal as well. His idea that works of art could contain a measure of truth in themselves—an idea voiced more explicitly by Plotinus in the third century AD—gave added strength to the symbolic view. Aristotle, following Plato, thought that music has power to mold human character, but he would admit all the modes, recognizing happiness and pleasure as values to both the individual and the state. He advocated a rich musical diet. Aristotle made a distinction between those who have only theoretical knowledge and those who produce music, maintaining that persons who do not perform cannot be good judges of the performances of others.

Aristoxenus, a pupil of Aristotle, gave considerable credit to the human listener, his importance and his powers of perception. He denigrated the dominance of mathematical and acoustical considerations. For Aristoxenus, music was emotional and fulfilled a functional role, for which both the hearing and the intellect of the listener were essential. Individual tones were to be understood in their relations to one another and in the context of larger formal units. The Epicureans and Stoics adopted a more naturalistic view of music and its function, which they accepted as an adjunct to the good life. They gave more emphasis to sensation than did Plato, but they nevertheless placed music in the service of moderation and virtue. A dissenting 3rd-century voice was that of Empiricus, who said that music was an art of tones and rhythms only that meant nothing outside itself.

The Platonic influence in musical thought was to be dominant for at least a millennium. Following that period of unquestioned philosophic allegiance, there were times of rededication to Greek concepts, accompanied by reverent and insistent homage (*e.g.*, the group of late 16th-century Florentines, known as the Camerata, who were instrumental in the development of opera). Such returns to simplicity, directness, and the primacy of the word have been made periodically, out of loyalty to Platonic imperatives, however much these "neo" practices may have differed from those of the Greeks themselves.

In the 20th century, the effects of Greek thought are still strongly evident in the belief that music influences the ethical life; in the idea that music can be explained in terms of some component such as number (that may itself be only a reflection of another, higher source); in the view that music has specific effects and functions that can be appropriately labelled; and in the recurrent observation that music is connected with human emotion. In every historical period there have been defectors from one or more of these views, and there are, of course, differences of emphasis.

Christian ideas. Much of the Platonic–Aristotelian teaching, as restated by the Roman philosopher Boethius (*c.* 480–524), was well suited to the needs of the church; the conservative aspects of that philosophy, with its fear of innovation, were conducive to the maintenance of order. The role of music as accessory to words is nowhere more clearly illustrated than in the history of Christianity, where the primacy of the text has always been emphasized and sometimes, as in Roman Catholic doctrine, made an article of faith. In the varieties of plainsong, melody was used for textual illumination; the configurations of sound took their cue from the words. St. Augustine (AD 354–430), who was attracted by music and valued its utility to religion, was fearful of its sensuous element and anxious that the melody never take precedence over the words. These had been Plato's concerns also. Still echoing the Greeks, Augustine, whose beliefs were reiterated by St. Thomas Aquinas (*c.* 1225–*c.* 74), held the basis of music to be mathematical; music reflects celestial movement and order.

Martin Luther (1483–1546) was a musical liberal and reformer. But the uses he envisioned for music, despite his innovations, were in the mainstream of tradition; music must be simple, direct, accessible, an aid to piety. His assignment of particular qualities to a given mode is reminiscent of Plato and Confucius. Calvin (1509–64) took a more cautious and fearful view of music than Luther, warning against voluptuous, effeminate, or disorderly music and insisting upon the supremacy of the text.

Music as mathematics

Later Western ideas. In reviewing the accounts of music that have characterized musical and intellectual history, it is clear that the Pythagoreans are reborn from age to age. The German astronomer Johannes Kepler (1571–1630) perpetuated, in effect, the idea of the harmony of the spheres, attempting to relate music to planetary movement. René Descartes (1596–1650), too, saw the basic of music as mathematical. He was a faithful Platonist in his prescription of temperate rhythms and simple melodies so that music would not produce imaginative, exciting, and hence immoral, effects. For another philosopher–mathematician, the German Gottfried von Leibniz (1646–1716), music reflected a universal rhythm and mirrored a reality that was fundamentally mathematical, to be experienced in the mind as a subconscious (unfelt) apprehension of numerical relationships.

Immanuel Kant (1724–1804) ranked music as lowest in his hierarchy of the arts. What he distrusted most about music was its wordlessness; he considered it useful for enjoyment but negligible in the service of culture. Allied with poetry, however, it may acquire conceptual value. Georg Wilhelm Friedrich Hegel (1770–1831) also extolled the discursive faculties, saying that art, though it expresses the divine, must yield to philosophy. He acknowledged the peculiar power of music to express many

nuances of the emotions. Like Kant, Hegel preferred vocal music to instrumental, deprecating wordless music as subjective and indefinite. The essence of music he held to be rhythm, which finds its counterpart in man's innermost self. What is original in Hegel's view is his claim that music, unlike the other arts, has no independent existence in space, is not "objective" in that sense; the fundamental rhythm of music (again an aspect of number) is experienced within the hearer.

Subjectivism in Hegel

After the 18th century, speculations upon the intrinsic nature of music became more numerous and profound. The elements necessary for a more comprehensive theory of its function and meaning became discernible. But philosophers whose views have been summarized thus far were not speaking as philosophers of music. Music interested them in terms extrinsic to itself, in its observable effects; in its connections with dance, religious ritual, or festive rites; because of its alliance with words; or for some other extramusical consideration. The only common denominator to be found, aside from the recognition of different types of music, is the acknowledgment of its connection with the emotional life; and here, to be sure, is that problematic power of the art to move men. Various extramusical preoccupations are today the raison d'être of "contextualist" explanations of music, which are concerned with its relation to the human environment. The history of music itself is largely an account of its adjunctive function in rituals and ceremonies of all kinds—religious, military, courtly—and in musical theatre. The protean character of music that enables it to form such easy alliances with literature and drama (as in folk song, art song, opera, "background" music) and with the dance (tribal, ethnic, "social," ballet) appears to confirm the wide range and influence that the Greeks assigned to it.

PROBLEMS OF MUSICAL MEANING

Musical theorists

Before the 19th century, musicians themselves seldom were theorists, if theorist is defined as one who explicates meaning. Musical theory, when it was something other than the exposition of a prevalent or emerging style, was likely to be a technical manual guiding vocal or instrumental performance, a set of directions for meeting current exigencies in church or theatre practice, or a missive advocating reforms. Prolific masters, such as Bach, produced not learned treatises but monuments of art.

The 19th century saw the emergence of composer–critics (Carl Maria von Weber, Robert Schumann, Hector Berlioz, Franz Liszt), versatile artists with literary proclivities who were not, to be sure, propounding comprehensive theories or systems of thought. Richard Wagner, an active theorizer, presaged a new species, the composer–author. But he did little to advance musical theory. He proposed a unity of music and drama (*Gesamtkunstwerk*)—a reflection of the "programmatic" preoccupations of 19th-century composers—but its multiplicity of musical and extramusical elements only added to the confusion of musical thought. The distinctly musical character of Wagner's genius, clearly discernible in *The Ring*, is in no way explained by his discursive credos. Igor Stravinsky, Arnold Schoenberg, and other composer–authors of the 20th century were to be somewhat more successful in elucidating their techniques and aims.

Origins of modern theories. Present-day ideas of music as a symbolism owe much to two German philosophers, Arthur Schopenhauer (1788–1860) and Friedrich Nietzsche (1844–1900), who brought to the theory of music a new concept, articulated by each in different ways and in divergent terms but faithful to the same principle—dynamism. Both saw in music an art that is not "spatialized" (hence not "objective") in the way that other arts are by the very conditions of their manifestation. Music is closer to the inner dynamism of process; there are fewer technical (and no concrete) impediments to immediate apprehension, for an entire dimension of the empirical world has been bypassed.

Schopenhauer looked upon Platonic Ideas as objectifying will, but music is

> by no means like the other arts, the copy of the Ideas, but the *copy of the will itself*. This is why the effect of music is so much more powerful and penetrating than that of the other arts, for they speak only of shadows, but it speaks of the thing itself.

In contrast to Kant he accords a special efficacy to music:

> The effect of music is stronger, quicker, more necessary and infallible. Men have practiced music in all ages without being able to account for this; content to understand it directly, they renounce all claim to an abstract conception of this direct understanding itself.

Schopenhauer acknowledged a connection between human feeling and music, which "restores to us all the emotions of our inmost nature, but entirely without reality and far removed from their pain." Music, which he is presenting an as analogue of the emotional life, is a copy or symbol of the will.

Nietzsche posed an Apollonian–Dionysian dichotomy, the former representing form and rationality, and the latter drunkenness and ecstasy. For Nietzsche, music was the Dionysian art *par excellence*. In *The Birth of Tragedy from the Spirit of Music*, Nietzsche anticipated the 20th-century discovery that symbol making (whether in dreams, myth, or art) is a necessary and to some extent even automatic human activity. The rich suggestiveness and prescience of his insights embraced the concept of the "symbolical analogue"—the artistic function of ordering and heightening the ingredients of the actual world—and the polarities of experience symbolized in the Apollonian–Dionysian conflict itself, which Stravinsky also explored. Nietzsche gave short shrift to mathematical aspects of music, and like Schopenhauer he deprecated blatantly programmatic music that abounds in obvious imitations of natural sounds. Discerning a power in music to create myths, he looked upon mere "tone painting" as the antithesis of its essential character.

Conflicting views about meaning. Efforts of theorists to account for the universal appeal of music and to explain its effects have, since the 19th century, been various, contradictory, and highly controversial. In pointing out the chief points of view that have emerged, it must be emphasized that there are no completely isolated categories and there is usually considerable overlapping; a single spokesman, the 19th-century English psychologist Edmund Gurney, for example, may incorporate formalist, symbolist, expressionist, and psychological elements, in varying proportions, to explain the phenomenon of music. Although some disagreements are more apparent than real because of the inherent problems of terminology and definition, diametrically opposing views are also held and tenaciously defended.

Referentialists and Nonreferentialists

Among those who seek and propound theories of musical meaning, the most persistent disagreement is between the referentialists (or "heteronomists"), who hold that music can and does refer to meanings outside itself, and the nonreferentialists (sometimes called formalists or absolutists), who maintain that the art is autonomous and "means itself." The Austrian critic Eduard Hanslick, in his *The Beautiful in Music* (German edition published 1854), was a strong proponent of music as an art of intrinsic principles and ideas; yet even Hanslick, ardent formalist though he was, struggled with the problem of emotion in music. Hanslick's views have been classified as a modified heteronomous theory.

One looks in vain for an extremist of either persuasion, referentialist or nonreferentialist. Stravinsky first achieved fame as a composer of ballet music, and his works throughout his career have been rich in extramusical associations. It would be a nice simplification to ally referentialism with "program" music and nonreferentialism with "absolute." But the problem cannot be resolved by such a choice, if only, first of all, because extramusical referents can vary in complexity from a mere descriptive title to the convolutions of the Wagnerian leitmotiv, in which a particular musical phrase is consistently associated with a particular person, place, or thing. The referentialist does not require an explicit program (which may, when present, be altogether too meagre by his canons), and the nonreferentialist does not necessarily denigrate program music, though he makes a point of distinguishing between the extramusical pro-

gram and the musical meaning. The contemporary U.S. theorist Leonard Meyer, in his *Emotion and Meaning in Music* (1956), speaks of "designative" and "embodied" meanings; he recognizes both kinds in music, but appears to give equal weight to the extrinsic and intrisic.

If there is intrinsic, or embodied meaning, one may well ask what meaning is embodied and how it is to be apprehended. An extreme formalist would say that the acoustic pattern itself and nothing more is the sense of music; Hanslick, indeed, said this, though he did not hold consistently to the view. But most nonreferentialists regard music as, in one way or another, emotionally meaningful or expressive. Referentialists, too, find expressive content in music, though this emotional content may be extramusical (even if not explicit) in origin, according to the American theorists John Hospers in *Meaning and Truth in the Arts* (1946) and Donald Ferguson in *Music as Metaphor* (1960). Meyer has made the observation that while most referentialists are expressionists, not all expressionists are referentialists. He makes the useful distinction between absolute expressionists and referential expressionists and identifies his own position as "formalist-absolute expressionist." In acknowledging that music can and does express referential (designative) meanings as well as nonreferential ones, Meyer exhibits an eclectic and certainly permissive view. But he has been criticized for failing to make clear the *modus operandi* of this referential meaning in music.

Intuition and intellect

Most theorists agree that music is an auditory phenomenon and that hearing is the beginning of understanding. Beyond this there is little agreement. There is bad blood especially between proponents of intuition, like Benedetto Croce, and champions of intellectual cognition, like Hospers. Gurney was constrained to postulate a special musical faculty that need not reside exclusively either in the mind or the heart. The main problem for theorists arises from the inveterate tendency to dichotomize thought and feeling. Henri Bergson (1859–1941) broke with this tradition when he spoke for "an intellectual act of intuition." Recently, a reawakened philosophic and artistic concern for the concept of organic unity has revealed strong affinities among such disparate works as Gurney's *The Power of Sound* (1880); the U.S. philosopher Susanne K. Langer's *Philosophy in a New Key* (1942) and her later works; John Dewey's classic *Art as Experience* (1934); and the U.S. composer Roger Sessions' *The Musical Experience* (1950).

It is apparent that music is connected in some way with the emotional life of man, but the "how" continues to be elusive. Sessions (echoing Aristotle) states the problem fairly:

> No one denies that music arouses emotions, nor do most people deny that the values of music are both qualitatively and quantitatively connected with the emotions it arouses. Yet it is not easy to say just what this connection is.

It was long fashionable to speak of the "language" of music, or of music as the "language of the emotions," but since a precise semantics is wanting in music, the analogy breaks down. Two or more listeners may derive very different "meanings" from the same piece of music, and since written and spoken language cannot render these musical "meanings," whatever they may be, in consistent and commonly recognizable terms, verbal explication often seems to raise more questions than it settles. Philosophic analysts who hold that all meaning is capable of rendition in language therefore pronounce music—unless it can be saved by the referentialists—without meaning, confronting the thoughtful listener, thereby, with a proposition that seems clearly to contradict (and trivialize) his own experience. The difficulty, of course, is a semantic one, and explains why some theorists substitute such terms as import, significance, pattern, or gestalt for meaning. Recognizing an incompatibility between the modalities of nonverbal arts and their treatment by discursive thought, it is hardly surprising that musical aestheticians have been few.

Symbolist contributions

Significant contributions to musical theory have been made in the 20th century by several investigators who may be classified as symbolists, though most of them exhibit formalist, expressionist, and psychological elements as well. The most influential (and controversial) work has been done by Langer. Her most adamant critics (such as John Hospers) have objected to her use of the term symbol that must, in their lexica, stand for something definite; she takes pains to ascribe this more limited usage to the term signal. The more general use of the term symbol that she endorses has a long history, notably in such 19th-century figures as Goethe, Carlyle, and the French Symbolist poets. Mrs. Langer is accused of having somewhat weakened her argument through a vacillating terminology, and she has described the musical symbol as "unconsummated" because of its ambiguity. But the validity of her theory does not depend upon the term symbol; her thought, indeed, has much in common with that of Edmund Gurney, who does not employ the term, and whose "ideal motion," if substituted for symbol, would remove most of her critics' objections. Her use of symbol is nevertheless defensible; she construes art as a "symbolic analogue of emotive life," rendering the "forms of sentient being" into intelligible configurations. She is a naturalist; she sees art as organic in origin and she echoes the view, long held among symbolists, that artistic form and content compose an indissoluble unity that each art manifests according to its peculiar conditions. The symbolism of music is therefore tonal (or, at its broadest, auditory) in character and can be realized only in time; in psychological experience, time assumes an ideal guise. (Painting and sculpture, in their distinctive modalities, embody ideal space.) Mrs. Langer embraces all the arts in her purview. The U.S. musical theorist Gordon Epperson's *The Musical Symbol* (1967) is an application of her concepts, with modifications, intensively to music.

In moving from symbolic to contextualist explanations of music, it is well to note that a source of great confusion, in the former, is the fact that tone painting (with explicit signals that yield, when the code is understood, designative meanings) is widely regarded as musical symbolism. An example of such tone painting is Bach's introduction of musical notes, corresponding to the letters of his own name, as a theme in the unfinished final fugue of the *Art of Fugue*. And surely it may be argued that this qualifies on one level. But the contention that there is an intrinsic symbolism in the musical meaning itself is a claim that referentialists are generally unwilling to honour. Yet many theorists, whose concern is with the sociological or psychological effects of music, are not so much opposed to the idea of inner or profound meaning as indifferent to such meaning per se. Even an absolutist, however, is unable to examine music in isolation from its human environment. Meyer deliberately eschews logical and philosophical problems of music and makes "no attempt to decide whether music is a language or whether musical stimuli are signs or symbols." (He does not defend the inference that such concerns are irrelevant to meaning.) Musical meaning and communication, he maintains, cannot exist in the absence of the cultural context. The statement is hardly open to dispute; a theorist is classified according to his proximity to the referential or nonreferential pole. If a referentialist emphasizes explicit aims and associations of a particular work (as in varieties of *Gebrauchsmusik* or "utility" music, written for specific social or educational purposes), the formalist can maintain that there is also an intrinsic or embodied meaning to which he attaches the greater aesthetic value.

Psychological approaches

Among contextualists, however, a simple referential view is the exception rather than the rule. Any theorist who examines musical perception is making a study of a complex human activity. He is dealing with the psychology of music, in which certain elements—*e.g.*, music, listener, mode of apprehension, cultural context—are indispensable, and in which characteristic processes recur. Specialists will emphasize one element or another: formalists the music itself, sociologists the listener and his milieu, psychologists the how of perception. Though psychology could survey the whole field, in practice psychologists, according to their persuasions, investigate the perception of measurable acoustic phenomena, the physical–mental

effects of musical sound, or—more rarely—the functional role of music in human experience; and pragmatists and analysts alike may leave something out of account. But it remains for the comprehensive theorist, probably one who, like Langer, is equipped to discern relationships among many departments of thought, to construct a valid hierarchical structure of musical meaning in all its ramifications.

Deryck Cooke, the British musicologist and the author of *The Language of Music* (1959), who may be classified as a referential expressionist, has offered a sophisticated argument for the notion of music as language. Concepts, however, may not be rendered by this language, only feelings. Cooke reaffirms the possibility, long disputed by many theorists, that such feelings may be recognized, identified, and even classified. But he confines his investigation to the last few hundred years of the Western tradition.

Information theory

The French theorist Abraham Moles's *Information Theory and Esthetic Perception* (1966), brings the new science of information theory to bear on musical perception, emphasizing that the concept of form is the essential thing; the "sonic message," whose dimensions vary from one composition to another, is a whole. Information theory thus proves to be a novel ally for organicists. The message, which is subjected to "atomistic" study of its components is (thanks to recording) concrete; there is a temporal sonic material, a *materia musica.* Moles gives reinforcement to the aesthetic theory of distance:

> The esthetic procedure of isolating sonic objects is analogous to the sculptor's or decorator's isolating a marble work against a black velvet draping: This procedure directs attention to it, alone and not as one element among many in a complex framework.

Information theory, which Leonard Meyer also discusses, begins its investigations without the help of traditional theory, which it finds to be untenable for its procedures. Musical messages discerned through information theory are not referential, yet Moles chooses to term the measurable elements in the sonic repertoire symbols: "each definable temporal stage represents a 'symbol' analogous to a phoneme in language." According to Moles, music must, as an art, obey rules; the role of aesthetics is to enumerate universally valid rules, not to perpetuate the arbitrary or merely traditional. He foresees experimentation with a much richer repertory of sounds, transcending musical instruments and drawing on whatever sources—certainly electronic ones—are available for realizing the "most general orchestra." A host of composers have set out to fulfill this desideratum. In order to increase the compass of possible sounds, various electronic "synthesizers" were constructed. In electronically synthesized music, the medium itself is indistinguishable from its message.

The quest for some distillation of musical meaning may be foredoomed to failure. Meanings, intrinsic and extrinsic, abound; meanings of all kinds, moreover, are revealed in and through the social setting. Church, theatre, and broadcasting affect music in characteristic ways. The modern concert is a device whereby formal, autonomous meanings are emphasized; further, the scope and available repertoire of the concert have been enormously increased through recordings, for any suitably equipped room may become, at the turn of a switch, a recital hall.

A number of important psychological and sociological considerations, closely related to problems of musical meaning, remain. These are discussed, in the context of 20th-century practice, in the following section.

Considerations related to performance practice. Listening to music for its own sake, apart from ritual or storytelling, is a recent historical development. There have always been impromptu song and dance; and performances of music in home, church, and theatrical productions have a long history. But there was no public opera house until 1637 when the first one opened in Venice; the first public concerts for which admission was charged appeared in London in 1672. During the next 50 years there were beginnings in Germany and France also, but the modern concert was not a significant feature of musical life until the late 18th century.

Of the forms that have characterized distinct periods of musical history, it is sufficient to remark here that the chief Renaissance forms—mass, motet, the polyphonic chanson, and madrigal—were allied to texts that strongly influenced their structure. Instrumental music was for the most part in the service of the voice, though instrumental church compositions, dances, and chansons arranged for organ were not uncommon. A strong alliance between voices and instruments has continued into the present, with musical theatre, the art song, and religious music. Instrumental music as a separate genre emerged in the 16th century, gaining considerable momentum in the 17th through a variety of idiomatic pieces. Increased attention to technical fluency was accompanied by greater complexity and sophistication in the instruments themselves. In response to stylistic demands for greater resonance and power, the modern forms of the violin appeared in the late 16th century, only gradually supplanting the earlier viols. The harpsichord did not finally yield to the pianoforte until the 18th century. The once-prevalent idea that early stringed and keyboard instruments were primitive precursors of their modern counterparts has been effectually demolished by research in medieval and Renaissance music and by dedicated performers, who seek to restore the sounds and spirit of those eras.

The development of opera, oratorio, and the cantata gave a prominence to vocal music throughout the Baroque era (*c.* 1600–1750) that made it equal in importance to instrumental music, with which these forms were closely allied. But instrumental-chamber and independent-orchestral ensembles, as they exist today, also had their beginnings during this period. A highly significant development of the late 18th century was the definitive appearance of the modern sonata (whether in the form of the solo and duo sonata, piano trio, string quartet, concerto, or symphony) with the Viennese classicists Haydn and Mozart and, later, Beethoven.

Since a vocal text is likely to be confused with intrinsic musical meaning, or at least to divert attention from a preoccupation with it, it is not surprising that modern aesthetic theory followed on the emergence of an autonomous instrumental music requiring greater concentration on the sound itself; its colour and intensity, and the intelligibility (in terms of tonal organization alone) of a composition. Moreover, the very concept of listening as an attentive (and sometimes rigorous), serious, and necessary activity of the music lover gained acceptance only slowly, following the inauguration of public concerts, and is still vigorously resisted. The expectation that the art should provide enjoyment without effort is, indeed, widespread and accounts for much of the opposition to new and demanding idioms. But even for the well-disciplined and eager listener there is the problem of quantity: he must cope as best he can with what Langer has called "the madhouse of too much art." If more effort is required, more discrimination is also needed. In music education, articulate voices ask that teaching be centred more upon qualitative aspects of the art ("aesthetic education"), less upon music making as an activity. This concern for musical value appears to reflect a more intensive search for meaning, which is not likely to be the exclusive property of a particular style or era; nor is it to be sought in an indiscriminate acceptance or rejection of novelty per se. A pronounced pedagogical interest developed in such contemporary popular music as "rock," "soul," and similar folk idioms with great numbers of followers, especially among the young, whose gigantic festivals generated feelings of religious exaltation. The texts of the songs are highly emotional and often deal with "protest" themes; accompaniments are provided by guitars, keyboard and percussion instruments, and are electronically amplified. Music educators became attracted by the intrinsic structural values of this music, especially its distinctive rhythmic and modal characteristics, its texts, and the qualitative levels that may be distinguished. A music so vital and widespread, moreover, was deemed by many to be worth studying in school. The "rock" movement became a musical-sociological phenomenon of large proportions.

Implied world views

Again, music proves its protean susceptibilities in the service of disparate world views. Among humanist psychologists (such as the Americans Gordon Allport and Abraham Maslow) music may be one among other means toward self-fulfillment, integration, self-actualization; for aesthetic Existentialists (such as the philosopher Jean-Paul Sartre) it is yet another crucial department of choice and freedom; for spiritual Existentialists (such as the philosophers Karl Jaspers and Martin Buber) it transmits transcendent overtones. For expressionists (such as the composers Schoenberg, Ernst Krenek, and René Leibowitz) music carries austere, and sometimes doctrinaire, moral imperatives. Theodor Adorno, a composer-philosopher and pupil of Alban Berg, writes powerfully of these and speaks for an awareness of dazzling lucidity, but the tone, notwithstanding his humour, is one of obligation. Only the expressionists, among those mentioned here, are committed primarily to music, though Adorno, in particular, considers music and musicians always in interaction with their environments. The aesthetic concept of play is virtually absent, except among such humanists as Maslow. With Sartre, no less a humanist, the tone is one of responsibility. Many educators long held the explicit aim (at least in part because of a misinterpretation of John Dewey) of presenting the content of a discipline as "fun"; the present concern for aesthetic education, an area of great interest to Dewey himself, eschews this trivial view. But play, in the aesthetic sense, follows rules, as information theory has demonstrated; even "controlled aleatory" composition observes some limits. And the play may be very serious indeed, as in the important 20th-century atonal style known as "12-tone technique," practiced by the Viennese expressionists and their successors.

Tonality and meaning

The most troublesome problem not only for the untutored listener but also for the professional musician has been, in much contemporary work, the loss of explicit tonality; and this accounts for the tardy popular response to Schoenberg and his school: the vocabulary is esoteric. Nineteenth-century compositions did, indeed, stretch the tonal system to its outer limits; but it is now clear that Wagner and Richard Strauss, and even the early Schoenberg, had not broken from it. As for Debussy, his use of "exoticisms" was filigree upon a secure tonal base. So were such practices as the juxtaposition of keys by Stravinsky. This is not to say that the tonality of the Western world, fecund though it has been, is superior or more natural than other systems. Ethnomusicology and comparative musicology have proved this to be a parochial view, though there are still those who champion harmonic practices based on the physical laws governing overtones—as Western tonality is—as the only "natural" source of development. It should not be irrelevant, however, to inquire if any folk music has been discovered that exhibits atonal characteristics.

Tonality in Western music, though a significant aspect, cannot be considered the crux of musical meaning. The tone-rows that are used in the 12-tone compositions of Schoenberg, like major–minor tonality in earlier music, are a technical substratum and must be no more explicit in the finished work than the chemical makeup of pigments in the "Mona Lisa." The devices selected may affect the comprehensibility or accessibility of the work, but they are not, per se, the determinants of its worth or quality. Similarly with musical colour, or timbre; the 19th century produced a great profusion of compositions, particularly in the orchestral repertoire (*e.g.*, works by Liszt and Berlioz) that exploited the unique sonorities of instruments; control of volume was, in itself, a rich source of colour. Works with literary or other extramusical associations were excellent vehicles for sonorous effects, but colour, like tonality, must be evaluated in musical context. Langer, among present-day aestheticians, regards words themselves as musical, rather than discursive, ingredients; they are "assimilated" by the song.

MUSIC TODAY

Music and the mass media. In the same way that recordings extended the concert, television and film enlarged the scope of musical theatre and the dance. Experiment characterized both media, though film makers became more enterprising. Television, being the later arrival, borrowed from its predecessor, and its own unique possibilities remained largely undeveloped. In the occasional "live" (though more often taped) concert or musical film, the medium functioned as a large theatre. In England and Canada, subsidized broadcasting made possible a comparatively large musical offering. But in the United States, television's tie with commercial advertising produced meagre fare, though some promise could be seen in developments in educational television and university broadcasting centres.

Creative ferment lagged in the musical theatre despite the success of works by Leonard Bernstein (*Fancy Free, On the Town*)—with their emphasis on ballet—and by Gian Carlo Menotti, which demonstrated the popular appeal of original and sophisticated composers and loosened the rigid distinctions once thought to prevail between "concert" and "commercial" music. Except in European countries, with a vigorous tradition or government support or both, the enormous expense of staging opera and ballet in metropolitan centres discouraged not only experiment but even new repertory, except in relatively obscure theatres and on college campuses. The belated fanfare for Alban Berg's opera *Wozzeck*, which was completed in 1921 but not extensively performed until mid-20th century, only emphasized the difficulties experienced by radical new works. Evidences of operatic vitality and imagination such as Germany exhibited—in repertory, staging, choreography—became important, even if not global, developments.

Music in films

In film making there was more widespread evidence of musical vitality, as well as public response. Here the Wagnerian ideal of the music-drama in which all the arts reach their apotheosis, was proclaimed by some practitioners, who held, as Wagner had, that the arts of music, dance, and theatre as separate entities would be absorbed and superseded in the new form. If this proved so, Wagner's Bayreuth would be of much greater compass and far more fluid than the one he envisioned. It seemed more probable, however, that the emerging art film, in which such unity is demonstrated, would achieve a place of its own, while in no way dispossessing other genres. Many British and American composers wrote music for commercial films, notably William Walton (*Henry V*, 1944; *Hamlet*, 1947; *Richard III*, 1954—all featuring Olivier); Virgil Thomson (*The Plow That Broke the Plains*, 1936; *Louisiana Story*, 1948); and Aaron Copland (*Our Town, Of Mice and Men*, 1939). Among Soviet composers Shostakovich scored music for *Hamlet*, and Prokofiev for *Alexander Nevsky* (1938) and *Ivan the Terrible* (1944), both directed by Sergey Eisenstein.

A more or less obtrusive "background music," to be sure, accompanied most commercial films; what Eisenstein envisaged (and what present-day art films strive to achieve) was "a unity of fused musical and visual images, *composing the work with a united audio-visual image.*" Ingmar Bergman, in whose films such a unity of sound and sight was often attained, believed that music and film have much in common because both affect the emotions directly, and are not mediated by the intellect: "Film is mainly rhythm; it is inhalation and exhalation in continuous sequence." Many observed that music and film are both arts of time and movement, with myriad possibilities of unification.

Recording, television, and film have their own potentialities as media; all are devices, as well, for extending the range of concert hall and theatre. The global communication that now takes place with such rapidity by means of these media weakened those cultural barriers, which long isolation imposed on human societies. East Indian and African music, and art, the Japanese Nō Theatre, the sounds of medieval viols—all these and much more were placed within reach of anyone who would take a little trouble. What further revelations lie ahead?

Future prospects

Leonard Meyer, in *Music, the Arts, and Ideas* (1967), suggested that a great many kinds of art and music would continue to exhibit vitality and even flourish, side by side,

in fluctuating equilibrium. For historical precedent, Meyer pointed to ancient Egypt, where a cultural "steady state," favourable to art but without radical innovation, endured for millennia. He thought that further experiment would follow avenues already opened up and that radical novelty in the media of art could hardly be expected after what has transpired in recent years. This was a recognition of possible "outer limits" of the human physiology and psyche, but imposed no moratorium on novelty.

It has happened again and again that music at first considered revolutionary is later regarded as conventional or passé. "The sated ear forever demands new harmonies, which it will fast learn to tolerate and then dismiss as threadbare and uninteresting," said the British physicist James Jeans (1877–1946). Debussy was once "modern" and Stravinsky avant-garde; later even Bartók was regarded as old hat. Impressionism (Debussy, Ravel) and Expressionism (Schoenberg, Berg, Webern) were contemporary manifestations of the early 20th century, succeeded by the Neoclassicism of Stravinsky and Hindemith in the third and fourth decades. Neoclassicism, stylish before 1940, was itself one illustration of the eclecticism that permeated 20th-century composition. The use of jazz idioms by Stravinsky and Hindemith (as well as Copland and many others) was another. After World War II, there was a belated upsurge of Expressionism; occurring not so much in its original spirit, however, as in the preoccupation with 12-tone technique as a compositional device. This was followed by experiments with other aspects of serialism (described as "total" in some contexts) and the bewildering proliferation of new idioms in which electronic music in its various manifestations played an ever more dominant role. The British critic Donald Francis Tovey foresaw this development with extraordinary prescience; but his confident prediction that the human voice (and its echo in the singing instrument) would never be superseded may be open to question.

But if highly abstract and perhaps dehumanized art music—conceived and consummated electronically—is already in process of realization, it need not be concluded that other modes of composition will or must therefore be replaced or invalidated. Though a prominent composer of electronic music may say that Bartók has had his day, it is also a fact that there are millions of listeners who have not "caught up" with Bartók. Many music lovers today, even among professional musicians, find the last Beethoven string quartets hard going. The same composer may be expected to produce works that differ in their dimensions and accessibility. Beethoven, for example, wrote easy short pieces for piano, as well as lengthy sonatas; and Bartók composed not only six demanding string quartets but also the carefully graded *Mikrokosmos,* designed for the instruction of beginning pianists.

Although a particular musical style, such as Impressionism, may reach its peak and cease to evolve, a change of fashion—to Neoclassicism or whatever—does not consign Debussy to the ash heap. The concert-hall repertory is testimony to the persistence of disparate styles; and it is evidence, what is more, of a preference for the old over the new. Public taste in classical music is conservative. There are, however, fluctuating allegiances in regard to established repertoire, as may be seen in the shifts in taste affecting a composer such as Liszt.

Folk music, popular music, jazz. The persistence of viable music is especially remarkable in the universal phenomenon of folk music. Essentially anonymous in character, folk music, transmitted orally, is often a repository of ethnic and cultural idiosyncrasies and affords them a precarious immortality. With the ending of the cultural isolation that geography for so long has imposed, whole traditions are now threatened with extinction. Meanwhile, comparative musicologists are busy recording those traditions, as Bartók and Zoltán Kodály did in their extensive collection of Hungarian folk music. Precisely because recording instruments were developed only recently, the origins and early characteristics of folk music, both vocal and instrumental, are largely conjectural. But present practice, in this domain, may be a good, if not altogether reliable, indication of what preceded it. There is a problem, however, in keeping genres distinct, since "folk" and "art" music have greatly influenced one another. In more than one instance, a composed song has found its way into the vernacular; and it is necessary only to recall a number of Schubert's songs to realize how slight the distinction between folk and art music can be.

Reciprocal influences

Authentic folk music is characterized by melodic and rhythmic directness and simplicity—a compass that will not tax the normal voice—and accent and repetition corresponding closely to indigenous speech patterns; all qualities that make folk tunes easily transferable from voices to instruments. But composed music may also (and often does) exhibit these characteristics. In the 19th century there was much intentional exploitation of native music (a symptom, not infrequently, of nationalistic loyalty) by such composers as the Bohemians Dvořák and Smetana; the Russian "Five" (Balakirev, Cui, Rimsky-Korsakov, Borodin, and Mussorgsky); and Chopin in his polonaises and mazurkas. But this was more often an aspect of eclectic skill and cosmopolitanism than of patriotic sectarianism. Dvořák, for example, wrote many melodies in the style of American Indian and Negro music as he understood them. German composers created melodies of unmistakable folk quality and sometimes made use of well-known tunes not only from their own folk tradition but from other countries also, as Brahms bears witness to in his Hungarian dances, and Beethoven in the "Russian theme" of his string quartet (*Razumovsky*) Opus 59, No. 2. Similar examples abound. Folk music has been a source of nourishment and inspiration for art music and its composers. It is a truism that every artist draws sustenance from his own culture; some composers, notably the Hungarians Bartók and Kodály, absorb and reflect regional musical idioms to an extraordinary degree in their work.

But if folk music finds its way into art music, the converse is also true. Every civilization in which music is highly developed exhibits sophisticated elements (structural, melodic, rhythmic) in its folk music. This has made it difficult to establish the authenticity of much European folk music, and many "examples" are undoubtedly of recent origin. A high degree of musical literacy in a culture influences every level of music making. In the United States, for example, the growing sophistication of "popular" music (and especially jazz) has been effected both by knowledgeable and imaginative practitioners and by a listening public made ready for their innovations.

In a broad sense, folk music best deserves the epithet "popular." But in common usage the term denotes a more ephemeral, commercially successful music; introduced, in many cases, in stage or film musical comedy, or by a reigning vocalist through nightclub and television performances or recordings, and disseminated by the varied and far-flung resources of the entertainment industry. The quality of much of this music is very high. Composers, arrangers, and performers devote considerable talents to its production, often combining in one person two or even three of these functions in a manner reminiscent of "serious" composers, until a fairly strict division of labour (though only within the past century) became the rule. The cohesive force in all this has been the dance band in its varied manifestations; "hit tunes" originate and are perpetuated in that setting.

Popular music, so identified, does represent a viable, though short-lived, folk music throughout much of the civilized world. But few "hits" hold their own for more than a few months against the relentless production of new tunes. Some of the best ones are revived, periodically, out of nostalgia or simply from the quest for variety; when more than a few years old, however, they sound "dated," though they may be highly evocative of another era and possessed of an archaic charm. Newness is one of the principal values of the genre. Such "all-time" favourites as "Dinah," "Stardust," and the "St. Louis Blues" only underscore this point.

Every human calling has its peculiar vocabulary. The speech of jazzmen is highly esoteric, with nuances that cannot be rendered in any terms whatsoever to outsiders.

Meanings, moreover, are often seasonal, and some words change their meaning altogether or become obsolete. Jazz itself defies definition. Nevertheless, some of the elements that have gone into jazz can be identified; they are largely folk elements, indigenous to the United States. Because the Negro contribution to jazz has been enormous, the appellation Afro-American has often been used; but the presumption that jazz is a genetic or ethnic phenomenon is a big one. African melodic patterns found their way, certainly, into plantation slave songs, spirituals, and the distinctive and emotionally fervent blues. As a cultural phenomenon, jazz, in its origins and in its evolution in the 20th century, is uniquely American, though its dissemination has been worldwide. The characteristic syncopation, the undeviating beat, and frenetic energy were constituents of ragtime, an important "jazz" manifestation early in the 20th century. But harmonic accompaniment is essentially European. Jazz was a hybrid art from the start; in the decades since its emergence, it has become increasingly complex. Two distinctive features have been present: the element of improvisation—at one time every skilled musician's stock-in-trade—which has otherwise virtually disappeared from Western musical practice; and, related to improvisation, melodic and rhythmic rubato, or fluctuation, over a steady beat.

Elements of jazz

New Orleans was the first important jazz capital; "Dixieland" and "Chicago style" were early variations of New Orleans jazz. Swing, the predominant popular style of the 1930s and early 1940s, was an aspect of jazz, though purists might wish to disclaim it. (Jazz influence, nevertheless, has been pervasive in all 20th-century popular music, not excluding "rock" and "soul.") Later decades saw the development of "hot" jazz, "cool" jazz, and the growth of a serious audience of intellectual analysts with a special jargon and stringent (though unofficial) requirements for membership in their ingroup. Jazz, like music itself, is exceedingly protean.

One of the most exciting manifestations among jazzmen is the jam session, in which a group of players improvise together, sometimes for hours at a stretch. At its most successful, the jam session produces that happy fusion—rare in modern music making—of rigorous discipline and spontaneous—even inspired—creativity. The complexities of such performance cannot be rendered in notation; nor are they repeatable.

Non-Western counterparts

The foregoing description of the jam session can apply, with little modification, to renditions of East Indian ragas —music adhering to strict traditional melodic patterns; there, too, skillful improvisation is essential to good performance, within prescribed limits. Vocalist, percussionist, and instrumentalists resemble the makeup of the "dance band"; and the lengths of performances (and the Dionysian atmosphere generated) are often strikingly similar. Another interesting parallel is seen in the capacity of jazz instrumentalists to suggest East Asian microtonality, using intervals smaller than half tones.

This interesting similarity between American jazz and East Indian "classical" practice, made even more striking by the rhythmic complexity of both, need not be given exaggerated importance, though it calls for a reappraisal of concepts of "new" and "old." Cultural exchange may be expected, among its other benefits, to effect just this kind of examination. The appetite in the West for East Asian (and other "exotic") music is a recognition of its intrinsic value. Of course, the influence is reciprocal; and the use of Western musical techniques, particularly in film scores and broadcasting, is steadily increasing; India and Japan offer dramatic evidence of this. East meets West, musically speaking, and both are changed. The rapprochement, moreover, proceeds on several levels, not only on the people to people, or "folk" level, but also among ethnomusicologists, a less numerous and more highly disciplined breed who ply their researches across the continents and are jealous in the preservation of traditional patrimonies. In between are those musicians (and active listeners) who are gradually becoming aware of a confrontation of cultures: sometimes dramatic, as in the musical dialogue that ensued when the Indian sitar player Ravi Shankar and the violinist Yehudi Menuhin met, but more often in the form of a subtle infusion of newly published works, recordings, public concerts, and the unrehearsed music of the streets.

BIBLIOGRAPHY

Historical purview: JULIUS PORTNOY, *The Philosopher and Music* (1954), gives a chronological account of what philosophers have contributed to musical theory; and GORDON EPPERSON, *The Musical Symbol*, ch. 2 (1967), discusses philosophical musical thought from classical antiquity through the 18th century. LIN YUTANG (ed. and trans.), *The Wisdom of Confucius* (1938, reprinted 1958), gives attention to ancient Chinese music.

Problems of musical meaning: ARTHUR SCHOPENHAUER, *Die Welt als Wille und Vorstellung* (1883; Eng. trans., *The World as Will and Idea*, 1961); and FRIEDRICH NIETZSCHE, "The Birth of Tragedy from the Spirit of Music," trans. by CLIFTON P. FADIMAN in *The Philosophy of Nietzsche* (1954). EDUARD HANSLICK, *Vom musikalisch Schönen* (1854; Eng. trans., *The Beautiful in Music*, 1957), remains the best single exposition of the formalist (or nonreferentialist) position in musical aesthetics. EDMUND GURNEY, *The Power of Sound* (1880, reprinted 1966), maintains a similar point of view but with considerably greater amplitude and subtlety. For background of the contemporary symbolist views of musical meaning, see ALFRED NORTH WHITEHEAD, *Symbolism* (1959); SUSANNE K. LANGER, "On Significance in Music," in *Philosophy in a New Key*, 2nd ed. (1951), and *Feeling and Form* (1953). LEONARD B. MEYER has made an important contribution to the aesthetics of music. His interest in the relevance of information theory to music has been evidenced in two articles: "Meaning in Music and Information Theory," *Journal of Aesthetics and Art Criticism*, 14:412–424 (1957), and "Some Remarks on Value and Greatness in Music," *ibid.*, 17:486–500 (1959), reprinted in his *Music, The Arts, and Ideas: Patterns and Predictions in Twentieth-Century Culture* (1967). JOHN DEWEY, *Art as Experience* (1934, reprinted 1959); and KARL JASPERS, *Von der Wahrheit* (1947; Eng. trans., *Truth and Symbol*, 1959), have given reinforcement to organic and symbolic theses, respectively.

Performance practice, styles, and musical forms: The best historical accounts of musical forms, styles, and performance practice are to be found in DONALD J. GROUT, *A History of Western Music* (1960); GUSTAVE REESE, *Music in the Middle Ages* (1940), and *Music in the Renaissance*, rev. ed. (1959); MANFRED F. BUKOFZER, *Music in the Baroque Era* (1947); ALFRED EINSTEIN, *Music in the Romantic Era* (1947); and WILLIAM W. AUSTIN, *Music in the 20th Century, from Debussy to Stravinsky* (1966). SIR DONALD FRANCIS TOVEY, *The Forms of Music* (1956), contains informative and engaging short pieces. ROBERT SCHUMANN, *On Music and Musicians*, ed. by KONRAD WOLFF (Eng. trans. 1947), is an example of the work by a 19th-century precursor of the phenomenon of the present-day composer-authors who have contributed to aesthetic theory by elucidating their own works and commentating on other composers and on the scene in general. See also IGOR STRAVINSKY, *Poetics of Music in the Form of Six Lessons* (1947); PAUL HINDEMITH, *A Composer's World* (1952); AARON COPLAND, *Music and Imagination* (1952). Discussion of music and film may be found in LEWIS JACOBS (ed.), *The Emergence of Film Art* (1969). Twelve-tone technique and varieties of serialism deriving from it are treated in ARNOLD SCHOENBERG, *Style and Idea* (1950); and RENE LEIBOWITZ, *Schoenberg, et son école* (1947; Eng. trans., *Schoenberg and His School*, 1949). Short pieces on electronic music appear often in periodical literature. See LEJAREN HILLER and JAMES BEAUCHAMP, "Research in Music with Electronics," *Science*, 150:161–169 (1965); CARLOS CHAVEZ, *Toward a New Music* (Eng. trans. 1937), anticipated by several decades the electronic revolution that is now in progress. On folk music, a classic work is BELA BARTOK, *Amagyar Népdal* (1924; Eng. trans., *Hungarian Folk Music*, 1931). Jazz is treated in LEONARD FEATHER, *The Book of Jazz* (1957); and in LEROY OSTRANSKY, *The Anatomy of Jazz* (1960), a balanced and comprehensive analysis of the subject.

(G.E.)

Music, East Asian

A study of the music of East Asia covers historical periods of changing styles from at least 2,000 BC to the present time, set in a context of regional cultures located as far south and west as Vietnam, Tibet, and Central Asia, north to Mongolia, Manchuria (Northeast Provinces), and Korea, and east to Japan. From this rich and varied offering the present article will centre primarily on the music of three countries: China, Korea, and Japan. After

a general introduction each country will be discussed separately in chronological order. Historical relations and musical similarities will be pointed out along with striking musical differences. This article is divided into the following sections:

I. The nature of East Asian music
  East Asian music vis-à-vis that of other major cultures
    Concepts of music
    Theoretical systems
  Musical traits common to East Asian cultures
II. The music of China
  Formative period
    Ancient artifacts and writings
    Aesthetic principles and extramusical associations
    Tonal system and its theoretical rationalization
    Classification of instruments
    Han dynasty: musical events and foreign influences
  T'ang dynasty
    Thriving of foreign styles
    Courtly music
  Sung and Yüan dynasties
    Consolidation of earlier trends
    Music theatre
  Ming and Ch'ing dynasties
    Further development of opera
    Other vocal and instrumental genres
  Developments since 1911
    Period of the Republic of China and Sino-Japanese War
    Communist period
III. The music of Korea
  Shaman music
  Court music
  Vocal music
IV. The music of Japan
  Music before and through the Nara period
    Early evidences
    Predominant musical traits
    The Nara period
  The Heian period
    Music of the left and of the right
    Musical notation
    Tonal system
    Vocal music
    Shintō music
    Buddhist music
  Kamakura, Muromachi, and Edo periods
    Nō music
    Koto music
    Schools of shakuhachi flute music
    Samisen music
    Kabuki theatre
    Biwa, vocal, and folk music
  The Meiji period and subsequent music
    Sources of Western influence
    Traditional styles
    Composers in Western styles

I. The nature of East Asian music

EAST ASIAN MUSIC VIS-A-VIS THAT OF OTHER MAJOR CULTURES

East Asian music can be viewed as one of the big four among the generally urban, literate cultural groups of the world. The other three are South Asia, the Near East, and Europe. Around each of these major regional cultures one can find many satellite musical systems known as national forms. In most cases, the fundamental musical concepts of such national forms reflect the basic ideals of the cultural core. For example, the musics of Iran and Egypt are of one family, as are those of France and Sweden or of China and Japan. A possible fifth addition to the "big four" concept is the Southeast Asian musical culture characterized by the use of knobbed gongs. Although it lacks the extensive written documents on music history and theory characteristic of the other four, it shares with them certain distinctive types of instruments and musical styles. Using instrument type alone as a measure, it is sometimes possible to note cultural influences and mixtures of the major traditions in smaller units. For example, the physical structure and playing positions of various bowed instruments in mainland Southeast Asia can often mark clearly Chinese influence, as in Vietnam, or Muslim and Chinese forms in confluence, as in the various bowed lutes of courtly ensembles in Cambodia and Thailand. By the same token, the appearance of flat gongs in mainland Southeast Asia shows Chinese connections, while the knobbed gongs clearly stem from Southeast Asian culture proper.

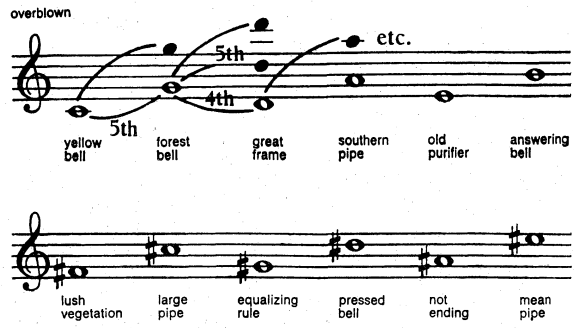
Intercultural influences

Concepts of music. If one turns to distinctions in musical style one of the first questions to arise is "What is music?" Two basic definitions will suffice for this article. The first definition is cultural: a sonic event can be called music if the people who use it call it music, regardless of one's own reaction to it. Similarly, certain events that sound musical to foreign ears are not music culturally if they are not accepted as such by native culture carriers. A good example of such a situation is found in the Near East, where singing is never allowed in the mosque, though one may hear performances and even buy records of "readings" from the Qur'ān. Such cultural and functional problems of definition seldom arise in East Asian music, and a more neutral definition is appropriate. A sound event may be considered and studied as music if it combines the elements of pitch, rhythm, and loudness in such a way that they communicate emotionally, aesthetically, or functionally on the levels that either transcend or are unrelated to speech communication. Those who have been moved by a love song or a lament can well appreciate some of the implications of such a view of music. When listening to "exotic" music—*i.e.*, that of a tradition outside one's own background—it is important to remember that such transcendental values are at work for the alien listener as well as for listeners familiar with the particular musical language in use.

There are many kinds of music in the world, the three most common terms being folk, popular, and art music. Folk and popular music have their special indigenous and mixed forms in Asia (as in all the world today), but it is in the literate art traditions of Asia that historical and musical distinctions can be made most clearly. In the context of this discussion, art music is defined as a tradition having, to some degree, a conscious theoretical basis and a sense of repertoire that is played against the highest standards of excellence held by informed native listeners. The performer is often—though not always—a professional, and there may be a known historical depth to the traditions. Thus, there may be art music in many nonliterate cultures such as that of the Australian Aboriginals and that of the tribal courts of Africa. In this article, however, the major concern is with one of the large urban, literate cultures and its three national variants. Before looking at these musical systems in detail, it is useful to compare the entire culture with those of the other major "big" three, South Asia, the Near East, and Europe.

Theoretical systems. It has already been implied that all four major literate cultures have written histories and music theories. All four, in their ancient forms, laid a strong emphasis on the extramusical qualities of music. For example, the study of such concepts as the power of vibrations (in ancient Indian music theory), and the relationships between music and other elements in the universe (found in Assyrian records as well as in the writings of medieval European scholars) can be matched in East Asia by the joint efforts of Chinese musicologists and astrologers to bring the music of the empire in tune with the universe. In addition, all four developed mathematically and acoustically based music theories. The Chinese system used end-blown bamboo tubes (*lü*) as the source of pitch data, whereas the other three systems tended to rely on the divisions of a string (thus the pictures of Pope Gregory I or monkish European scholars strumming a monochord). Although their Chinese equivalents are pictured with panpipes instead, one of the most fascinating facts is that both Chinese and Western systems, unlike the other two, emerged with a cycle of fifths (the fifth being a musical interval such as C to G in the West; see also below under *The music of China*), creating 12 tones that were mathematically correct (in their acoustical relationships) but not equidistant musically (shown in the musical notation labelled I). Thus it would seem that the two geographical ends of ancient urban music cultures came up

I



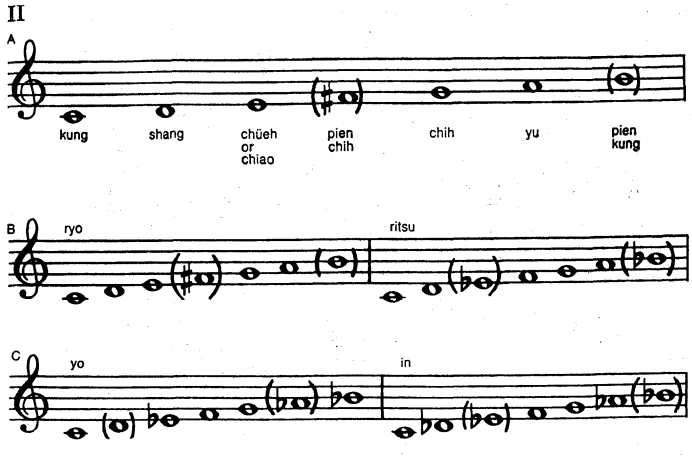
with very similar systems. One of the reasons that they do not sound the same today is that in the 17th century, Western musicians and acousticians found a formula that allowed them to bypass the "natural" system and hence make all pitches equidistant. The same formula was discovered by a Chinese, Hsü Tsai-yu, about 1596; but such "well-tempered" tuning was not accepted in Chinese music practice until very recent times. There are, of course, many other differences in the sound, the instrumentation, and the forms of Western and Eastern music. Many will be noted in the subsequent chronological study of East Asian music. But if the wonder of such variants is to be fully appreciated it must be understood that music is not in fact an international language. It consists of a whole series of equally logical but sometimes very different closed systems. The word closed is used to mean that the musical facets mesh perfectly within a given system, but they often may prove difficult or impossible to transfer to another system. In this light, a given passage of Chinese music when analyzed or judged with the logic of Beethoven is chaos, but Beethoven seems equally illogical when viewed in the context of Chinese, or for that matter Indian, music theory. Such intercultural clashes can be constructed between almost any of the larger systems. In this context, one can see that Chinese music is tonally more foreign to Near Eastern or Indian music than to Western, though historically it had closer relations with the other two. There are, of course, many other musical concepts and styles that travelled over the silk route between China and other parts of Asia; but these must be held aside until the discussion of the history of Chinese music.

Internally logical, closed systems

MUSICAL TRAITS COMMON TO EAST ASIAN CULTURES

In these primary considerations a view of some general aesthetic traditions common to much of East Asian music is also requisite. The tonal vocabulary of 12 tones generated in a cycle of fifths is the first common factor. From this tonal vocabulary various scales of five to seven notes are chosen, as they are in the West. The total number of

II



notes in an East Asian scale is often seven, but each tends to have what could be called a five-tone (pentatonic) core, as seen in notation II. The one scale in which no half steps appear (the so-called anhemitonic pentatonic; the first scale in II-A) is common all over the world, although casual listeners often mistake it as being characteristically "Oriental." A study of other East Asian scales, such as the *yo* and *in* scales shown in II-C, in addition to a few moments spent listening to authentic East Asian recordings, will reveal clearly that Oriental music comprises a great deal more than the black notes on the piano augmented by a few bangings of cymbals or gongs.

Although gongs and cymbals are well-known elements in much modern Chinese opera music (*i.e.*, after 1500), such instruments are not particularly common to East Asia in general. Indeed, there are a great variety of combinations, many of them historically related. Perhaps the only fundamental parallels between the multitude of East Asian performance practices lie in two common characteristics: (1) the emphasis on melody supported by rhythm, with little interest in harmonic accompaniment such as is found in the Western Classical style; and (2) a preference for the chamber-music sound ideal. The latter term means that no matter how large an ensemble may be, the individual instruments are meant to be heard. This differs from the orchestral sound ideal, popular in 19th-century Western music, in which the intention is to merge the sounds of the individual instruments into one musical colour.

Melodic and sonic ideals

The degree to which the melodic and sonic ideals are applied in East Asian music will become evident in the examination of specific genres. At this point it is necessary only to note the rather common tendency for East Asian music to be word or pictorially oriented. Such orientations are obvious in the many vocal and dance forms of Asia; but with a few notable exceptions one also finds that purely instrumental music will be attached through its title to a literary moment or a scene. Perhaps this relates to a general sensitivity to nature in East Asian culture as a whole. Whatever its source, it has produced many sonorous and pleasing results.

II. The music of China

One always approaches any survey of Chinese music history with a certain sense of awe—for what can one say about the music of a varied, still active civilization whose archaeological resources go back to 3000 BC and whose own extensive written documents refer to endless different forms of music in connection with folk festivals and religious events as well as in the courts of hundreds of emperors and princes in dozens of different provinces, dynasties, and periods? If a survey is carried forward from 3000 BC it becomes clear that the last little segment of material, from the Sung dynasty (960–1279) to today, is equivalent to the entire major history of European music. For all the richness of detail in Chinese sources, it is only for this last segment that there is information about the actual music itself. Yet the historical, cultural, instrumental, and theoretical materials of earlier times are equally informative and fascinating. This mass of information will be organized into four large chronological units: (1) the formative period, from 3000 BC through the 4th century AD; (2) the international period, from the 4th through the 9th century; (3) the national period, from the 9th through the 19th century; and (4) the world music period of the 20th century.

FORMATIVE PERIOD

Ancient artifacts and writings. Chinese writings claim that in 2697 BC the emperor Huang Ti sent a scholar, Ling Lun, to the western mountain area to cut bamboo pipes that could emit sounds matching the call of the phoenix bird, making possible the creation of music properly pitched for harmony between his reign and the universe. Even this charming symbolic birth of music dates far too late to aid in discovering the melodies and instrumental sounds accompanying the rituals and burials that occurred before the first historically verified dynasty, the Shang (1766–1122 BC). The beautiful sounds of music are evanescent, and before the invention of recordings they disappeared at the end of a performance. The remains of China's most ancient music are found only in those few instruments made of sturdy material. Archaeo-

logical digs have uncovered globular clay ocarinas (*hsün*), tuned stone chimes (*ch'ing*), and bronze bells (*chung*); and the word *ku*, for drum, is found incised on Shang oracle bones. The earliest surviving written records are from the next dynasty, the Chou (1122–221 BC). Within the famous five classics of that period, it is in the *Li chi* ("Collection of Rituals") of the 2nd century BC that one finds an extensive discussion of music. The *I Ching* ("Classic of Changes") is a diviner's handbook built around geometric patterns, cosmology, and magic numbers that indirectly may relate to music. The *Ch'un Ch'iu* ("Spring and Autumn" Annals), with its records of major events, and the *Shu Ching* ("Classic of History"), with its mixture of documents and forgeries, contain many references to the use of music, particularly at court activities. There are occasional comments about the singing of peasant groups, an item rare even in the early historical materials of Europe. The *Shih Ching* ("Classic of Poetry") is of equal interest, for it consists of the texts of 305 songs that are dated from the 10th to the 7th centuries BC. Their great variety of topics (love, ritual, political satire, etc.) reflect a viable vocal musical tradition quite understandable to modern radio or record listeners. The songs also include references to less durable musical relics such as the flutes, mouth organ (*sheng*), and, apparently, two forms of the zither (the *ch'in* and the *se*).

Chou dynasty musical documents

Aesthetic principles and extramusical associations. Despite the controversial authenticity and dates of ancient Chinese written sources, a combined study of them produces tantalizing images of courtly parties, military parades, and folk festivals; but it does not provide a single note of music. Nevertheless, in keeping with the prehistoric traditions of China, the philosophies of sages, such as Confucius (K'ung-fu-tzu, 551–479 BC) and Mencius (Meng-tzu, 371–289 BC), and the endless scientific curiosity of Chinese acousticians furnish a great deal of rather specific music theory as well as varied aesthetic principles. The straightest path to this material is found on the legendary journey, mentioned earlier, of Ling Lun in search of bamboo pipes. The charm of such a tale tends to cloud several interesting facts it contains. First, it is noteworthy that the goal of the search was to put music in tune with the universe. This extramusical need was noted earlier in the general discussion of world music history. It is upheld in theory in the "Annotations on Music" ("Yüeh-chi") section of the *Li chi* with such comments as "Music is the harmony of heaven and earth while rites are the measurement of heaven and earth. Through harmony all things are made known, through measure all things are properly classified. Music comes from heaven, rites are shaped by earthly designs." Referring back to the previously given general definition of music, it can be seen that such cosmological ideals may be not merely ancient superstitions but actually cogent insights into the cultural function of music in the societies of men. Confucius, as pictured in the *Analects* written long after his death, had a similar view of music, including a concern for the choice of music and modes proper for the moral well-being of a gentleman. It is an open question as to how much performance practice followed the admonitions and theories of the scholars; but centuries later one finds numerous pictures of the wise man standing before some natural beauties while his servant follows closely behind him carrying his seven-stringed zither (*ch'in*) for proper use in such a proper setting.

Another point to be noted in the legend of the origin of music is the fact that Ling Lun went to the western border area of China to find the correct bamboo. It shall be noted as this article progresses how often cultures from Central and West Asia or tribal China influenced the growth and change of music in Imperial China. Finally, it is significant that, although the emperor in the myth was primarily concerned with locating pipes that would bring his reign into harmony with the universe, the goal was also the creation of precise, standard pitches.

Tonal system and its theoretical rationalization. The cyclic creation of a 12-tone system in China has been mentioned before. Under the Chinese classical system this cycle began from a basic first pitch called yellow bell (*huang chung*), produced by blowing on the first bamboo tube (*lü*). This concept is of special interest because it is the world's oldest information on a tone system concerned with very specific pitches as well as the intervals between them. The precise number of vibrations per second that created the yellow bell pitch is open to controversy (between middle C♯ and the F above) because the location of this pitch could be changed by the work of new astrologers and acousticians in behalf of a new emperor, in order that his kingdom might stay in tune with the universe. (The note C was used in the previously given example in deference to Western readers; it should not be assumed that a pitch identical to C♯ was necessarily central to ancient Chinese music). The choice of the primary pitch in China had extramusical as well as practical applications, for the length of the yellow bell pipe became the standard measure (like a metre); and the number of grains of rice that would fill it were used for a weight measure. Thus, the pipe itself was often the property not of the Imperial music department but of the office of weights and measurements.

Practical applications of Chinese music

Mathematical relationship of pitches. The bamboo *lü* pipe is closed at the bottom by a node in the bamboo, with the result that another pitch a fifth and one octave higher could be produced on it by blowing more strongly (overblowing) as shown in notation I. This new pitch could be produced an octave lower by constructing a separate pipe two-thirds the size of the first one. If one then continued to construct pipes alternately four-thirds and two-thirds the length of the previous ones, an entire system of 12 notes could be generated, which is, with the exception of the means of creation, acoustically and proportionately in the same relation as is found in the Greek Pythagorean system. The English versions of the Chinese names for the 12 pitches seem quite fanciful; but they represent theoretically correct pitches, as do terms used in the Western traditional system, such as C or A♭. The source of each name in the Chinese system is conjectural; but Chinese classical acousticians, like modern Western scientists, no doubt found value in creating a professional nomenclature that was divorced from everyday speech and potentially descriptive of the nature of the object. For example, the use of bell names may relate to the gradual preference for tuned bells over pipes in the music division of the courts. Names like "old purifier" and "equalizing rule" may refer to aspects of the fact that, in both the Greek and Chinese "pure" systems, tones created beyond the first 12 will be out of tune with the original set. In the West, the degree of out-of-tuneness—the famous so-called Pythagorean comma—caused great problems as music became more involved in harmonic textures and modulations. Chinese academicians toyed with the same mathematical and acoustical problems of the comma, but in general they were not of a major concern for practicing musicians because the Chinese did not use the 12 tones in the same way as did the Westerners.

Scales and modes. For both traditions, the 12 pitches are merely a tonal vocabulary from which a specific ordering of a limited number of pitches can be extracted and reproduced on different pitch levels. Such limited structures are called a scale. With a set scale it is possible to emphasize different notes in such a way that they seem to be the pitch centre. Such variations of pitch centre within a scale are called modes. In the Western traditional systems most scales use seven tones that can be transposed and that contain modes. For example, C major (C–D–E–F–G–A–B) can be made a Dorian mode by using D as the pitch centre without changing the pitches used (D–E–F–G–A–B–C); and the whole scale and its modes can be transposed to a higher or lower pitch level (F major, E♭ major, etc.). The Chinese system concentrates in a similar way on a seven-tone scale but with a five-tone core (*wu sheng*) plus two changing (*pien*) tones, as shown in notation II-A.

Pitch centre

The notes of a scale (a set of intervals not tied to specific pitches) are often indicated in Western music with syllables such as *do re mi*. The Chinese equivalent terms for

notes in their classical scale are given in notation II-A. As in the Western system, modes can be constructed in Chinese music, and the scale can be transposed. From these comments it can be seen that the mythical emperor Huang Ti seems to have founded a very thorough system indeed. Throughout the Ch'in (221–206 BC) and Han (206 BC–AD 220) dynasties Imperial systems were tuned and retuned to meet Imperial and heavenly needs. Theoretical sophistications and experimentations continue on to the present day. How far back they may go in time is unknown, but in recent times there have been discovered stone chimes from the 2nd millennium BC that imply by their tunings that the Chinese classical tone system tradition may actually be as ancient as the legends claim. It is a pity that the music was not equally durable.

Extramusical associations. Returning to the extramusical aspects of the Chinese system, one finds that the five fundamental tones are sometimes connected with the five directions or the five elements, while the 12 tones are connected by some writers with the months of the year, hours of the day, or phases of the moon. The 12 tones can also be found placed in two sets of six on Imperial panpipes (*pai-hsiao*) in keeping with the female–male (Yin and Yang) principle of Chinese metaphysics. Their placement is based on the generation of the pitches of each pipe by its being either four-thirds larger or two-thirds smaller than the previous one, the smaller ones being female.

Classification of instruments. The Chinese talent for musical organization was by no means limited to pitches. Another important ancient system called the eight sounds (*pa yin*) was used to classify the many kinds of instruments used in Imperial orchestras. This system was based upon the material used in the construction of the instruments, the eight being stone, earth (pottery), bamboo, metal, skin, silk, wood, and gourd. The sonorous stones, ocarinas, and flutes mentioned earlier are examples from the first three categories. The bells are obvious metal examples. Another ancient member of the metal category is a large bronze drum (*t'ung ku*), which is of special interest because of the widespread distribution of archaeological examples of it throughout Southeast Asia. Equally intriguing are the designs and sounds of the bronze head of the drum as well as the frequent statues of frogs around the rim of the head. Han dynasty military expeditions to the south report that bronze drums among southern peoples represented the spirit of rain and water and rumbled like bullfrogs. The possession of such bronze drums or later gongs was, and still is, prestigious among tribal groups in Southeast Asia.

Stringed instruments of ancient China belong to the silk class because their strings were never gut or metal but twisted silk. Drums are skin instruments, whereas percussive clappers are wood. The most enjoyable member of the wooden family is the *yü*, a model of a crouching tiger with a serrated ridge or set of slats along its back that were scratched by a bamboo whisk in a manner recalling the various scratched gourds of Latin American bands. The Chinese category of gourd is reserved for one of the most fascinating of the ancient instruments, the *sheng* mouth organ. Seventeen bamboo pipes are set in a gourd or sometimes in a wooden wind chest. Each pipe has a free metal reed at the end encased in the wind chest. By blowing through a mouth tube into the wind chest and closing a hole in a pipe with a finger the reed will be made to sound, and melodies or chord structures may be played. Many variants of this instrumental principle can be found in Southeast Asia, and it is not possible to know with assurance where this wind instrument first appeared. Western imitations of it are found in the reed organ and, later, in the harmonica and the accordion.

The mouth organ

Han dynasty: musical events and foreign influences. The extensive work in theory and classification in ancient times implies that there must have been an equally large amount of performance practice. Modern information on all of these elements of music has suffered because of the destruction of many books and musical instruments under the order of Ch'in Shih Huang Ti, emperor of the Ch'in dynasty (246–209 BC). Yet there are several survivals from the Han dynasty (206 BC–AD 221) that do give some insight into how the musical events took place. In the court and the Confucian temples there were two basic musical divisions: banquet music (*yen yüeh*) and ritual music (*ya yüeh*). Dances in the Confucian rituals were divided into military (*wu wu*) and civil (*wen wu*) forms. The ensembles of musicians and dancers could be quite large, and ancient listings of their content were often printed in formation patterns in a manner analogous in principle to those of football marching bands in America today. Rubbings from Han tomb tiles show more informal and apparently very lively music and dance presentations at social affairs. The early Chinese character for dance (*wu*) implies movement by the body more than by the feet. The folk sources of many of the songs from the *Shih Ching* and later books show that courtly musical life was not without its gayer and more personal and secular moments. The stringed instruments, notably the seven-stringed *ch'in* zither, apparently were popular as vehicles for solo music.

The Han dynasty empire expanded and at the same time built walls between its national core and western Asia. But these actions were paralleled by an increasing flow of foreign ideas and materials. Buddhism entered from India to China in the 1st century AD, whereas booty, goods, and ideas came from Central Asian Gandharan, Tokharian, and Iranian cultures along the various desert trade routes via the cities of Khotān to the south (3rd to 5th centuries), Kucha in the centre (4th through 8th century), and Turfan to the north (5th through 9th century). Desert ruins and Buddhist caves from this period and later reveal a host of new musical ensembles and solo instruments. Two stringed instruments of particular interest are the angle harp (*k'ung-hu*) and the pear-shaped plucked lute (*p'i-p'a*). The harp can be traced back across Central Asia to the ancient bas reliefs of Assyria. The lute also seems to have West Asian ancestors but is a more "contemporary" instrument. Variants of this instrument have continued to enter or be redesigned in China down to the present day. A delightful symbol of the long-term musical and commercial value of such a plucked lute is found on a 10th-century clay statue of a caravan Bactrian camel with two different styles of *p'i-p'a* tied to the saddle post on top of the rest of the cargo.

Diffusion of instruments in the Han dynasty

New percussion instruments are evident in the celestial orchestras seen in Buddhist iconography. One apparent accommodation between old Chinese and West Asian tradition is the *fang-hsiang*, a set of 16 iron slabs suspended in a wooden frame in the manner of the old sets of tuned stones. Knobless gongs related to the present-day Chinese *lo* seemed to have entered the Chinese musical scene before the 6th century from South Asia, while the cymbals (*po*) may have come earlier from India via Central Asian groups. One of the most sonorous Buddhist additions was a bronze bell in the form of a basin (*ch'ing*) that, when placed rim up on a cushion and struck on the rim, produces a tone of amazing richness and duration. Among the varied new instruments pictured in heavenly ensembles one can still find occasional "old time" instruments such as a set of narrow wooden clappers (*ch'ung-tu*) tied together on one end like ancient wooden books. The clappers were sounded by compressing them quickly between the hands. Variants of this Chou dynasty instrument are still heard in all three major East Asian countries.

Not all of the new influences in China came via religious or trade activities. During the Six Dynasties period (AD 220–589) China was rent by internal strife and border wars. The constant confrontations with the Tatars of the north caused an increased interest in the musical signals of the enemy via drums, trumpets, and double reeds. Although related instruments were equally evident to the south and west, there can be little doubt that the creation of cavalry bands with double kettledrums are direct imitations of the musical prowess of the horseback terrorists against whom the walls of China were built. With great effort and much blood, China gradually reunified under the Sui dynasty (581–618), and older courtly music and the latest musical fads were consolidated.

T'ANG DYNASTY

Thriving of foreign styles. The few centuries of T'ang dynasty existence (618–907) are supersaturated with brilliant Imperial growth and cultural flourishing as well as military and natural disasters. Such a rich loam of good and bad nourished one of the most fascinating eras of music history in the world. The more formal Imperial ceremonies revitalized the ancient orchestras of bells, stone chimes, flutes, drums, and zithers plus large bands of courtly dancers. In reality, Imperial power was based perhaps less on the Mandate of Heaven than on the "liberation" of neighbouring countries, a development of more thorough tax systems and more and more trade cities and harbours. Into all these power sources flowed foreign goods and foreign ideas. Persians, Arabs, Indians, and Malayans were found in the foreign quarters of port towns, while every trade caravan brought in masses of new faces and modes of living. Perhaps it is not surprising that an 8th-century poet, Yüan Chen, should lament about air pollution created by western horsemen, about the ladies who studied western fashions and makeup, and about the entertainers who devoted themselves to only "western" music. (One must remember that the term western refers to the land west of the China wall.)

Tavern entertainers

There was hardly a tavern in the capital of Ch'ang-an (now Sian, Shensi Province) that could compete without the aid of a western dancing or singing girl with an accompanying set of foreign musicians. Popular tunes of the period included "South India" and "Watching the Moon in Brahman Land," while beautiful, exotic dancing boys or girls were ever the rage. One set of girls from Sogdiana (centred in modern Uzbek Soviet Socialist Republic) won the support of the emperor Hsüan Tsung by being costumed in crimson robes, green pants, and red deerskin boots and twirling on top of balls. Other girls from the area today called Tashkent inspired a poet of the 9th century, Po Chü-i, with their dance, which began by their emergence from artificial lotuses and ended by the pulling down of their blouses to show their shoulders, a style not unfamiliar to old Western burlesque connoisseurs. A study of the lithe bodies and flying sleeves on T'ang clay dancing figurines is an even more compelling proof of the style of the era. In such a context one can understand how eventually an additional character was added sometimes to the word for dance to indicate the movement of the legs as well as of the body.

In addition to all the commercial musical enterprises of the T'ang dynasty period, there was another equally extensive system under government supervision. The T'ang emperor Hsüan Tsung (712–756) seemed particularly keen on music and took full advantage of the various musical "tributes" or "captives" sent to him by all the nations of Asia. This plethora of sounds was further enriched by the special area in Ch'ang-an called the Pear Garden (Li Yüan), in which hundreds of additional musicians and dancers were trained and in which the emperor himself was most active. Such trainees were naturally often female. They followed in an earlier tradition of court girls (*kung nü*) whose basic duties were to entertain distinguished guests. A remnant of this tradition is found in the Japanese geisha girl of today.

The mass of different foreign musical styles in the capital was too much for the government musical bureaucracy. A distinction already had been made between court music (*ya yüeh*) and common music (*su yüeh*); but T'ang nomenclature added a third kind—foreign music (*hu yüeh*). Eventually officials organized Imperial music into the ten kinds of systems (*shih pu chi*). Of these categories, one represented instrumentalists from Samarkand, whereas another group came from farther west in Bukhara. Kashgar, at the mountain pass between the East and West, sent yet a different group. Musical ensembles also were presented to the emperor from the eastern Turkistan trade centres of Kucha and Turfan. India and two recently defeated kingdoms of Korea provided still other musicians. Chinese and Kucha music were blended by different musicians. One group was supposed to maintain the old styles of Chinese folk music, and there had to be one special group for the performance of formal Chinese court music. These ten types by no means completed the picture, for nearly every Asian culture took its chance at musical goodwill in Ch'ang-an. Nothing from further west appears in T'ang China, for culture hardly existed in Europe at that time. Nevertheless, one can sense in T'ang musical culture an internationalism not matched until the mass communications of the mid-20th century provided radio and phonograph owners with the delights of a similarly exotic and extensive choice.

Courtly music. The only music that can be discussed in a survey of a repertoire so large is the more official courtly music. Ritual presentations are generally divided into two types: so-called standing music, performed without strings and apparently in the courtyard; and sitting music, for a full ensemble played inside a palace. There are lists of the names of some pieces in these categories with their authorship usually credited to the emperor or empress of the time. For example, "The Battle Line Smashing Song" was said to be by the T'ang emperor Tai Tsung (626–649). The accompanying dance is listed for 120 performers with spears and armour. A similarly grandiose piece is the "Music of Grand Victory" credited to the next T'ang emperor, Kao Tsung (649–683). Empress Wu (died 705) is said to have written "The Imperial Birthday Music," in which the dancers form out the characters for "Long Live the Emperor" in the best modern marching-band tradition. Music inside the palace includes a concert version of "The Battle Line Smashing Song," with only four dancers, "A Banquet Song," and a piece supposedly composed by Empress Wu in honour of her pet parrot, who always called out "Long Live Her Majesty." Those familiar with music in the courts of Henry VIII and Louis XIV or with the songs always ending in praise of Queen Elizabeth I may recognize the cultural context of such music.

Courtly extravaganzas

Later dynasty copies of T'ang paintings show ladies entertaining the emperor with ensembles of strings, winds, and percussion; and many of the choreographic plans of the larger pieces are also available in books. According to some sources, court orchestra pieces began with a prelude in free rhythm that set the mood and mode of the piece and introduced the instruments. This was followed by a slow section in a steady beat, and the piece ended in a faster tempo. Documents also tell much about the instrumentation and the colour and design of each costume of the musicians and dancers. No orchestral scores are to be found, however. One solo piece for *ch'in* survives and 28 ritual melodies for *p'i-p'a* were discovered in the hidden library of the Tun Huang (Cave of the Thousand Buddhas), but the grand musical traditions of T'ang remain frustratingly elusive. Major clues to their actual sounds will be found in marginal survivals of such music, which will be discussed in the Korean and Japanese sections. The original traditions waned with the decline of T'ang good fortune, and the conflicts of the Five Dynasties and Ten Kingdoms period (907–980) brought the international period to an end.

SUNG AND YUAN DYNASTIES

Consolidation of earlier trends. Despite the chaos of kingdoms in the 10th century, or perhaps because of it, cultural traditions solidified, so that by the Sung dynasty (960–1279) one can speak of a national rather than an international cultural mood. Many of the short-lived usurpers of regional governments were of "barbarian" (*i.e.*, Turkish) origin, but their general cultural efforts were to appear Chinese rather than to import further foreign fads. But one significant foreign musical addition of the period was from the northern Mongols in the form of a two-stringed fiddle, or bowed lute—the "foreign lute" (*hu-ch'in*). It became an important feature of the plebian theatre and teahouse world, which grew stronger and larger as more musicians and dancers were dropped from government payrolls. With the establishment of the Sung court, Confucian ceremonies and similar "old-fashioned" musical events were revived; but Imperial contributions to music of the period were primarily in the creation of gigantic historical or encyclopaedic works. For example, the official *Sung shu* ("Sung History," 1345) contained

496 chapters, of which 17 deal directly with music, and musical events and people appear throughout the entire work. The *Yü hai* encyclopaedia (*c.* 1267; "Sea of Jade") has 200 chapters, with ten on music. It is interesting that the *lü* pipes are discussed separately under the topic of measurements. Manuals on how to play the seven-stringed *ch'in* zither also survive, as well as rare music collections such as the "Songs of Whitestone, the Taoist," based on the poems and songs of Chiang Kuei (1155–1221) and first printed in 1202. Many Sung poets continued to use the five- and seven-syllable-line *shih* form perfected by T'ang writers, which was believed to have been chanted to tunes strictly adhering to the word tones of the Chinese language. The singing girls (*chi kuan*) of the teahouses and brothels and the general growth of urban, mercantile life inspired the creation of *tz'u* poems, which were free of word-tone restrictions, filled with colloquial phrases, and capable of freewheeling musical settings. A major source for music based on both the old and new forms is found in the rising world of public theatre.

Music theatre. Chinese drama can be noted as far back as the Chou dynasty, but it was really the T'ang period Pear Garden school that quite literally set the stage for Chinese opera. Regional music-drama flourished throughout the Sung empire, but the two major forms were the southern drama (*nan ch'ü*) and the northern drama (*tsa ch'ü* or *pei ch'ü*). The *tz'u* poetical form was popular in both, although the southern style was held to be softer, with its emphasis on five-tone scales and flute and percussion accompaniments. The northern style is said to have preferred the seven-toned scale, to have used more strings, and in general to have been bolder in spirit. According to period writers, each of the four acts of a northern drama was set in a specific mode in which different tunes were used, interspersed with dialogue. The southern style was more lyrical.

Northern and southern styles

The Mongols under Genghis Khan and later Kublai Khan finally succeeded in invading China, and the foreign Yüan dynasty (1279–1368) was founded. The two styles of drama noted above continued and intermixed under Yüan drama (*Yüan ch'ü*), while the basic poetical form became *san-ch'u*, popular songs of even freer style. On stage there appeared standard songs for specific situations or emotions that could be used in any opera, thus making it easier to communicate a story to mass audiences who may have spoken in many different dialects. Additional appeals to the general public were made by bringing onto the stage several forms of dancing and acrobatics, events that had been, along with several forms of puppet theatre, such gay parts of Chinese city life during the Sung dynasty.

MING AND CH'ING DYNASTIES

Internal Mongol struggles, natural disasters, and peasant revolts permitted the return of Chinese rule and the founding of the Ming dynasty (1368–1644). It in turn gave way to Manchu invasions from the north under which the last dynasty, the Ch'ing (1644–1911) was formed. Though there is much history and much blood involved in all such changes, one can view the music of these eras together under their two most active styles—theatre music and instrumental pieces.

Further development of opera. *Forms of the 16th–18th centuries.* The flourishing of regional music-drama has continued throughout the centuries since the Sung dynasty until the present day. Musically they vary greatly in their instrumentation and particularly in their voice qualities. But all tend to follow a tradition of using either standard complete pieces (*lian ch'ü*) or stereotyped melodic styles (*ban ch'iang*) in every opera. The complete-piece approach of Yüan drama survives today primarily in a 16th-century form called *k'un ch'ü*. Nurtured in a more aristocratic form of theatre, the music of *k'un ch'ü* was less bombastic than that of the popular theatre. The major instruments were the horizontal flute (*ti tsu*) and the notched vertical flute (*hsiao*). The flutes often produce a special mottled tone by the presence of one hole that is covered by thin rice paper that buzzes quietly as one plays. The *sheng* mouth organ and the *p'i-p'a* plucked lute could also be found in *k'un ch'ü*, along with a single free-reed pipe, *kuan*. The term *kuan* usually stands for one of several forms of double-reed woodwinds with cylindrical bore and no bell. Survivors of its ancient forms are found in Korean and Japanese court music. Variants of the single-reed *kuan* are found throughout Southeast Asia, where it is equally appreciated for its mellow, clarinet-like tone. A plebian instrument found in some *k'un ch'ü* is the three-stringed plucked lute (*san-hsien*) with a snakeskin soundboard. Plucked with a bone pick, it enjoys great popularity in folk music as well as theatre music, and it developed in two sizes, the shorter one prevalent in the south and the longer one in the north. The shorter form is of particular historical interest, for it was imported into the Ryukyu Islands as the jamisen and from there moved to Japan, where it evolved into a samisen.

The *k'un ch'ü* form

The vocal style of *k'un ch'ü* matched the mauve accompaniment and was usually performed by a male singing falsetto. Another style of opera from the same period, *i-yang ch'iang*, seemed more appealing to the general public and is noteworthy for its use at some point in its development of a chorus (*pang-ch'iang*) as well as of soloists. In addition, passages in colloquial speech were often interpolated between lines of classical poetry in order to explain them. Such lines were often sung. Still another Ming music-drama genre of considerable influence in the myriad regional forms is the clapper opera (*pang-tzu ch'iang*). In addition to the rhythmic importance of the clappers, the instrumental accompaniment of this form is noted for its emphasis on strings, the principal form being the moon guitar (*yüeh-ch'in*), a plucked lute with a moon-shaped wooden body and four strings in double courses. An interesting addition to this instrument is the presence of a thin strip of metal tied at both ends inside the body to give the instrument a richer tone. Among the endless variants of style and accompaniments in Chinese regional opera one must add the sounds of the extremely large flat gongs heard in the southwest and the *yang ch'in* (western zither), particularly popular in Cantonese music. The latter is often called a butterfly harp, though it is neither a harp nor a butterfly but a hammered dulcimer derived from a Near Eastern instrument (*santūr*) brought into China in the 18th century. For the rest of the materials concerning theatre music it is best to turn to the primary music-drama form since the 18th century, Peking opera (*ching-hsi*).

Peking opera. Credit for the beginning of Peking opera is given to actors from Anhwei appearing in Peking in the 1790s. But Peking opera really combines elements from many different earlier forms and, like Western grand opera, can be considered to be a 19th-century product. In addition to all the instruments mentioned above, many others may be found. Barrel drums with tacked heads (*ku*) and a double reed with a conical bore and bell (*so-na*) are used in military scenes, along with cymbals (*po*) and large flat gongs. The two most common percussion instruments are a small flat gong (*lo*) and a *pan ku* drum. The small gong is some eight inches in diameter; the face is slightly curved except for a flat centre spot. It is designed in this manner in order that the tone and pitch of the gong will rise quickly each time it is hit. This "sliding" gong effect is characteristic of the Peking sound. The *pan ku* or *tan-pi ku* is equally unique in construction. The skin is stretched over a set of wooden wedges strapped in a circle with only a small spot in the middle completely hollow. This allows the performer to produce a very dry, sharp sound. Such a tone is practical as well as aesthetic, for the *pan ku* player is the leader and his signals are essential to the coordination of the performance. In combination with the large and small gong, cymbals, and a clapper, the *pan ku* also controls the very important rhythmic elements of the drama. In all East Asian music one must remember that harmony and harmonic progression are not parts of traditional music. The functions of harmony—such as underlining expression, providing sonic contrast, and creating a sense of forward motion—are handled with equal efficiency by rhythm in

Role of rhythm

East Asia, although the methods and sounds are very different. In both traditions, the choices are not arbitrary, and with cultural exposure one comes to recognize the musical intention, even though it is not necessary to know precisely what chord or what rhythm pattern produces an appropriate musical effect. For example, very few listeners to Western music know that a doubly diminished chord (C–E♭–G♭–A) played tremolo means danger, although all would recognize the danger signal by ear. By the same token, a Peking opera fan hearing the large gong played alone in the rhythm

would know that the situation is a similar moment of confusion but probably would not know that the pattern is named the scattering hammer (*luan-chueh*). Pattern names are for specialists, but pattern sounds and "meanings" are for attuned listeners. Other aspects of the functions of rhythm in East Asia will emerge in the examination of other cultures. For the moment, attention will be given to the melodic side of Peking opera.

Like any theatrical music, the tunes of Peking opera must conform with the text structure and the dramatic situation. In the latter case, one finds that a majority of Peking aria texts are based on series of couplets of seven or ten syllables each. Although there may be several verses set in strophic form (*i.e.*, music repeated for each strophe, or stanza), part of the musical tension is maintained by the interjection of comments or short dialogue between the two lines of each verse. These leave the listener waiting for the completion of the line. The tune aids in this forward motion and tension by playing what could be considered an incomplete melodic cadence (point of resolution) at the end of line one, which is brought to a final resolution at the end of the second line. From a dramaturgical standpoint, the arias of Peking opera can be categorized into different types whose style is recognizable in the same way that one can tell, without language ability, the mood of a love, farewell, or vengeance aria in Italian opera.

Prototypes of Peking melodies

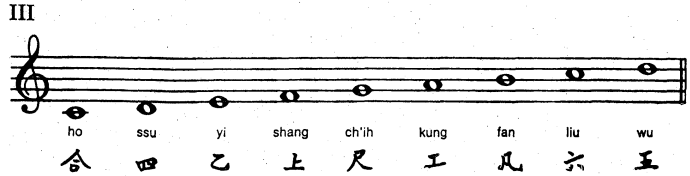
Peking melodies themselves tend to fall into two prototypes called *hsi p'i* and *erh huang*. Within each of these general types there are several well-known tunes, but the word "prototype" has been used to define them, as each opera and each situation is capable of varying the basic melody greatly. For example, if the scale shown in notation II-A is used as the tonal vocabulary of a given scene, one will find that *erh huang* melodies in it will tend to cadence on G or the C above it, whereas a female aria in the *hsi p'i* style will prefer E and A. It is not possible in a short explanation to refine the musical details further, except to add that the particular rhythm in which a melody is set will also allow its character and meaning to change. There are also other classes of melodies (as there are in Western opera of any one historical period), and there are action pieces (as there is ballet in some Western opera) and borrowings from other traditions (as there is church music in Western opera). In all, Chinese opera is filled with a musical and dramatic logic that, even at first introduction, helps to explain why it remains more popular than Western forms for more than a fourth of the world's population.

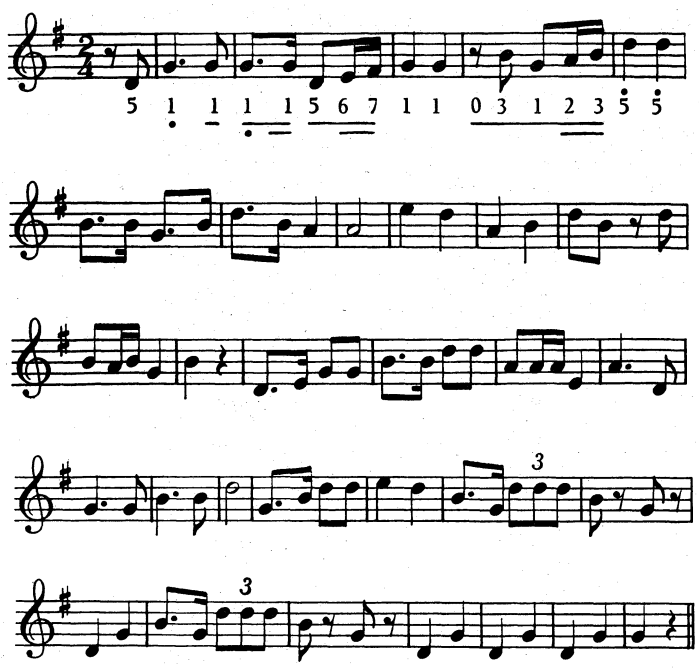
Other vocal and instrumental genres. The emphasis here has been on opera because it is best known, but there are many other popular forms from the Ming and Ch'ing periods. Musically, the various shadow- and hand-puppet plays are similar to the opera tradition except that, like Southeast Asian puppetry, a manipulator must often be the singer-narrator as well. The most common melodic instrument for both opera and puppetry is some form of fiddle, or bowed lute. It comes in several different forms, such as the *hu-ch'in, erh hu,* and *nan hu*. Although the shape of the body may be different, all traditional Chinese fiddles exhibit certain specific structural characteristics. The small body has a skin or wooden soundboard and an open back. The two strings pass over a bridge and then are suspended above a pole to the pegs, which are inserted from the rear of the scroll (not from the sides as on a Western violin). Such a system places one string above the other rather than parallel to it (as on a banjo or a *p'i-p'a*). Because of this, the bow passes between the strings, playing one string by pressing down and the other by pulling up. The fingerings of tunes are done by sideways pressure, along the strings; they are too far from the pole for it to serve as a fingerboard, which, because of the vertical stringing, would be a nuisance in any case. It is this unique manner of fiddle construction that helps one determine the source of many of the bowed lutes of Southeast Asia.

A simpler form of vocal accompaniment is found in a popular narrative idiom called *ta ku tz'u,* which uses a small flat drum with an occasional plucked string instrument such as the *san-hsien.* A *p'i-p'a* player may also be heard accompanying himself in a *tan tz'u* narrative. Leaving the endless forms of vocal music, it is appropriate to turn briefly to the instrumental. The 25-stringed *se* zither, with movable bridges, and the seven-stringed *ch'in,* with permanent upper and lower bridges (like a piano), were well-known for solo music in ancient times. During the last dynasty collections of *ch'in* music and instruction books flourished as part of certain neo-Confucian revivals. Many musical notations were developed, perhaps the most interesting variety for the *ch'in* being one in which Chinese characters were artificially constructed by combining symbols for the notes with indications of fingering technique, such as up strokes, down strokes, or harmonics. Although most of the music was based on vocal pieces or evoked some scene, there were several examples of variation forms that had an important influence on Korean and Japanese forms that followed. The *p'i-p'a* likewise developed an extensive repertoire of solo pieces, many of them quite virtuosic and pictorial. For example, anyone hearing a *p'i-p'a* battle piece needs to know very little Chinese to recognize the musical interpretations of the action. In recent times there has been a considerable revival of solo literature for the *cheng,* a zither with 16 strings and movable bridges whose popularity spreads as far south as Vietnam. The strings are apparently influenced by the Near Eastern dulcimer mentioned above (*yang ch'in*), for they are metal.

Music for the *ch'in* zither

Chamber music exists in many styles, functions, and locations. Some of it can be considered folk music played by farmers or working people for festivals or private entertainment, as in the American bluegrass tradition. Music of this type can still be heard at weddings or funerals in Chinese communities all over the world. During the Ming and Ch'ing periods, small ensembles of courtiers or professional musicians could be found at palaces, but the major sources for this kind of chamber music were in the world of the musically inclined businessman or trader. Because of this, certain regional forms of chamber music such as Amoy "southern pipe" and Shantung music survive in such locations as Taiwan, Manila, Singapore, and San Francisco. In this context it is noteworthy that even during Japan's isolation period (1600–1868) Chinese vocal and chamber music, known in Japanese as *minshingaku* (Ming and Ch'ing music), was played in Nagasaki, the only open port in Japan. Examples of such dispersed regional music are of great value in the study of the oral history of Ming and Ch'ing music and of the distribution and development of various musical instruments. Much of the repertoire of such stylistic groups is derived from theatre music, but there are many examples that may imply the sounds of older lost traditions. There are a variety of notation systems, particularly for the solo music. The one most commonly used in tune books of the last dynasties is *kung ch'ih,* which

III

indicates notes in a scale as shown in notation III. This system is still popularly used, although mainland sources prefer the number system shown in the first line of notation IV. It is based on the 19th-century French *chevé* system (which used numerals 1–7 for the notes of the scale) and, unlike other Chinese notations, shows rhythm by the use of dots and beams borrowed from Western eighth and 16th notes. Percussion accompaniments can also be found in a similar style, as can larger ensemble scores, but both are more characteristic of 20th-century China.

IV

DEVELOPMENTS SINCE 1911

Period of the Republic of China and Sino-Japanese War. Under the influence of missionary and modernization movements, many musical experimentations occurred in the last dynasty, but these were greatly increased by the rise of the first republic in 1911 and the establishment of Communist rule in 1949. During the period of the republic and of the Japanese war, a plethora of new songs were created in "modern" style, the most famous being shown in notation IV. The piece was written in 1934 by Nieh Erh to text by the modern Chinese playwright T'ien Han as a patriotic march and was adopted as the national anthem in 1949. It is an excellent example of a mixture of new and traditional Chinese music. The first phrase implies a major mode with its use of F♯. But after that point the entire piece is Chinese pentatonic. The first phrase also leads one to expect symmetrical four-bar phrases, but the tune quickly takes a more flexible Chinese course. Chinese and Western composers continued to try out bits of each other's traditions with only occasional success, and individual Chinese artists have become famous for their performance on Western instruments. Chinese instruments in turn have been subjected to many modernizations, such as the building of a family of *erh hu* fiddles by the creation of bass and alto versions. In conjunction with this movement there was the appearance of concertos for such instruments accompanied by a mixed Western and Chinese orchestra.

The "modern" style

Communist period. As was noted earlier, many completely traditional forms continued, particularly in foreign Chinese communities. The special point of interest since 1949, however, is the application of Marxist doctrine to the musical scene of China. The first obvious area of change is found in the ever popular forms of regional and Peking opera. Although the appeal of traditional tales of emperors, princesses, or mythological characters could not be suppressed, the emphasis of all new operas was on workers, peasants, soldiers, and Socialism. Thus, the *Romance of the Three Kingdoms* (*San Kuo chih yen-i*) or the *Ruse of an Empty City* tend to be replaced by the *Raid on the White Tiger Regiment* or *The Red Guards of Hung Lake.* Aria topics also vary, such as "Looking Forward to the Liberation of the Working People of the World" or "Socialism is Good." The accompanying orchestra is usually larger than the traditional one and is set in families of instruments, sometimes with complete violin sections. Choruses often appear, representing the people or the army in the victorious advance of Socialism. Acrobatics are still popular, but one also sees ballets strongly influenced by the more energetic sections of Soviet productions. Ballet shoes are combined with peasant or Red Army uniforms, and nearly every finale includes large numbers of red flags and a portrait of Mao Tse-tung. For all these modernizations, a surprising large number of new operas follow the previously mentioned musical forms of prototype melodies set in appropriate rhythms.

As part of the encouragement of people's music the national government emphasized regional folk music. But, in general, folk music was "reconstructed" away from its former individualistic nature into collectives of choruses or folk orchestras. The topics of such regional songs were also reconstructed so that they reflect the new Socialist life. The most famous new folk song from Shensi Province is "Red is the East," while the Miao people are credited with "Sing in Praise of Chairman Mao." Every "minority" group in China has at least one song directly in praise of Chairman Mao, while other songs deal with local industries and accomplishments. Such songs are sometimes performed in regional style with traditional accompaniments, although they may often be found arranged Western-style for use in the public schools of the nation. This effort, in addition to the number of recordings that are also available, make it possible for a Chinese citizen to become aware, perhaps for the first time in history, of the great variety of local music traditions in his large country, even though such music appears now in Marxist reconstructions. Marxist defense of this changed folk music is that music of a given period must reflect the views and aspirations of the masses (as understood by the government) and must be based on idioms of the people. Composers of concert music have produced many folk orchestra compositions along with symphony, piano, and military band music based on this basic Marxist musical principle, called Socialist realism. When dealing with traditional instruments and vocal styles the composers have sometimes created extremely original and interesting pieces despite the general conservatism of government aesthetics policies. Vocal and choral music are preferred because of their ability to communicate specific national goals more efficiently than, for example, *The Sacred War Symphony*. Two excerpts from the words of "My Home Is the Collective Farm" represent typical goals of Socialist realist music:

Socialist realism

> Our land strips are fused into one, No boundaries, no distinctions of 'yours' or 'mine'; If you ask me what my name is, All of us are named collective farmers, All of us are named collective farmers. . . . People present us with prize certificates and banners, We hang them all up in our parlour; In the centre is a portrait of Chairman Mao. His heart must be glad. His heart must be glad.

The social philosophy implied in this message may not be appealing to all peoples, but it is as real and meaningful to 20th-century Chinese as was "A Mighty Fortress Is Our God" to 17th-century German Protestants, "A Hymn to Reason" to 18th-century French revolutionists, or "We shall overcome" to 20th-century American protesters. Most modern Western musical sales are for material goods, but China has concentrated its musical media on national messages.

It must be remembered that music exists in a cultural context and that it has never remained static since the world began. In the late 20th-century, music of all periods from every society is available to those with sufficient mass communication sources, although China has wished

to suppress what is called capitalist formalistic arts. It does seem unlikely that the tuning of the *lü* pipes for rulers will ever be a major concern of Chinese musicians again, but the ability of China to preserve so many historical facts, materials, and idioms along with modern changes is sufficient to keep the musical world in awe for some centuries to come.

III. The music of Korea

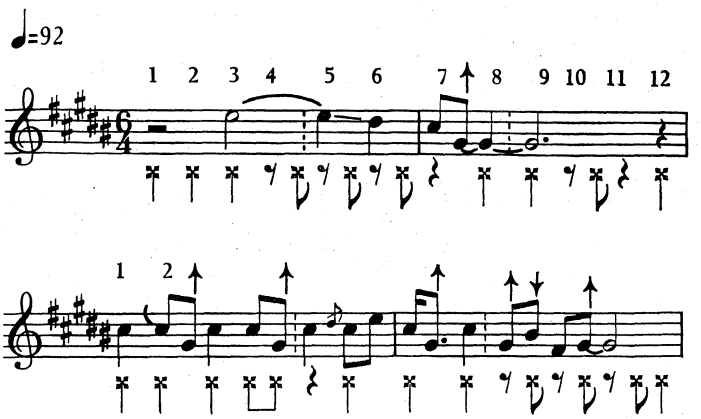
On a map Korea looks like a finger pointing from the top of China down to the lower part of Japan. Thus, one would expect its music to reflect its "bridge" position between two such powerful traditions. The movements of foreign, particularly Chinese, armies and cultures are indeed major factors in Korea's tradition. But beneath these reflections lies a deeper core of indigenous musical styles that, at first hearing, seem most strange to the ear of listeners with preconceived notions as to what East Asian music sounds like. A possible additional factor in the growth of Korean music is the country's position as a peninsula jutting out from Manchuria and from the spawning ground of many Mongolian hordes. It may be that part of Korean music is the result of a cultural cul de sac left by the Mongolian hordes as they dashed across the steppes of Asia to their eventual demise in the mountains of Europe. Archaeological sources indicate that various Mongol peoples from north Asia did indeed occupy areas of Korea from at least 2000 BC, and Chinese writings show that their people and armies were active in Korea from the period of the Chinese Han dynasty (206 BC–AD 220) on. Obviously, a study of Korean music contains riches extending far beyond its geographic borders.

SHAMAN MUSIC

The earliest references to music in Korea are found in a 3rd century AD Chinese history book that comments about agricultural festivals (*nong-ak*) with singing and dancing among the tribes of northwestern Korea. Such events are still a strong part of Korean life. Another ancient but long-lived tradition in Korea is shamanism, or communication with the unseen world by medicine men in a state of trance. This is of special interest because such a belief is characteristic not only of all north Asian tribes but also of other peoples (such as Eskimos) who live in the northernmost regions of the world. Korea is one of the few countries south of the Arctic area that maintain strong shamanism in the face of several foreign religious adoptions such as Buddhism, Confucianism, and Christianity. A few Korean shamanistic events, however, show still other possible prehistoric connections. For example, sometimes the head of an animal may be placed in the centre of a Korean ritual ground, and mythology says that the first ruler of Korea was created by the union of god and a bear. Divine origin is part of Japan's Imperial mythology, and the head of a bear is central to the important rites of the Ainu tribal culture in northern Japan.

Today, a Korean shaman (*mudang*) may use many combinations of musical instruments. The simplest and potentially most significant accompaniment is a small, flat gong with a slight rim. It brings to mind the single-headed pan drum with a wooden or bone hoop found in the shamanism of most of Central Asia and in the Arctic Circle as far away as Lapland and Hudson Bay. A drum sound itself is produced in Korea by the most popular percussion instrument, the *changko*, an hourglass-shaped, two-headed drum struck by the hand and a stick. In Korean shaman rituals flutes, double reeds, fiddles, and other gongs and drums may be used that at first sight may appear rather Chinese. The sound, however, creates a totally different impression. The music shown in notation V represents a flute and gong excerpt from a shaman ensemble that also contains a drum, zither, fiddle, and oboe in a driving polyphony (combination of simultaneous voices, or parts) that seems closer to Dixieland jazz than to Chinese music. The seven pitches of the flute tune, plus its microtonal slides, its use of a fourth (c♯″–g♯″) "out of tune" with the Chinese *lü* pipes, and its rhythm, produce a very un-Chinese, jazzlike sound. The style is totally traditional and is notated to show the special Korean proclivity for a six-beat unit, which in this excerpt is particularly strong in the accompaniment of the other instruments. The dotted bar lines in the transcription imply a kind of polymetric syncopation that often gives Korean folk and popular music its special appeal. One part seems to be in four whereas others are in six, so that they come together only after 12 beats. Triplets and even five-beat forms are found as well. Thus, it is evident from a brief look at one example from one type of Korean shaman music, in addition to a discussion of other forms and characteristics, that Korean cultural materials continue to reflect fascinating mixtures and mysteries. The total style cannot be called purely Chinese or north Asian but simply Korean. The characteristics of the first example are typical of the kind of music best known and loved by the general Korean populace. Studies in Korean music, however, have tended to concentrate on the less familiar styles of court music, which are maintained by dedicated national music institutes and Korean scholarship.

Polyphony, tuning, and rhythm

V

COURT MUSIC

The three kingdoms of Koguryŏ in the north, Paekche in the southwest, and Silla in the southeast were established in the century before AD 1 along with a small Japanese-related enclave (Kaya or Mimana). The organization of courts and introduction of Chinese religions resulted in an ever-increasing importation of Chinese musical materials. Indications of this can be found in such sources as paintings in an AD 357 tomb near An-ag, a colony of China at that time. A horseback band of the Chinese Han dynasty style is seen with drums and a small bell hit with a hammer. One brave rider apparently is able to play the Chinese panpipes in transit. Deeper in the tomb a zither, lute, and a very long end-blown flute can be seen accompanying a dancer whose long nose and costume imply that central Asian traditions may have travelled even as far as Korea by the 4th century. The Silla dynasty domination coincided closely (668–935) with the heyday of the Chinese T'ang period, and the subsequent Koryŏ (935–1392) and Yi dynasties (1392–1910) also tended to match parallel Chinese periods. Thus, Korea's court-music traditions tended to reflect those of China.

Being on the border of Chinese culture, Korea was able to maintain certain ancient traditions during periods of barbarian domination in China proper. Such marginal survivals are of particular importance because many have continued to the present day, thus giving extremely rare examples of music traditions long gone from the land of their origin. For example, in the Silla period, court music was divided into *hyang-ak*, Korean music; *T'ang-ak*, T'ang Chinese music; and *a-ak*, Confucian ritual music. The instruments used for these ensembles were those described earlier in the discussion of Chinese music, such as sets of tuned stones (in Korean *teuk kyeng*) and bells (*pyen chong*), mouth organ (*saing*), and instruments in all the other eight categories of Chinese classical traditions (*e.g.*, those based on the materials used in construction; see above). Unlike China, ensembles using such

Survivals of ancient Chinese music

instruments can still be heard in the national institutes of north and south Korea and in Confucian rituals. Among the many instrumental treasures still played in Korea is the *a-chaing*, a zither—with seven strings and movable bridges—that is not plucked but, more remarkably, is bowed with a resined stick of wood. The globular flute (*hoon*), mentioned as one of the very earliest artifacts of Chinese music, has been played in Korean Confucian temples since the 12th century, as has a *tchi* flute, which has a bamboo mouthpiece plugged into the blow hole with wax. In addition to five finger holes it has a cross-shaped hole in what on other flutes is the open lower end. The lower end of the *tchi* can thus be closed by the little finger of the left hand. This unique flute is known to have been in Korea by at least the 11th century and, like the previous two examples, has totally disappeared from the rest of East Asia. By contrast, the *piri* cylindrical double-reed aerophone (wind instrument) has many relatives in Asia, but the rich saxophone-like tone produced by its deceptively narrow tube body and large reed are not heard elsewhere. It is heard in many other forms of Korean music, from folk festivals to party music.

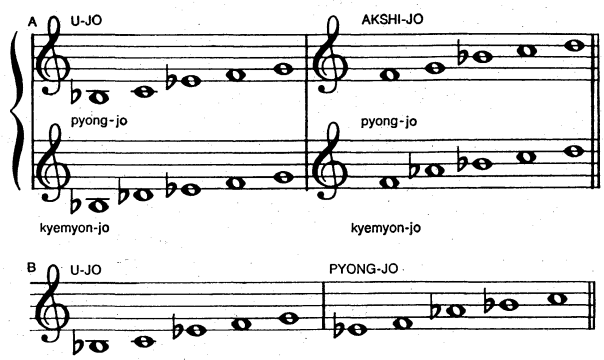
Indigenous instruments

Not all the instruments of Korea are Chinese imports. The *taekeum* flute with six finger holes, a membrane-covered hole for a buzzing sound, and open holes near the end would seem to be a Chinese instrument, except for its spectacular length of two feet five inches (74 centimetres) and its gigantic blow hole. A Korean musician, Wang San-ak, is credited with the invention in the 7th century of a *komungo* zither with six strings. Two strings on one side and one on the other have movable bridges, whereas the central three strings pass over 16 bridges. It is played by striking the strings with a wooden stick clenched in the fist with the thumb stretched along the top of the plectrum. One of the other mysterious Korean instruments is the *haekeum* two-stringed fiddle. It would seem to be an obvious relative of the Chinese equivalent, with the bow passing between the strings, except that the neck is bent toward the strings (rather than away from them, as in the rest of the world), and the pegs seem to be inserted backward, so that the strings are wrapped around the large round part of the pegs instead of the narrower end, which sticks out, unused, from the back of the neck. The *kayakeum* board zither with 12 strings and movable bridges is surprising in sound to those accustomed to Chinese and Japanese zither melodies. It is held that the instrument was created in the 6th century in the Japanese-dominated Kaya area—thus the survival of one example in the 8th-century Japanese Shōsō-in treasure house. The *kayakeum* is regarded as Korea's favourite native instrument and it can be heard in all levels of Korean music and dance. The *sanjo* variation forms for this zither represent one of Korea's most famous purely instrumental genres.

Much of what is known about the origin of instruments is derived from Chinese and Korean historical books and administrative documents, such as the grand list of presents sent by the Chinese Sung emperor to Korea in the year 1111. The list includes ten sets of stone chimes and ten bell sets, along with five iron equivalents and numerous other instruments. Korean musicians performed successfully at the Chinese court, and Korean monks attended the international training centres in China to learn Buddhist chant. During the reign of Sejong (1419–50), new Imperial shrine music and traditional Confucian ritual music were emphasized along with musical settings of epics written in the newly developed Korean alphabet. The grand traditions of China were preserved under the guidance of the court master of music, Pak Yon (1378–1458).

Proof of the diligence and concern of these and later efforts are found in the *Akhak kwebon* ("Music Handbook"), first appearing in 1493. The nine volumes of this work contain pictures of all the court instruments along with their fingerings or tunings, costumes and accessories for ritual dances, and the arrangements of dance designs and orchestral seatings. The first three volumes deal with music theory and contain ample evidence of the continuation of the complete Chinese classical tradition discussed earlier. The only differences are the pronunciations given to the Chinese characters in which the terminology is written. There are, however, additional explanations of tone systems that reflect Korean-created idioms. For example, the tone systems for Imperial shrine and banquet music are given in modes and transpositions of the pentatonic scale as shown in notation VI-A. The specific manner in which Korean pentatonics are described varies from period to period. Later theories tend to make *pyong-jo* the same mode as *u-jo* transposed a fourth higher (notation VI-B), considering the *kyemyon-jo* variants to be the other mode on those two pitches (analogous to the pattern in VI-A). The exact pitch on which these modes start also varies over the centuries.

VI



Balance of Korean and Chinese traditions

From the short theoretical discussion above it should be evident that Korean musicians maintained a balance between native and Chinese traditions. Such a balance is seen in the standard instrumentation of the three major court orchestras. By the 15th century the Chinese-style *T'ang-ak* and the Confucian *a-ak* ensembles concentrated on Chinese instruments such as bell and stone chime sets, and the texts of surviving *T'ang-ak* pieces such as *Lo-yang-chun* (*Spring in Loyang*) follow the Chinese *tz'u* poetical form. Processional military music (*chui-ta*) begins in the style seen in ancient drawings, with drums, gongs, accompanying conch shell and straight trumpets, in addition to a "barbarian" oboe with a conical body. This ensemble is followed by a softer one with the more typical Korean hourglass drum (*changko*) and cylindrical oboe (*piri*), the unusual Korean fiddle (*haekeum*) and flutes (*taekeum*). The softer ensemble can also be heard in dance music (*sam-hyon*), whereas chamber music (*chong-ak*) softens this group further by using a smaller oboe along with the later addition of the Chinese "Western" dulcimer (*yang ch'in*, or in Korean, *yang-keum*). The most famous suite of movements in this and in orchestral traditions is the *Yongsan Hoesang*, which consists of nine to 11 pieces taking some 30 minutes to play. The title is based on a former religious chant about Buddha preaching on Mt. Yongsan, but the pieces attached to this general name have since lost their vocal tradition. When the Korean ensemble (*hyang-ak*) plays pieces like *Su-je-chon* (*Long Life as Immeasurable as the Sky*), one hears a more indigenous combination of the hourglass drum, oboes, flutes, fiddles, and the special bowed zithers (*a-chaing*). Though the style is still slow and "ancient," the sounds seem less Chinese.

The survival of so many old traditions is partly due to the preservation of notation books. Many are in the traditional Chinese forms. In the late 15th century, however, a Korean mensural system (*i.e.*, a notation showing time values) was created that, through the use of columns of 16 squares, gave a clearer indication of rhythm and tempos than do most Chinese notations. This system, usually with modifications into six- or 12-square groups, is used today to notate the six-beat rhythms and can be read as easily as Western notation if one can read Korean.

VOCAL MUSIC

Vocal music is another important side of the Korean tradition. One of the longest and rarest older forms is the

kagok, in which a man and a woman sing 26 songs, each with five lines of different lengths (32 beats, 27, 37, 27, and 41). Accompaniments and interludes are provided by a small ensemble. *Sijo* is a classical three-line form of Korean poetry that also can be sung to the accompaniment of the hourglass drum. Narrative songs are found in the genre called *kasa* accompanied by a flute and drum. The dominant narrative form today is *p'ansori*. It was originally performed simply to the accompaniment of a barrel drum (*puk*), but today a small ensemble is used. In addition, the performer moves in a stylized dance form, holding a fan in a manner vaguely similar to Japanese Nō drama. It may be that this music-drama tradition emerged from ancient shamanistic entertainments of the gods, although it is now purely a classical narrative. As with so much Korean folk music and songs of the teahouse girls (*gishang*), the most striking aspect of the *p'ansori* sound is the female voice quality and melodic ornamentation. It reminds one more of traditions at the Arabic end of the long trails of Mongolian movements, such as North Africa and flamenco Spain, than it does of other East Asian traditions. Korean music is indeed an enigma but a very beautiful and exciting one.

Voice quality and ornamentation

IV. The music of Japan

A reader of the previous sections of this article will know already that Japanese music can be considered as a national tradition set in the satellite category of the general East Asian music culture. One also knows that Korea served as a bridge to Japan for many Chinese musical ideas as well as exerting influence through its own forms of court music. A comment has been made as well about the presence of north Asian tribal traditions in Japan in the form of Ainu culture surviving on Hokkaido island and the Shakalin islands. It should further be pointed out that the island isolation of Japan allowed it to develop its own special characteristics without the intense influences of the Chinese giant and the Mongols so evident in other mainland cultures. Therefore, in the ensuing discussion all the "foreign" elements will be placed in the matrix of traditions and styles that are characteristically Japanese.

MUSIC BEFORE AND THROUGH THE NARA PERIOD

Early evidences. As in the case of the history of mainland traditions, ancient Chinese sources and modern archaeological data provide the earliest surviving insights into Japanese music. Archaeologists have discovered materials of Stone Age people in Japan and pottery remains of the so-called Jōmon culture dating back as far as the 3rd millennium BC. Among the items recovered from the subsequent Yayoi period (3rd century BC–3rd century AD), the musically most significant finds are *dotaku* bronze bells. They show that the native population had adopted Chinese metallurgy. The shape of the bells and the locations of their remains indicate that they may have entered the Japanese islands with tribes migrating from north Asia. The gradual domination in Japan of one group called the Yamato clan became more evident in the Tumulus period (*c.* AD 250–*c.* 500) and led to the present Imperial system. Specific evidence of its musical life is found first in certain tomb figurines (*haniwa*), which were substitutes for the earlier Asian tradition of human sacrifices at the death of a leader. One *haniwa* has been found playing a barrel drum with a stick, while another figure is seated with a four- or five-stringed board zither across his lap. Crotal bells (pellet or jingle bells) are found on costumes, and some statues seem to be of singers. The zither is of special interest, for it is related to the Korean *kayakeum* mentioned earlier as appearing in the Japanese section of Korea (Kaya) by at least the 6th century. It also may be the earliest example of the *wagon*, or *Yamato-goto*, a six-stringed zither with movable bridges found in Japanese Shintō music. The crotal bells survive in the form of the *suzu* bell tree, an instrument characteristic of Shintō dances only. The interpretation of another figure as a singer and the presence of a drummer are rather too general for conclusions, although a Chinese history book of the 3rd century (*Wei chih*, AD 297) does speak of the natives of Japan as singing and dancing during a funeral. This source also notes two actions well-known in Shintō today: a concern for purification and the use of hand claps in praying before a shrine. The mention of shamanism is also found in Chinese accounts and is of particular interest in those concerned with the north Asian aspects of Japanese culture. In this context it must be remembered that the Ainu were as populous and strong as the new Japanese people at the time of the founding of the Yamato dynasty. Battles between the Japanese and Ainu are noted in 6th-century Chinese books such as the *Sung shu* (513) and, rather like 19th-century American Indians, Ainu were found as mercenary troops in a group of Japanese forces sent to assist the Korean Silla dynasty in the 7th century. The Chinese *Sui shu* history book (630) mentions tatooed people like the Ainu, as well as a five-stringed zither and a flute. Ainu culture today maintains a Jew's harp and no flutes, but it does have a *tonkori* zither with two to five strings. It is unlike the zither on the lap of the earlier tomb figure in both its shape and playing position, being held like a banjo and played open stringed with both hands. The surviving shamanism of the Ainu has equivalent forms in early Shintō and in a few surviving Japanese folk "mountain women" traditions. But the guttural vocal style and the frequent polyphonic textures of modern Ainu music today seem culturally to point north rather than south or west. Perhaps the Ainu are a living link between present-day civilization and the life pictured in ancient Chinese documents.

Archaeological evidence and early references

As the Japanese people gradually drove the Ainu northward they solidified their own internal structure and established stronger ties with continental culture. Records show that a Korean Silla (in Japanese, Shiragi) emperor sent 80 musicians to the funeral of a Japanese ruler in 453. Chinese Buddhism was officially accepted as a religion in Japan in the 6th century, selected converts being sent to China for proper training in the rituals (hence the music) of that faith. In 612 a Korean musician, Mimashi, is believed to have introduced masked dances and entertainments (*gigaku*) and southern Chinese music (*kuregaku*) into the Japanese court. Finally, by the 8th century Japan produced its own first written chronicles, the *Koji-ki* (713) and the *Nihon-gi* (720), which recount the mythological origin of music as the form of an entertainment used by the gods to tempt the sun goddess out of her hiding in a cave. Indirect references to music appear in semi-historical accounts of early court activities in the books. In addition, the *Nihon-gi* contains the texts of some 200 poems, many of which seemed to have been derived from the oral musical tradition.

Predominant musical traits. It is apparent that by the 8th century the documentary history of Japanese music had begun. Although this claim predates an equal state of Western music history by some 100 years, certain interesting parallels between the two traditions can be made. Both seem more clearly established in the same general 200-year period, a short time when compared to Chinese music studies outlined above. Both developed a musical nomenclature heavily influenced by the music of religious organizations: the Roman Catholic Church in the West, Buddhism in Japan. Both traditions were equally influenced by theories of a foreign culture from over the nearest sea: Greece in Italy and China in Japan. Herein many differences arise, one of the most significant being that in the Japanese case, the foreign tradition of China at the time of its first major influence was alive and strong and could apply practical musical information and instrumentations as well as theories, whereas the Greek tradition was long dead by the same period, when the European monks turned to it for guidance. Nevertheless, one can see that the general length and beginning of each history is comparable. Before discussing Japanese music in chronological detail, an attempt should be made to envision general characteristics, realizing that in doing so the tendency is to apply aphorisms to music that stretches over a series of styles as old and varied as the music of Europe from Gregorian chant through Claude Debussy. With that caveat, general guidelines for the appreciation of Japanese traditional music can be put forth.

Aesthetic and formal ideals. These guidelines fall under three general concepts: (1) the sound ideal, (2) the structural ideal, and (3) the artistic ideal; but these three things are not clearly separate in any musical event.

In general one can say that the most common sound ideal of Japanese music is to produce the maximum effect with a minimum amount of material. For example, the *taiko* drum of the Nō drama consists of a barrel-shaped body over which are lashed two cowhide heads some 20 inches in diameter stretched over iron rings. Wooden sticks are used to hit one head. Obviously, the sound potentials of the drum are many, but they are deliberately suppressed. For example, the sticks are made of very soft wood, and the strokes are applied only to a small circle of soft deerskin in the centre of the head. The *taiko*, like Japanese ink paintings, accomplishes a great deal by concentrating on very carefully chosen limitations of the medium.

Instrumental potential and tone colour

Another feature of much Japanese traditional music could be called the chamber music sound ideal. No matter how large an ensemble may be, one finds that the various instruments are set in such a way that the timbre, or tone colour, of each can be heard. This can be understood in Western chamber music and contrasts with the Western orchestral sound ideal, in which the primary intention is to merge all the instrumental sounds into one glorious colour. The colour separation of Japanese music is quite evident in the large court ensemble (*gagaku*), as well as in drama music and actual chamber ensembles such as the *sankyoku*, for koto zither, samisen plucked lute, and the end-blown *shakuhachi* flute.

The structural intents of Japanese music are as varied as those of the West, but one of special interest is the frequent application of a three-part division of a melody, a section of a piece, or an entire composition. This is in contrast with the more typical two-part division of Western music. Of course, examples of both ideals can be found in the music of both cultures; the concern here is with broad generalities. The fundamental terminology of the Japanese tripartite form is *jo-ha-kyū*, the introduction, the scatterings, and the rushing toward the end. A Western musician might wish to compare this with sonata allegro form and its three parts (exposition, development, recapitulation). But the Western example relates to a complete event and involves the development of certain motives or melodic units (such as first and second theme), whereas the Japanese concept may be applied to various segments or complete pieces that are generally through-composed (*i.e.*, with new material for each segment). Japanese music reveals its logic and its forward motion not by themes but by a movement from one section to another different one until the final section is reached. Forward motion in motivic Western music was often derived during the classical periods from the tension created by chord progressions. In Japanese music, such sonic events generally are not used. Nevertheless, the need for aurally recognizable patterns falling into a progression that the informed listener can anticipate is necessary in all music. In Japan such stereotyped patterns are melodic or rhythmic, not harmonic. They will be discussed in detail later; but the recognition, whether intellectual or aural, of the existence of such recurring patterns is essential to the appreciation of any music.

Word orientation. One of the artistic ideals of Japanese music is equally clear in all of East Asia. It is the tendency for much of the music to be word oriented either through actual sung text or through pictorial titles to instrumental pieces. With the exception of variation pieces (*danmono*) for the Japanese koto zither, one can seldom find a purely instrumental piece in the spirit of, for example, the Western sonata or symphony. Japanese ensemble pieces, like those mentioned earlier in China and Korea, are either dance pieces, instrumental versions of songs, or descriptive. This ideal in all of East Asia was not weakened until recent times, when such music was forced to compete with Western idioms.

Guilds. By the same token, the ideal of the composer as genius, so dear to 19th- and 20th-century Western hearts, had little place in earlier East Asian music. In Japan, as in China and Korea, the names of many composers are known, but the actual setting of their music was and still is often done by a group of fairly anonymous people. One may know who was helping out at a given time and in a given place; but in any written form of the music their names, or even the name of "the" composer, may often be missing. In the Orient, particularly in Japan, the performer is often the person remembered and noted. Such an ideal is understood in the West by fans of popular music. Although this ideal has given way to the Western composer "star" system in modern Japan, it does depict an important social setting for any appreciation of the older Japanese classical traditions. In keeping with this artistic ideal, one should add that often there is not one "correct" version of a given piece. Most traditional music is organized under guild systems, and thus each guild may have its "secret" version of a well-known piece. A given guild will play its version precisely the same way in each performance, for improvisation has practically no role in any of the major genres of all East Asian music. Differences are maintained between guild versions, however, in order to identify a given group's musical repertoire as separate from all the rest. A fascinating example of the tenacity of this ideal is found in the present-day Imperial court orchestra. It was formed in the late 19th century by the combination of previously separate ensembles from the courts and temples. The musicians played the same-named pieces but could not always agree on many musical details, so they continued to play their disparate versions at the same time, creating a kind of polyphony that is now accepted as the "correct" ancient tradition in Japan, although, as with the U.S. composer Charles Ives, it sometimes actually represents two things going on at once. The separation of guild styles can be carried further to one more artistic ideal, which holds that it is not just what one plays on an instrument, it is how one plays it. For example, in the case of the *taiko* drum mentioned above, the manner in which a player sits, picks up the sticks, strikes the drum, and puts the sticks away will reveal the name of the guild to which he belongs and also can be used to judge his skill in performance. No Japanese instrument is merely played. One could almost say that its performance practice is choreographed. Such distinctions exist in the music of other East Asian cultures as well, although the clues to their understanding have not yet been revealed to outside listeners and viewers. This brief discussion of their existence in Japanese music will serve to enhance the appreciation of at least one Asian tradition as the discussion turns to a chronological study of its many styles.

Manner of handling instruments

The Nara period. *Codification of court music.* The previously mentioned documents from the Nara period (710–784) demonstrate how very active music was in the newly established capital in Nara. The general term for court orchestra music, *gagaku*, is merely a Japanese pronunciation for the same characters used in China for *ya yüeh* and in Korea for *a-ak*. As Japan absorbed more and more of the outside world, the music of the court, like that of T'ang dynasty China of the same general centuries, received an increasing variety of styles. In 702 these styles were organized under a music bureau (*gagaku-ryō*) and by the early 9th century an additional *Ōuta dokoro* (the "Big Song Hall") was created for handling Japanese-composed additions to the repertoire. Among foreign genres, the musical styles of the nearby three kingdoms of Korea have already been shown to be some of the first imports, Silla music being called in Japanese *shiragi-gaku*, Paekche music, *kudaragaku*, and Koryŏ music, *kōkurigaku*. Music from these three kingdoms was sometimes called collectively *sankangaku*. Under all these terms were found still other Chinese and north Asian traditions, in addition to music purported to have come from India as early as 736. Evidence of such a distant import can be found in a surviving court dance (*bugaku*) called "Genjōraku," whose story about the exorcising of a snake can be traced to an ancient Indian Vedic tale. The date of 736 is also assumed for the entrance of music from Indochina, which survived for several centuries in a form of music called *rinyūgaku*. Al-

though this tradition is now lost, there are extant detailed pictures of the ensemble along with other ancient instruments and a variety of dances in sources such as the 14th-century copy of the 12th-century *Shinzeigakuzu* scroll.

Influence of T'ang dynasty China. The dominant musical style of early *gagaku* was, naturally, from China and was called T'ang music (*tōgaku*). In Japan, as in Korea, the establishment and maintenance of such a music has made it possible for modern listeners to hear foreign versions of famous pieces long forgotten in the country of their origin. For example, there are names of pieces played and dances performed in Japan that are also found in T'ang Chinese lists. Unlike China, however, many of these works are still played in Japan, and a few of the original costumes and masks used at that time are preserved. Perhaps the most valuable treasure in Japan for such materials from the ancient traditions of all of East Asia is the Shōsō-in, a storehouse built for the household goods of the emperor Shōmu after his death in 756. In this collection (which includes a few later additions from temples) one can find some 21 percussion instruments, 12 strings, and 12 winds, in addition to dance masks, notation, and drawings. Some of the materials are Chinese or Korean imports, while others are Japanese-made. The Chinese variant of the arched harp of the ancient Near East (in Japanese the *kugo*) is best preserved here. The very decorations of certain instruments can also be historical gold mines. For example, the protective cover across the face of one plucked lute (*biwa*) contains the picture of a performer riding a camel near a palm-treed oasis. Another such cover depicts a group of foreign (*i.e.*, not East Asian) musicians accompanying an energetic dancer, all on the back of an elephant. Etchings along a hunting bow show scenes of dancing and music performances connected with a popular imported art of acrobatics and juggling called *sangaku*.

Instruments preserved in the Shōsō-in

THE HEIAN PERIOD

Music of the left and of the right. Further images of Japanese musical life can be captured from the Heian period (794–1185). In the very first chapter of the 10th-century *Ochikubo monogatari*, Japan's earliest novel, the sad fate of the heroine is noted by the fact that she was never able to learn how to play the Chinese seven-stringed *ch'in* zither, although she did have some training in Japanese koto zither music. The famous 11th-century works, such as Murasaki Shikibu's *The Tale of Genji*, are filled with romantic koto, *biwa*, and flutes, as well as *gagaku* and *bugaku* performances and the singing of many songs. Diaries also show that the courtiers, now moved to Kyōto, found music to be a useful and frequent adjunct to their insular courtly life. It was in this period that the many forms of official court music were organized into two basic categories. The so-called music of the left was called *tōgaku* and contained the Chinese- and Indian-derived pieces. The music of the right was called *komagaku* and contained all Korean and Manchurian examples. In both categories there were pieces that by this time may have been Japanese arrangements or original compositions. The terms left and right were derived from the Confucian-based administration system of the new capital, which divided the entire government into such categories. In *bugaku* they controlled the costumes of the dancers, left dances emphasizing red, right dances, green. In *gagaku* these two major divisions standardized the instrumentation of the ensembles. When playing dance accompaniments, stringed instruments were deleted, but the two orchestras for purely instrumental performances were complete. Each used plucked 12-stringed zithers with movable bridges called *gaku-sō* or by the generic term koto. The string section was completed by a four-stringed plucked lute, the *gaku biwa*. A small hanging gong (*shōko*) and a large hanging drum (*tsuri daiko*) were found in both. The leader of a *tōgaku* piece would use a barrel drum (*kakko*) with two lashed heads struck with sticks, while a *komagaku* piece would be led by an hourglass *san no tsuzumi* drum similar to the Korean *changko*. The standard melodic instrument for both was the double-reed *hichiriki*, a *komabue* flute being added in *komagaku* and a *ryuteki* flute in *to-gaku*. The Japanese *shō* mouth organ appears in both.

Dance costume and instrumentation

Musical notation. In modern performances the *shō* plays a fascinating cluster of harmonies, although there is some feeling that the original Chinese interpretation of its notation was melodic, with little or no harmonic addition. Part of the performance problem, outside the impressive age of the music, is that *gagaku* notations came only in part books, which were often rather like Western "lead sheets"; *i.e.*, they served as memory aids rather than detailed guides. For example, the *shō* notation gives only one note for every four beats of a standard piece. The modern interpretation is that each note represents the bottom of a chord, but the notes might actually have been the skeleton of a melody. The earliest surviving form of instrumental notation in Japan is a book, dated 768, of lute (*biwa*) music found in the Shōsō-in. The earliest flute part book is dated 966, and a few additional wind or string books remain from the 10th to the 13th centuries. More frequent sources can be found from the last 300 years. Scores did not exist until modern times. The traditional part-book notations reflect the importance of oral, rote learning and the guidance of a teacher. For example, flute and *hichiriki* notations in their standard forms consist of columns first marked off by dots representing major percussion time markers (usually every four beats, though there are five- and six-beat pieces). Next, one finds a column of syllables called *soga*, which were used to help one memorize the instrumental part by singing it. With this system it was even possible to substitute a vocal rendition of one part in an ensemble if that instrument were missing. Finally, there is another parallel column that contains a skeleton of notes or fingerings on the instrument itself. Obviously one can comprehend fully the subtle ornamentations and nuances of any melody notated in this manner only through the guidance of a teacher. In the case of the string notation, one generally finds only the names of stereotyped patterns along with occasional notes. The percussion notations likewise consist of names for stereotyped patterns. If both the strings and percussion are played as written, they appear to be merely time markers, in accordance with the colotomic principle found in much East and Southeast Asian music (the demarcation of time intervals by the entrance of specific instruments in prescribed order, a procedure contributing to the musical structure and imparting a sense of progression).

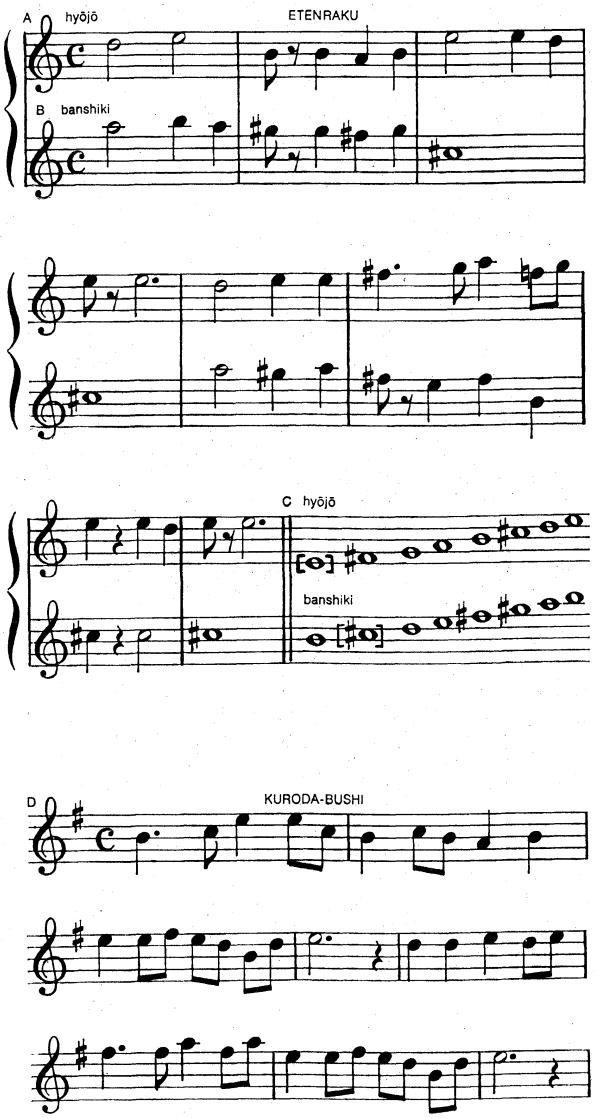
But as in jazz, there must have been more to the music hidden in the oral tradition. Further clues as to the performance practice of this music in addition to its underlying music theory and its practical uses are found in several important sources. In 735 an ambassador, Kibi Mabi, brought back from China a 10-volume *Digest of Musical Matters* (in Japanese *Gakusho yoroku*), which implies the Chinese foundation of the art. In 1233 a court dancer, Koma Chikazane, produced another 10 volumes—the *Kyokunsho*, describing Japanese *gagaku* matters. Of equal value is the *Taigen-shō*, written by a *gagaku* musician, Toyohara Sumiaki, in 1512, when court music seemed on the verge of extinction.

Tonal system. By a combination of these sources—Buddhist music-theory tomes, part books, and present-day performance practice—it is possible to understand many of the basic principles upon which ancient Japanese music was founded. From what has already been said about the beginnings of Japanese court and religious music it is not surprising to find that the complete tone system of both consists of the Chinese 12 tones shown in notation I, the only difference being the Japanese pronunciations of the characters for each pitch name. The scales in II-B show that Japanese ancient music followed the East Asian tradition as well in the use of two seven-tone scales, each with a pentatonic core. The *ryo* scale (set on C for the sake of comparison) shows no great difference from the Chinese scale in II-A; but the *ritsu* scale seems to reveal the early presence of an indigenous Japanese tonal ideal with the placement of its half steps. Japanese *gagaku* and Buddhist music theories contain most of the

Principal tonalities

classical Chinese ideas concerning transpositions and modes, but in practice the two scales shown in II-B could be constructed on only three pitches each: *ryo* on D (*ichikotsu*), G (*sōjō*), E (*taishiki*); and *ritsu* on E (*hyōjō*), A (*ōshiki*), and B (*banshiki*). Note that the pitches for such transpositions form a classic pentatonic (D–E–G–A–B). The two names for the pitch E are present in order to make a distinction between the two scales possible on that same tone. In the unaccompanied court songs and the chants of Buddhism one can observe the use of other transpositions, for all oral traditions in the world "adjust" notated pitches to the preferences of given singers. In *gagaku* instrumental music, the six tonalities are observed, part books for each instrument being organized in sections by the tonalities of the compositions. A few pieces are found in more than one tonality. A transcription of part of the basic melody for such a composition, "Etenraku," is shown in notation VII-A/B. Although set in two *ritsu* tonalities (*hyōjō* and *banshiki*; VII-C), it is obvious from this example that the piece, which is a "crossover" (*watashimono*), is more than merely transposed.

Although the scale has been transposed, yet the pitch centre of the melody also has been changed (from E to C♯). This is one of the few clear examples in performance practice of the mode systems spoken of in music theory. A glance at the related folksong "Kuroda-bushi" (VII-D; see below) and at the *in* scale (II-C) provides a preview of an emphasis on such a different mode centuries later.

VII

Vocal music. In the poetry-oriented court life of Japan, secular vocal music would obviously be important. Many of the poems in classical collections seemed originally to have been song texts. One of the oldest secular song forms is *saibara*, which was first inspired by the singing of pack-train drivers. Among the new fads of Heian period vocal music (called collectively *eikyoku*) were *rōei*, songs based on Chinese poems or imitations of them, and *imayō*, contemporary songs in Japanese. Many *gagaku* melodies were given texts to become *imayō* songs, while others were derived from the style of hymns used by Buddhist missionaries. Little of these vocal traditions remain, but memories of their importance are preserved in nearly every novel and diary of the period. For larger surviving repertoires it is necessary to turn to religious music.

Shintō music. The indigenous religion of Japan, Shintō, was closely connected with the legendary legitimacy of the emperor. Thus, special Shintō music was devised for use in Imperial shrines, a tradition already familiar from the discussion of China and Korea. In Japan such Shintō music is called *kagura*. The kind of music and ritual used exclusively in the Imperial palace grounds is called *mi-kagura*, that in large Shintō shrines, *o-kagura*, and Shintō music for local shrines, *sato-kagura*. The *suzu* bell tree, mentioned before as among the earliest known Japanese instruments, is found in all such events; and the equally ancient *wagon* zither can be heard in the palace rituals and sometimes in the larger shrines. General Shintō chanting (*norito*) is rather straightforward, whereas the surviving music of *mi-kagura* is more complex. Unison choruses of men are accompanied by the *hichiriki* oboe, a *kagura-bue* flute, the *wagon* zither, and the periodic rhythmic markings of a pair of long, thin *shaku-byoshi* clappers. The music for *mi-kagura* ceremonies is divided into two types: one to praise the spirits or seek their aid (*torimono*), the other to entertain the gods (*saibari*) in the tradition, mentioned earlier, of the mythological amusements given before the sun goddess. Perhaps the most famous surviving dance suite from the Shintō tradition is *Azuma asobi* (*The Entertainment of Eastern Japan*), which can be seen as a courtly reflection of the agricultural base of Japan in its annual performances during the spring equinox and the summer solstice. The work is said to be an imitation of the dance of a heavenly maiden who performed on the beach of Suruga in the 6th century. *Azuma asobi*, along with *bugaku* dances, may be seen at many other Imperial, national, and shrine occasions—dim but nevertheless impressive reflections of the colourful courtly life of Japan of centuries ago. *Mi-kagura* is exclusively a male event, but Shintō female dancers (*miko*) are found in other shrines. Historical documents show that the Heian court, like ancient China or, for that matter, all other affluent societies of the world, appreciated the value of female dancers and their music. In later times the Heian-originated *shirabyōshi* female dancer-musicians became important elements in the transfer of courtly and religious traditions into later theatrical forms. The major source of religious musical influence is found elsewhere in the Buddhist temples.

Buddhist music. There are many forms of Buddhist hymns, such as *saimon*, as well as semireligious dance songs, such as *goeika*, *nembutsu odori*, and the *bon odori* performed to folk festivals. But the basis of Buddhist classical music and hence the core of Buddhist influence on Japanese art music is found in the theory and practice of chanting known generically as *shōmyō*. Such a tradition came originally from foreign Buddhist missionaries and next from Japanese converts studying in China. Noted sources from the many Japanese interpretations of this tradition are the *Shōmyō yojinshu* by Tanchi (1163–1237) of the Tendai sect and the *Gyosan taikaishu* (1496) of the Shingon sect. The theoretical bases of these studies are similar to the ones already discussed under the topic of *gagaku*. Here need be added only comments about Buddhist notation systems. Most early chant notations used neumes, squiggle-like signs that, like those of the early Christian traditions, served primarily as memory aids with which an initiate could recall the de-

Systems of notation

tails of a given melody. The most influential system was the so-called *go-in hakase,* attributed to Kakui (born 1236) of the Shingon sect. Under this method the five notes of each of three octaves of a pentatonic scale were indicated by the angle of a short line, rather like the hands on a clock. Variations of this method were of great influence in the notation of all vocal music of the period and continue to be used in Buddhist chant today. It has been conjectured that the system may be based on the positions of hand *mudra* (ritual gestures) in ancient Indian Buddhism. Whatever its true origins, it remains as one of Japan's more novel innovations in the preservation of music. Of equal value is the continuing vitality of Buddhist classical chant in modern Japan.

KAMAKURA, MUROMACH, AND EDO PERIODS

Nō music. The Kamakura period (1192–1333) marks the end of Heian court splendour and the start of a new military government located in Kamakura, far away from Kyōto. In such a context it is not surprising to find the development of long narratives of military history and the flourishing of plebian theatricals. The story of the defeat of the Heike clan (the *Heike monogatari*) was known in mansions, war camps, and temple grounds primarily as sung by *biwa*-playing bards. As in the traditions of ancient Greece and Europe, these minstrels were often blind or built their style in that of the blind priest lute tradition (*moso biwa*) in which mendicant monks used to recite sutra (scriptures) from house to house or at temples. More lucrative forms of entertainment grew under the circus acts that developed out of the *san-gaku* (folk theatricals) mentioned above, its companion comic acts, *saru-gaku* (literally, monkey music), and theatricals derived from folk rice-planting dances, *den-gaku.* Street parades (*fūryū*) and Buddhist entertainments (*en-nen*) also were part of the colourful scene. By the subsequent Muromachi period (1338–1573) the terms *saru-gaku-no-Nō* and *den-gaku-no-Nō* had become the dominant terms for temple and shrine pantomime and dialogue dramas, while the comic interludes of such plays were called *kyōgen.* Through the support of the military rulers and the efforts of individual artists such as Kan-ami (1334–85) and his son, Zeami Motokiyo (1363–1443), the first major form of Japanese theatre developed. It became known eventually as Nō.

The music of Nō as it is performed today consists of vocal music (*yōkyoku*) with an instrumental ensemble known collectively as the *hayashi.* The singing is done by the actors or by a unison chorus (*jiutai*). The four instruments of the *hayashi* are a flute (*Nō-kan*), the *taiko* stick drum described earlier, a small hourglass drum (*ko tsuzumi*) held on the right shoulder and a larger one (*o tsuzumi*) placed at the left hip.

Melodic principles. The writings of Zeami, such as the *Kaden-shō,* contain terms reflecting the traditional tone systems and terminologies of former times. A distinction was made between the recitative section (*kotoba* or *serifu*) of a play and melodic parts (*fushi*). The melodies of Nō can be categorized into two basic styles, the strong (*tsuyogin*) and the lyric (*yowagin*). Their differences are most evident in the placement of fundamental tones and the use of auxiliary tones around them. In the lyric style the three basic tones (*jo, chu,* and *ge*) are a fourth apart (see notation VIII-C, below). The movement to and from each note is regulated in a manner comparable to the regulated approach to certain intervals in 16th-century Western counterpoint. Similar but different laws are applied to the strong style, the fundamental tones of which are much closer, the standard procedure today being that *jo* and *chu* are one pitch and *ge* is approximately a minor third below (VIII-B). Melodies contain various formulas and ornamentations the names of which often reflect earlier traditions in Buddhist chant and court secular songs. The choice of combinations depends on the musical needs of the given dramatic text as well as the position of the music in the general form of the piece. Once more, such restrictions in an "exotic" music should seem reasonable to students of Western art music with its needs for stylistic restraints. The notation of Nō singing (sometimes called *utai*) is derived from simpler Buddhist and early *biwa* forms that used teardrop-shaped neumes along with important pitch names to remind singers of the performance practice of a given passage. This so-called sesame-seed notation (*goma-ten*) remains basic to Nō vocal music today, and there are many detailed books in modern Japanese to help the initiate follow the music with the aid of a teacher. Variations in notation style and in the interpretation of specific passages are maintained by the various schools of Nō along with the "secret piece" tradition basic to much of Japanese traditional music since its beginnings.

Fundamental and auxiliary tones

Song types. The musical-dramatic form of Nō has as many variations as any other creative genre, such as an opera or a symphony. Table 1 shows the outline of the form of one play, "Yuma Yawata." In it one can see the manner in which the concept of *jo-ha-kyū* or tripartite form, is applied in the context of sections (*dan*) along with typical placements of Nō musical styles within such a form. The *shidai* is usually an introduction and the *na-nori* allows the first character to identify himself. The travelling song (*michiyuki*) is followed by a song emphasizing higher tones (*ageuta*). Songs with lower melodic emphasis (*sageuta*) and other styles (*issei, kuri,* and *rongi*) mix with more recitative sections (*sashi, kotoba*) and with entrances (*deha*), closings (*kiri*), and dances (*kuse* and *mai*).

Table 1: Tripartite Form Concept (Jo-Ha-Kyū) in the Nō Drama "Yumi Yawata"

sections (dan)	placement of musical styles
Jo (mae-dan)	shidai na-nori michiyuki (ageuta)
Ha	
First dan	issei sashi sageuta ageuta
Second dan	kotoba sashi ageuta kotoba
Third dan	kuri sashi kuse rongi
Kyū (ato-dan)	ageuta deha sashi issei mai rongi kiri

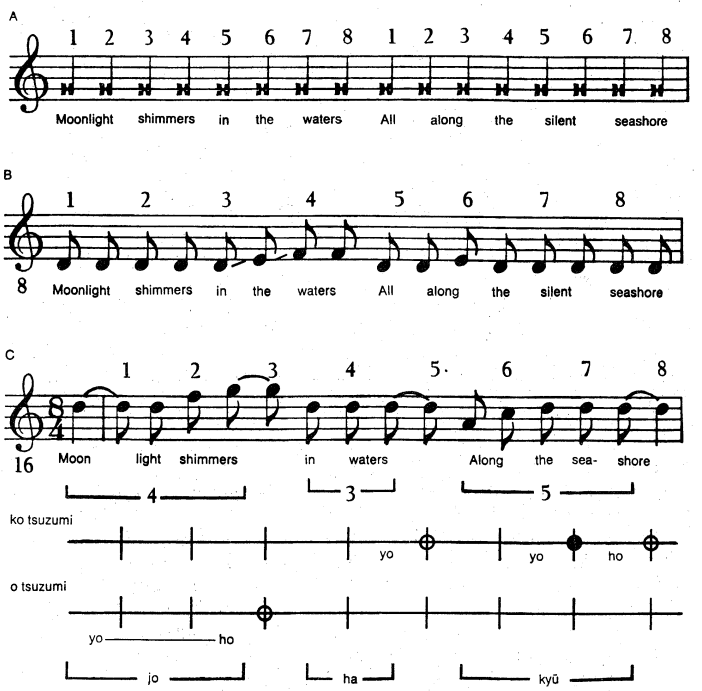
Setting of texts

Within each of these sections and subsections one must remember that drama and text have their influence as well. Though the drama is not all poetic, the earlier discussion of Chinese theatre should prepare one for the fact that many lines of the text are indeed actually poems or are influenced by poetic form. Although there is great variety in the syllable lengths and combinations, the most common divisions in Nō are into lines of seven or five syllables. Such a 12-syllable line has been constructed in English in notation VIII along with a 16-syllable variant so that the reader can compare certain basic principles in the settings of Japanese Nō texts. The placement of a text rhythmically can be done in three major ways: *o-nori,* with one syllable per beat (VIII-A); *chu-nori,* with two syllables per beat (VIII-B); and *hira-nori,* in which 12 syllables are worked into an eight-beat frame (VIII-C). The setting shown in VIII-A is in the spoken (*kotoba*) style, VIII-B, in the strong (*tsuyogin*) singing, and VIII-C is lyric (*yowagin*).

Note in VIII-C that, although there are two sections of seven and five syllables, the actual division of their presentation (4 + 3 + 5) is tripartite; *i.e.,* in the form of *jo, ha,* and *kyū.* Lines of greater numbers of syllables in *hira-nori* require more beats, and 12-syllable lines themselves can be handled in many other ways. Thus, one must not take this artificial example to be representative of the only style of Nō text setting any more than one can consider a given eight-beat phrase from one Mozart opera as being sufficient to show all there is to aria form. Nevertheless, the settings given in notation VIII show in their syllable displacements and melodic contours that such "Oriental" music, like that of all cultures, does have a thorough internal logic. Nō music shares with Western art music the extra convenience of a complete written music theory, of which this short example demonstrates only a few elementary principles. The principles may be taken one step further by constructing a typical drum accompaniment for the *hira-nori* version of the texts.

Function of drum patterns. The lower portion of VIII-C represents one possible setting of the text by the

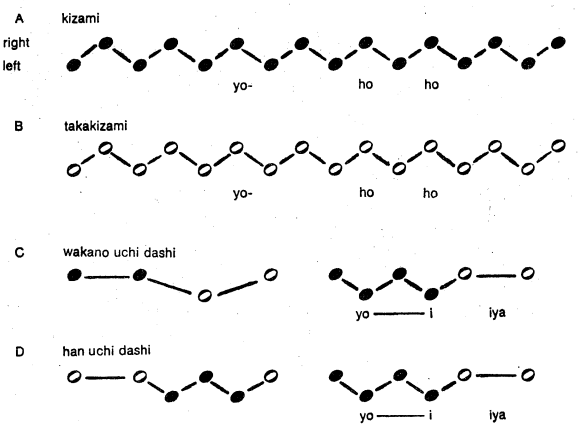
VIII



music of the two *tsuzumi* drums. The music of the Nō drums consists of a series of named, stereotyped patterns that are aurally perceivable and that tend to progress in given orders. The pattern supporting the vocal line in VIII-C is called in both drum parts *mitsuji*, although they do not always play a pattern with the same name. The circles represent moments in which the drum is struck, and the words are drum calls (*kakegoe*) uttered by the drummers. These calls are as essential to the performance and recognition of a given pattern as are the drum sounds themselves. They help to give each pattern a unique aural image. Also, the manner in which the calls are performed by the player helps to signal and control the timing of each beat in a music that is often very elastic in rhythm. Note in addition that, in VIII-C, the first call of each drummer builds up to his first actual striking of the drum, which in turn marks the divisions between the three parts; *i.e.*, the *jo, ha,* and *kyū* phrasing of the vocal line. In short, the vocal line and text have been "rhythmicized" in a manner analogous to the way in which Western music might harmonize a given moment in a song. There are no chords in most Japanese music, but this example shows one way in which many of the basic functions of harmony are handled in a totally parallel yet sonically different

Rhythmic patterns and listener expectation

IX



way. This point may be taken one step further by showing (in notation IX) a set of named patterns for the *taiko* stick drum, an instrument used only in dance sections of a play. Placement of the dots shows right- and left-hand strokes; black dots indicate the softer—and light dots the louder—strokes. The few patterns (*tetsuke*) in the example are from a set of 59 found in a *taiko* instruction book. A study of them shows that they belong to families (*tegumi;* the *kizami* family in A and B and the *uchi dashi* group in C and D). Once more the principles of harmony in Western music come to mind; *i.e.*, a C-major and C-minor chord are related because they share some common aural traits (the tones C and G in this case). If one looks into a lesson book of *taiko* music one will find that, as in many Western harmony books, the student will be told what patterns may appear before or after each item. In Western music one learns in a similar way that certain chords may be preceded and followed by others. In both the Japanese and Western cases there is a selection of permissible choices of earlier or following events for each pattern. Thus, it would seem that the concept of prediction and anticipation is fundamental to the listener's sense of logical progression in the music of both. The major difference in this case is that the aurally perceivable, named, stereotyped pattern of Western traditional music is a vertical sonority called a chord, whereas in Nō music it is a horizontal time unit called a *tetsuke.*

Role of the flute. The music of the Nō flute (*Nō-kan*) has many stereotyped patterns, but it functions in rather different ways from the drums of the *hayashi.* Although it seems originally to have related to the vocal line, today it does not play the singer's melodies. The Nō flute looks like the *ryuteki* flute of *gagaku*; but a short cylinder has been inserted inside its tube so that the upper holes of the flute overblow a seventh (as A–G) rather than the octave, as with all other flutes. This unique characteristic of the flute seems to reflect the tonal principles of Nō mentioned earlier, whose basic notes outline a seventh (A–D–G in notation VIII-C). The flute may give a specific pitch to set the tonality of a vocal entrance, but its normal functions today are to signal sections of the form and to play one of the dozen standard dance pieces used in various Nō dramas. The dance pieces are set in sections (*dan*). Each piece is learned by rote or by a notation consisting of flute mnemonics placed in a frame of eight squares representing beats in a manner rather like the Korean notation mentioned earlier. The flute's seven holes are fingered by the middle joint of the fingers instead of the tips, producing an impressively fluid melody that would not fit into the graphic notation system of traditional Western music. At the same time, it does not compete with the Nō vocal line by being too melodically clear. Indeed, a discussion of the Nō flute seems to be a particularly appropriate finale in this brief survey of Nō music (*Nō-gaku*); for it shows that the form of each musical item—whether performance practice, pitch relations, structure, or notation—follows the needs of its functions. The music of Nō drama seems on first hearing to be one of the most puzzling of East Asia's exotic sounds, but a study of its principles can make it become as reasonable and as beautiful as a Bach cantata.

Koto music. *Schools and genres.* The 13-stringed zither with movable bridges called the koto has been mentioned as one of the basic instruments of the court ensembles as well as a common cultural accoutrement for court ladies. The development of independent solo and chamber music genres for this instrument becomes more evident as one moves into the Muromachi period. One of the earliest schools of solo koto music is the *Tsukushi-goto,* supposedly founded by a priest, Kenjun, in the latter half of the 16th century on the southern island of Kyushu. Its repertoire was first based on the *imayō* song tradition and thus seems to reflect the presence of refugees from the court who fled south during the difficult civil wars. Sets of songs accompanied by the koto and sometimes by the three-stringed plucked lute, or samisen (*shamisen* in Tokyo dialect), were part of the tradition. They were called *kumiuta,* a term applied to much of the early chamber music that developed in subsequent years. The *Tsukushi-goto* school was restricted to priests and Confucian scholars. But Hosui, a samisenist and student of Kenjun on koto, took the tradition to Kyōto in the early 17th century, combined it with more popular music, and

Roots in earlier song traditions

taught it to more lowly people, such as blind men, as he was himself. In Kyōto he founded a new school, *Yatsuhashi-goto*, which mixed the older repertoire with many new compositions. The title *Yatsuhashi* was adopted later by another apparently unrelated school to the far south in the Ryukyu Islands.

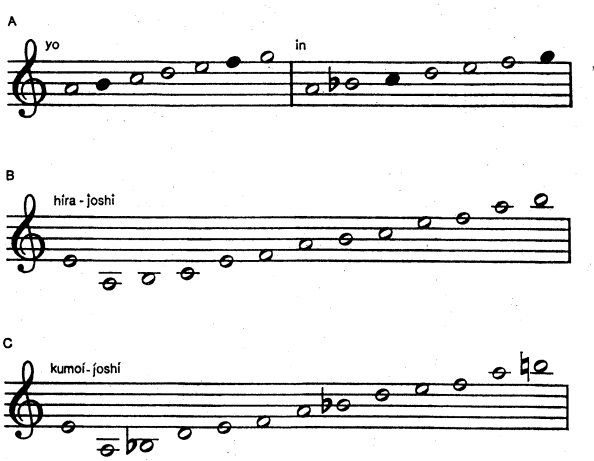
The new schools of popular, or "vulgar," koto (*zokuso*) are proper reflections of the mercantile life of the new Edo period (1603–1867). In 1695, a student of one of Hosui's students took the professional name Ikuta Kengyō and founded another school, the Ikuta. The term *kengyō* had been one of the basic ranks of musicians under the guild system and so is frequently found in professional names, but the name Ikuta remained as one of the primary sources of koto music until the creation of still another school by Yamada Kengyō (1757–1817). In present-day Japan, the Ikuta and Yamada schools remain popular, whereas the earlier traditions have faded considerably. Both schools have provided famous composers; and there are several pieces from their schools, as well as a few earlier works, that are now shared by the guilds as part of the classical repertoire of the koto. The slightly longer and narrower shape of the Ikuta koto produces a tone easily distinguishable from that of the Yamada school.

Koto music is known in general as *sokyoku*. In the koto solo instrumental music (*shirabemono*), the most important type is the *danmono*, a variation piece in several sections (*dan*), each normally of 104-beat length. The term for koto chamber music, *san-kyoku*, means music for three. The standard instrumentation today consists of a koto player who also sings, along with performers on a three-stringed plucked samisen lute and an end-blown *shakuhachi* flute. In earlier times a bowed variant of the samisen called the *kokyū* was used more often than the flute. The basic genre of chamber music is called *jiuta* and combines the earlier *kumiuta* tradition of accompanied song with instrumental music by alternating sections with singing (*uta*) and instrumental interludes (*tegoto*). After the 19th century a second embellishing koto part (*dana-wase*) often was added to the instrumental interludes. Twentieth-century innovations will be covered later.

Tunings and notation. Each school of koto music from the courtly tradition to the present time involves changes in the structure of the instruments, and changes in playing method and notation. The ancient court koto (*gaku-so*) is similar to the modern koto and is played with picks (*tsume*) on the thumb and first two fingers of the right hand or with bare fingers, although, unlike the Ikuta and Yamada styles, the left hand is not used to alter the tone by pressing the string on the other side of the movable bridges. Its notation consists primarily of the names of basic patterns in addition to occasional melodic fragments and the text. The survival of such music is dependent on a continuing viable rote tradition; thus, most of the tradition is lost.

The tunings of the 13 strings of the court koto were derived from the modes of the *ryo* and *ritsu* scales of the earlier periods. The tunings used in the Edo koto traditions, however, reveal hitherto new, apparently indigenous, tonal systems. These concepts were eventually categorized under the two scales called *yo* and *in*, shown in notation X-A. The tunings in X-B and X-C reflect the new kinds of pentatonism of the period with their use of half steps. The *hira-joshi* tuning appears in such famous early works as *Rokudan* (*Six Dans*) ascribed to Yatsuhashi Kengyō, the "founder" of the modern koto styles. In all, there are some 13 standard tunings for the koto and many variants. Like all the other popular music in Japan from the 17th century on, these koto tunings are based either on the older tradition preserved in part in the *yo* form or on the more "modern" *in* scale. One can note in the 19th century occasional pieces deliberately written in the previous *gagaku* mode style as well as the use of the Holland tuning (*oranda-choshi*) that the Western major scale derived from the Dutch business area on Deshima in Nagasaki. Nevertheless, the *yo–in* system remains the fundamental tonal source for new Japanese music from the 17th century on, exceptions being revived court music, new Nō plays, and the work of avant-garde composers after World War II.

X



The earliest printed notations of koto, samisen, and flute pieces from the Edo period are found in the *Shichiku shoshinshu* (1664), the *Shichiku taizen* (1685), and the *Matsu no ha* (1703). Although many sections of such collections contain only the texts of songs, one can find certain pieces that parallel the line of words with numbers representing strings on the koto or finger positions on the samisen, names of stereotyped koto patterns, or mnemonics for the particular instrument with which the piece is learned. In the late 18th century both the koto and the samisen traditions developed more visually accurate notations. The koto version (first seen in the *Sokyoku taisho*, 1779) used various sized dots to indicate rhythm. In the early 19th century, string numbers were placed in columns of squares representing rhythm, as in the system mentioned earlier in Korea. The numbers and squares eventually were combined with the $\frac{2}{4}$ bar-line concept of the West; so that the notations of both schools today, although separate systems, maintain a balance of traditional and Western ideas. Their modern compositions attempt to do the same as well; but before they can be treated, attention must be given to the traditions connected with the other major instruments of the Edo period.

Printed collections

Schools of shakuhachi flute music. The *shakuhachi* end-blown flute is a variant of the Chinese *hsiao*, and examples of it can be found in the famous 8th-century Shōsō-in storehouse mentioned earlier. During the Muromachi period (1338–1573), a smaller Japanese version called the *hitoyogiri* became popular as a solo instrument, but the best known form of the *shakuhachi* is the one developed in the Edo period. The instrument was used by *komusō*, priests who begged or sometimes spied while wandering through the streets playing the flute incognito, their heads covered by a special wicker basket hat. With the changes in Edo society, many former warriors no longer carried their swords, whereas young merchants carried more money. One curious side effect of such changes was the occasional appearance of a *shakuhachi* tucked in the back of one's belt for use as a musical device or as a club.

The major schools of *shakuhachi* music today come from guilds, the Meian and Kinko, whose origins derive from two sects of an earlier Fukeshu guild of *komusō* priests. The Meiji era (1868–1911) abolished the monopoly rights of the various Edo music guilds; and a Tozan school was founded for teaching the music to amateur musicians, a custom soon adopted by the other guilds. The instruments of all schools may vary in size and the number of finger holes for the purpose of pitch as well as differences in timbre ideals. The standard *shakuhachi* has four finger holes along the front and one thumb hole behind. A bell is formed by the bamboo root stems at the end of the flute. The mouthpiece is cut obliquely outward, and a small piece of bone or ivory (*a tsu*) is inserted at the blowing edge in order to help produce the

great subtle variety of tones typical of *shakuhachi* music. The basic repertoires of the music are divided into three types. Original pieces (*honkyoku*) are those claimed to be composed by the founders or early teachers of a given school, whereas outside pieces (*gaikyoku*) are taken from other genres or other schools of *shakuhachi* music. New pieces (*shinkyoku*) continually appear and are kept in that category. *Shakuhachi* notation varies with each school, but all are based on mnemonics with which the music is taught. Given the exceptional subtlety of tone changes and ornamentation in all traditional *shakuhachi* music, such a notation system seems quite logical. The beautiful introverted sounds of *shakuhachi* music seem closer to Buddhist chant than to other instrumental forms and are best learned by the ear and heart rather than the eye and brain.

Samisen music. The three-stringed plucked lute of Japan is known as the *shamisen* in the Tokyo area or the samisen in the Kansai district around Kyōto. It seems to have arrived in Japan as an import of the *sanshin*, or jamisen, from the Ryukyu Islands in the mid-16th century. The Ryukyu form of the instrument, with its oval body and snakeskin, is obviously derived in turn from the Chinese *san-hsien*. Such an origin is reinforced by collections of early Ryukyu music, which use a so-called *kukunshi* notation similar to the Chinese symbols shown in notation III. The Japanese samisen underwent considerable physical change, its body being rectangular and the skins coming from a cat or dog. Apparently under the influence of contemporary *biwa* lute traditions, the plectrum of the instrument was changed from the talon-like pick of the Ryukyus to a wooden *bachi* with a thin striking edge. In addition, the lowest string was kept off the small metal upper bridge near the pegbox so that it produced a buzzing sound (*sawari*) reminiscent of the tone of a *biwa*. The three basic tunings of the Japanese instrument are *hon-chōshi* (b–e′–b′; b = the B below middle C, b′ the B above); *ni agar* (b–f♯′–b′); and *san sagari* (b–e′–a′). They have remained standard to the present day, although there are occasional variants.

Greater variety is found in the many genres of samisen music. The earliest types seem to have been played by old *biwa* entertainers around Ōsaka, a city then called Naniwa; hence the name of the new genre was naniwa-bushi. Samisen was used for folk music and party songs, but, in keeping with the *biwa* origin of the first performers, narrative music was of prime importance. Such music became known as *jōruri*, the term derived from the title of a famous story of the princess Lapis Lazuli (*Jōrurihime monogatari*). As different guilds of samisen evolved it was possible in modern times to divide them into two basic styles: narrative traditions (*katari-mono*) and basically lyrical musics (*uta-mono*). Table 2 is an outline of the development of these two styles in terms of genre names. *Sekkyō* was an earlier form of Buddhist ballad drama for the general populace and thus is placed at the beginning of the narrative style, for *sekkyō-bushi* was eventually done with samisen accompaniment. The term *jiuta* has already been mentioned as one of the early chamber music forms and thus starts the lyrical list. Turning to the narrative list first, one finds a mass of names, most of which after *naniwa* and *jōruri* are derived from the professional name of the musician who began the style. Except for the term *ogie* and *utazawa*, the names for the lyrical styles are more descriptive. It has already been noted that *kumiuta* means a set of songs. The terms *hauta* and *kouta* stand for short lyrical pieces such as would be heard in a teahouse or at a banquet. *Nagauta* means a long song and represents the major genre in this category, which will be described presently. Each of the styles listed in Table 2 uses a samisen of different size with different weight bridges and design of plectrums. The voice quality of the singers is quite different as well. For example, a professional *shinnai* singer would find the performance of *gidayū* as difficult as would a French opera specialist attempting to sing Wagner. The most famous and perhaps most demanding of the narrative styles is *gidayū*, named after Takemoto Gidayū (1651–1714), who worked with Chikamatsu Monzaemon in the founding of the most popular puppet-theatre tradition (known as *bunraku*) of Ōsaka. The *gidayū* samisen and its plectrum are the largest of the samisen family, and the singer-narrator is required to speak all the roles of the play, as well as to sing all of the meditations and commentaries on the action. The part is so melodramatic and vocally taxing that the performers are often changed halfway through a scene. There is little notated in the books (*maruhon*) of the tradition except the words and the names of certain appropriate stereotyped samisen responses. The samisen player must know the entire drama by heart in order to respond correctly to the interpretations of the text by the singer. The two musicians sit on a platform to the stage left of the theatre and through the intensity and skill of their performance help bring life and pathos into the wooden characters who move with frighteningly realistic gestures in the hands of three puppeteers. The power of *gidayū* is such that it can be heard in concert versions as well. In the 19th century, a school of female performers (*onna-jōruri*) carried on the concert tradition with equal ability.

Jōruri, or narrative, music

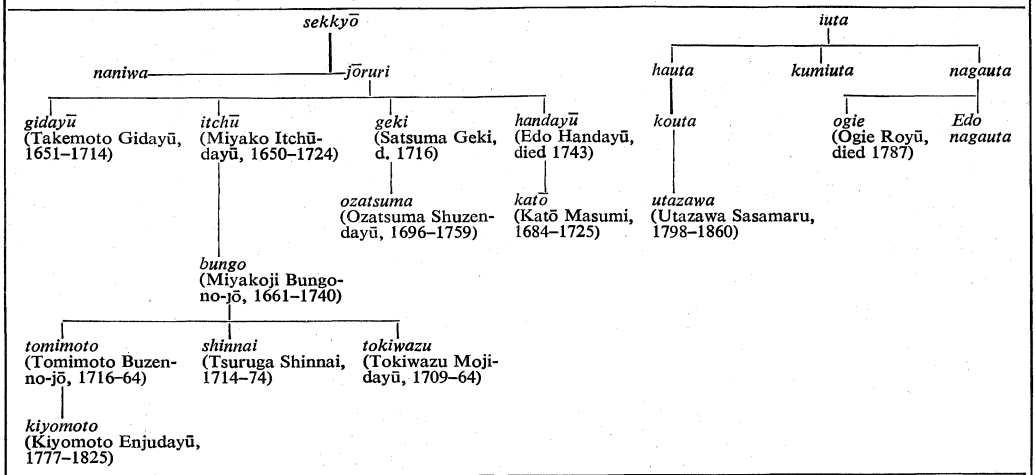
Music for the puppet theatre

Kabuki theatre. The *nagauta* form of lyric music, like most of the other narrative forms, began with a close relation to the Kabuki popular theatre of the Edo period. The first Kabuki performances used instruments (*hayashi*) from the Nō drama. Because Kabuki was related to the flourishing demimonde of the major cities, however, the music of the party houses and brothels was soon added to the theatre. By the mid-17th century the names of *nagauta* singers and samisen players were listed on posters along with the cast. In the same manner, the names of musicians in many of the other genres listed in Table 2 were adopted to denote parts of a play. Although nearly all of the music listed can be heard in concert forms today, the major genres still included in Kabuki productions are *gidayū*, *tokiwazu*, and *kiyomoto* from the narrative styles, and *nagauta* from the lyrical. Rather than discussing them individually they will be viewed in the total theatrical context and later brief reference will be made to their concert forms.

Onstage music. Kabuki as theatre is discussed elsewhere in the encyclopædia. Its musical events can be divided into onstage activities (*debayashi*) and offstage groups (*geza*). In plays derived from puppet dramas, the *gidayū* musicians, called here the *chobo*, are placed on their traditional platform offstage left or behind a curtained alcove above the stage-left exit. If other genres are used, the performers are placed about the stage according to the scenery needs of the play. There are some plays in which several different kinds of onstage music are required, a situation called *kake-ai*. The most common dance scene today, however, is one in which the onstage group consists of *nagauta* musicians and the Nō *hayashi*. The samisen and singers are placed on a riser at the back of the stage, and the *hayashi* sit before them on floor level —thus, their other name of the *shitakata*, meaning "the ones below." There are as many different types of dances that require different kinds of music as there are in Chinese or Western opera. In a general view, perhaps the most intriguing side of this variety is the relation of the older drum and flute parts to the Edo-period vocal and samisen melodies. In totally Kabuki-style pieces, the *tsuzumi* drums play a style called *chirikara* after the mnemonics with which the part is learned. The patterns of this style follow closely the rhythm of the samisen part. If the Nō flute is used as well, it is restricted to cadence signals; or a simple bamboo flute (*take-bue* or *shino-bue*) is substituted, which plays an ornamented (*ashirai*) version of the tune. There are many sections, however, in which the drum patterns and Nō flute melodies discussed earlier are combined with samisen melodies. In a classical repertoire of hundreds of set pieces, there are many different combinations, but to many listeners these situations seem rather puzzling at first hearing, with apparently two kinds of music going on at the same time. If the situation is from a play derived from a former Nō drama and uses the full *hayashi*, one notes first that the flute is not in the same tonality as the samisen nor is it playing the same tune. The drums in turn do not seem to relate

Tension between simultaneous styles

Table 2: Genres of Shamisen Music*



*Names of founders of some styles are given in parentheses.

rhythmically to the melody, as they do in the *chirikara* style. If one analyzes a given situation in detail it is found that the drums and flute are playing named stereotyped patterns normally of eight-beat length as in the Nō. The essential difference between them and the samisen melody is that they do not seem aurally to have the same first beat. Many times a given samisen melody will make room through silence for an important vocal call in the drum patterns, but the deliberate lack of coordination of beat "one" creates a vital rhythmic tension that makes the music drive forward until resolve is accomplished at a common cadence. This important compositional device can be called a slide-rule effect: each part is internally rigid and progressive, but its conflict with the other parts forces the music (and the listener) to move the musical event through a time continuum toward a mutual completion. The Nō flute music is frequently related to the *taiko* stick-drum rhythm, so that they can be considered as a common unit rather than separate parts. There are situations in which the *tsuzumi* play *chirikara* patterns in support of the samisen melody, while the *taiko* and Nō flute play either Nō patterns or later Kabuki-named drum patterns "out of synchronization" with the other music. At such moments one can see that in Kabuki dance music, as in Western Classical music, there are three kinds of musical needs. In the West they are melody, rhythm, and harmony. In this music they are melody, rhythm, and a third unit of one drum and a flute that functions like harmony although its sound is totally different. If this third Japanese feature is called the dynamism unit, then it can be said that *nagauta* dynamism and Western traditional harmony both serve to colour the line, to create tension that drives the music onward and help to standardize the formal design of the piece by clarifying cadences or by creating the need for them. All of this brings back the earlier point that music is not an international language. The equally logical but different aspect of this music is certainly most obvious and striking.

The formal aspects of Kabuki music are as varied as the plays with which music is connected. In dance pieces derived from Nō plays, many of the sectional terms of the Nō mentioned above are found. The classical Kabuki dance form itself often consists of sections divided into the traditional tripartite arrangement as shown below:

deha or *jo*
oki michiyuki
chūha or *ha*
kudoki, monogatari, odori ji
iriha or *kyu*
chirashi, dangire.

Generally speaking, the *oki* represents all kinds of introductory instrumental sections (*aigata*, or in this case *maebiki*) or vocal parts (*maeuta*) before the entrance of the dancer. The *michiyuki* usually incorporates the percussion section as the dancer enters. The term *kudoki* is found in the early history of samisen music as a form of romantic music and is used here for the most lyrical section, in which the percussion is seldom heard. The *monogatari* (story) relates to the specific plot of the dance, and the *odori ji* is the main dance section, rather like the *kuse* or *mai* of the previous Nō form. During this section, the bamboo flute may appear for contrast and, in Nō style, the *taiko* drum may be important. The *chirashi* contains more active music, and the final cadence occurs during the *dangire*. There are endless variations and extensions of this form, but the many specific instrumentational and stylistic traits found in each of these sections help the listener to become aware of the logical and necessary progression of a given piece through a moment of time to its proper ending.

Changes in notation in the 18th century

Most early collections (*shōhon*) of onstage music consisted of the text and samisen mnemonics (*kuchi-jamisen*, mouth samisen) of instrumental interludes (*ai-no-te*). In the 18th century some of the lyrical forms began to use syllables to represent fingering positions on the instrument, a system called the *iroha-fu*. In 1762 a set of circles with various extra markings along with the string number were combined in a book called the *Ongyoku chikaragusa* to create a more accurate if complicated system. Further rhythmic refinements were created in the 1828 *Genkyoku taishinsho*, but it was not until the modern period that Arabic numbers in the French *chevé* style (apparently learned in Germany by Tanaka Shōhei) were combined with Western rhythmic and measured devices to create notations that could be sight-read without the aid of a teacher. Three variations on this technique form the basis of most modern samisen notations, although occasional pieces can be found in Western notation as well. Thus, it is possible to purchase large repertoires of *nagauta, kouta*, or *kiyomoto* music for performance or study alone. Motivating such notational changes was the increased interest during the mid-19th century in samisen music composed for concert performance (*ozashiki*) rather than as dance accompaniment. Such a tradition is common practice for all the samisen genres today.

Music and dramatic content

Offstage music. Returning to the theatre, one finds rather different music offstage. This *geza*, or *kagebayashi* (shadow *hayashi*), music is normally placed in a small room on stage right with a view of the drama through a bamboo curtain. The music consists of special samisen and vocal pieces and a great variety of percussion signals. For example, a huge *o daiko* barrel drum with two tacked heads signals the beginning of a program, in keeping with the sounds given by the same drum from a tower over the entrance of very early Kabuki

theatres. Other drums, bells, gongs, and clappers are used to reinforce stage action, and special offstage songs may set the mood or location of a scene, particularly in those scenes in which onstage musicians do not appear. For example, the singing of the offstage song "Eight Miles to Hakone" will tell an audience that the scene is set along the old Tokaido road, whereas the sound of waves (*nami-no-oto*) beat on the *o daiko* drum indicates that the scene is on the road near the sea. A type of offstage song called *meriyasu* may be used to reflect the silent thoughts of the stage character, while the call and response of occasional beats offstage of two *ko tsuzumi* drums will place a scene in a mountain area with its echoes. As in Western musical theatre and films, many of the sounds are naturalistic, whereas others are traditional means of evoking desired responses from an audience. Theatregoers in both traditions are often unaware consciously of the means used for such reactions even though familiarity has made their dramatic value very real indeed. The specific musical devices used in a given Kabuki play are under the control of a headman, *hayashi gashira*, who works with the first samisenist, the actors, and the director to produce the desired results. Thus, the musical contents of a given play may change with different productions. In Kabuki, the combination of offstage and onstage music creates a total atmosphere that has few parallels in other world theatres. Perhaps it comes as close as anything to the composer Richard Wagner's ideal of the all-embracing art form (*Gesamtkunstwerk*).

Biwa, vocal, and folk music. During the late 19th century the *biwa*-accompanied narratives enjoyed a revival. The blind-priest *biwa* (*moso biwa*) tradition had originally been divided into two schools named after the provinces in Kyushu from which they came, Chikuzen and Satsuma. The tradition declined greatly over the years. When the Imperial restoration began in the Meiji Era (1868–1912), many members of the new administration were from those provinces. Thus new schools of narrative *biwa* music arose under those two names, influenced at this time by several samisen narrative traditions. The topics of the new *biwa* pieces were often military and appropriate to the modernization period. The 19th century also was one of Japan's periodic revivals of interest in things Chinese, reduced somewhat with the advent of the Sino-Japanese War of 1894–95. Another late Edo period style was *shigin*, the singing of Chinese poems in an intense solo style quite unrelated to the Heian *rōei* tradition of Chinese-based songs. *Shigin* was later accompanied by *shakuhachi*, and during the increased military spirit of the Meiji era it was combined with a posturing sword dance, *tsurugi-mai*. It also appeared in *biwa* concerts and could still be heard on rare occasions in the early 1970s.

Work songs, theatricals, and festivals

Courtly writings have left little information about the music of the peasants in any detail, but some folk songs and theatricals of the Edo period remain for modern study. The rice-planting, harvesting, and other work songs that survive may retain ancient melodies and may also be evidence of the indigenous origins of the *yo–in* scale systems to which most such music belongs. In this context, the first phrase of the folk song "Kuroda-bushi" is shown in notation VII-D as it is said to have been derived originally from a Heian period *imayō* based on the *gagaku* piece shown in VII-A. Most folk songs are, of course, regionally functional but historically vague and subject to the normal changes of any oral tradition. Viewing as a whole both the performance practice and voice qualities of Japanese folk music, one finds a great variety of styles. Such richness may reflect the long periods of Japanese feudalism, which fostered many different musical dialects.

The many processionals and pantomimes of folk theatricals are accompanied by flutes and percussion, the generic term for such ensembles being *hayashi*. During the Edo period, the Shintō shrines of Tokyo developed festival ensembles (*matsuri bayashi*) for the various major districts of the city. Most of these combine a bamboo flute with two folk-style *taiko* stick drums, an *o daiko* barrel drum, and a small hand gong called the *kane*, or *atarigane*. When such groups are playing general festival music, they all use a suite of five pieces: *yatai, shoden, kamakura, shichome*, and another *yatai*. But their versions of each piece can be very different. When dance or pantomime is involved, the *sato-kagura* music implied earlier is used. There are in addition many forms of folk festival dances (such as the *bon odori*) that either have a *hayashi* ensemble, are unaccompanied, or, since the Edo period, use a samisen. Given the oral base of all folk music, many songs are lost with the demise of another old farmer or worker. There has been a considerable revival of interest in national songs in Japan since World War II, however, and there are hopeful signs that such colourful traditions will survive.

THE MEIJI PERIOD AND SUBSEQUENT MUSIC

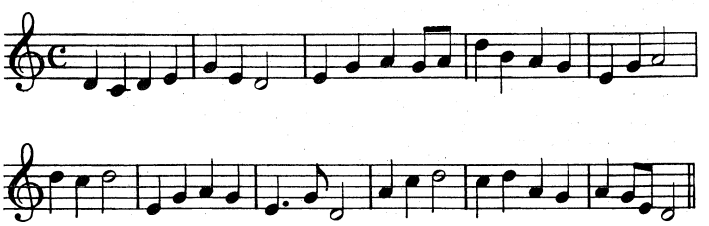
Sources of Western influence. The period of Japanese history after 1868 is often thought of primarily in terms of its Westernization. The three major sources of Western music in Japan were the church, the schools, and the military.

Religious and military music. Christian music had, in fact, been introduced into Japan as early as the mid-16th century with the arrival of Portuguese merchants and Roman Catholic priests. With this importation came Catholic music and Western musical instruments, the most lasting of which was the double-reed shawm, which survives today as the tuneful accessory of itinerant noodle sellers. The bowed *rebeca* lute may have combined with the Chinese *hu-ch'in* in the creation of the bowed *kokyū* of 17th-century Japan. But the suppression of Christianity in that century destroyed the bamboo organs, choirs of mass singers, and most of the other direct Western musical imitations until the Meiji restoration. The official doctrine of new religious freedom in 1872 brought large numbers of Protestant missionaries into action, and collections of hymns with Japanese text were printed by 1878. Interdenominational editions were necessary by the 1890s. Since that time, standard Catholic and Protestant musical activities can be found and, with the international growth of Tokyo, one can even add the sounds of synagogues and a mosque. But the growth of musical acculturation in Meiji Japan is better seen in its other foreign imports.

Band music, as part of a military table of organization, had already been tried in Dutch style at a military school in Nagasaki during the early 19th century. After Adm. Matthew C. Perry's arrival in 1853, every foreign delegation to Japan did its best to impress the natives with marching bands (Perry added a minstrel show). Thus, the various Japanese regional and national military leaders were quick to add such organizations to their modernized armies. The Emperor was equally aware of the Western musical values displayed by the first foreign missions and ordered that the *gagaku* musicians be trained in band music as well. A navy band from the Satsuma clan gave the first Japanese public performance of this new music at the opening of the railroad in 1872, and in 1876 *gagaku* musicians made their debut as band musicians on the occasion of the Emperor's birthday. The training of the many new ensembles was in the hands of English, French, and German bandmasters, and new music was created by them or by their Japanese students to match the spirit of Meiji modernism. The most famous case is the national anthem, "Kimi ga yo," which was one of the few successful early attempts at combining Western and Japanese traditions. A British bandmaster, William Fenton, teaching the Japanese navy band, worked together with *gagaku* musicians through several unsuccessful versions; and the search continued through his German successor Franz Eckert. A court musician, Hayashi Hiromori (1831–96), is credited with the melody shown in notation XI, which was given its premiere in 1880 and has remained the national anthem since that time. Hayashi first wrote it in traditional *gagaku* notation; and Eckert "corrected" it with Western harmonization, noting that it fit in both a *gagaku* mode (*ichikotsu*) and one from the Western church tradition (Dorian). As Japan's military prowess grew, standard Western-style marches and patriotic pieces dominated the repertoire. They also influ-

Collaboration of Western and traditional Japanese musicians

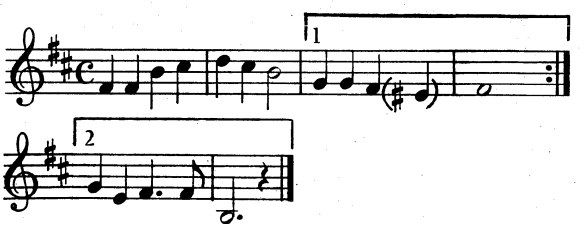
enced popular music with such genre as *rappa-bushi* (literally, "bugle songs") as well as music in the schools.

XI

Music education. Public-school music in Japan was organized by a member of a Meiji educational search team, Izawa Shūji (1851–1917), and a Boston music teacher, Luther Whiting Mason (1828–96). Mason was brought to Japan in 1880 to help form a music curriculum for public schools and start a teacher-training program. Though there was much talk of combining the best of East and West, the results of the sincere efforts of an American late-Victorian and a Japanese bureaucrat were less than glorious. The first children's songbook, the *Shogaku shokashi shoden* (1881), contained either Western pieces with Japanese words or songs newly composed by Mason. The primary sources of Western tunes were those pieces from Boston school books that appeared to be pentatonic. Through this method songs like "The Bluebells of Scotland" spoke of beauty ("Utsukushiki"), "Auld Lang Syne" concerned fireflies, and Stephen Foster became the major composer of songs known to educated Japanese children. The newly composed songs with their artificial tunes and moralistic words quickly faded away and eventually were replaced by more popular children's school songs based on military music (*gunka*) from the Sino- and Russo-Japanese wars. The teacher-training school became the Tokyo Music School by 1890 and included instruction in koto and, because of the lack of proper violins, the bowed *kokyu*. The music department of the modern University of Fine Arts is still located at the spot of the original school in Ueno Park, Tokyo, with a bust of Beethoven beside the entrance. Koto, samisen, Nō music, and Japanese music history are now found there along with extensive offerings in Western music. But until very recent times music education was totally Western in orientation. Japanese music was presented in middle-school music appreciation courses only some ten years after World War II. The teaching of Western-style singing and the use of choruses have become fundamental to a proper education in Japan, with the results that youth and workers' choruses of the 20th century are cut off from original Japanese music. It was only with the rise all over the world in the mid-20th century of searches for cultural or ethnic identities that the Western nature of Japanese music education has been bypassed by some youth. Such a move should be quite clear to followers of Euro-American folk- and minority-group music revivals. In Japan, 20th-century activists, right or left, have attracted youth by the use of the public-school choral tradition in new textual contexts. Behind the robust volume of such functional, harmonized tunes lies the equally viable if quieter sounds of older, traditional music.

Traditional styles. The pre-Meiji period of 19th-century Japanese traditional music, known generically as *hōgaku* vis-à-vis Western music (*yōgaku*), was generally strong. It has been noted that certain styles of samisen music had been able to create concert repertoires disconnected from dance or party accompaniment. Koto teachers and composers also flourished; and *biwa* music began to return along with court music, paralleling the restoration of Imperial power. The most devastating effect of the restoration was the cancelling of monopoly privileges previously held by the various guilds, including those in the music fields. This temporary economic–social setback was overcome by admitting students from all classes of people and, at the same time, by a concerted effort on the part of more imaginative musicians to make some compromise between their old traditions and the new sounds flowing in from the West. In general, the evaluation of Western music by Japanese traditionalists showed that it differed from *hōgaku* in the following ways: it used other tone systems; it was thicker in texture, with more high and low notes going on at the same time; part of this thickness was sets of chords; it was generally considered better if it were faster and louder and the instruments were played more fancifully; it used more instruments at a time; it used different kinds of metres; and it had other forms, often organized by the concept of first and secondary themes. A survey of late-19th-century and early-20th-century musical experiments in Japan shows that every one of these characteristics was tried out, particularly in koto and samisen music. Perhaps the most obvious and successful composer in the new traditional music (*shin hōgaku*) following World War I was Miyagi Michio (1894–1956), a blind koto teacher in the Ikuta school. In 1921 he composed a piece "ochiba no odori" ("Dance of the Falling Leaves"), which used two koto, samisen, and a 17-stringed bass koto of his invention. Later works by Miyagi combine orchestras of traditional instruments, sometimes with strikingly successful results, although concertos for koto by some composers, with their mass koto and *shakuhachi* accompaniments, rather negate the entire sound ideal of the original idioms. The 1929 duet for *shakuhachi* and koto, "Haru-no-umi" ("Spring Sea"), has proven Baroque-like in its performance practice, for it is often heard played by the violin, with koto or piano accompaniment. Its style equals the French composer Claude Debussy in his most "orientale" moments. The Japanese traditionalist's view of Western music listed above continued to be employed after World War II with such works as multimovement pieces using mixed orchestras in other contemporary idioms, including electronic manipulations. Such trends are best seen in the context of Western-style Japanese composers.

Evaluation of Western music by Japanese traditionalists

Composers in Western styles. Although graduates of the Tokyo Music School and modernized court musicians were involved in many of the first concerts and compositions in Western classical music, the major Japanese forces in this direction came from young men who studied in Europe. The most famous surviving composition of this era is *Kojo-no-tsuki* (*The Ruined Castle*), written in 1901 by Taki Rentarō after his training in Germany. The first line, shown in notation XII, reveals, with its use of E or E♯, a conflict between the Western minor and the Japanese *in* scales. In its piano-accompanied version it recalls the style of Franz Schubert, and as sung in the streets it sounds Japanese. Yamada Kōsaku was training in Germany when the Meiji Era ended (1912) and returned to Japan with a new, more composer-like name, Koscak, and a strong interest in the founding of opera companies and symphony orchestras, as well as in the teaching of Western music. His famous 1940 opera, *Kurobune* (*The Black Ships*), deals with the opening of Japan to the West and reflects his knowledge of Wagnerian style. Attempts at nationalistic operas can be represented better by the 1952 work *Yuzuru* (*Twilight Crane*) by Ikuma Dan. The plot is a Japanese folktale, and, although the musical style is a mixture of the music of Maurice Ravel and the late works of Giacomo Puccini, one finds as well deliberate uses of folk songs and idioms. Shimizu Osamu is perhaps more successful nationalistically in his choral settings of Japanese and Ainu music, in which the style of vocal production and chordal references seems to be a more honest abstraction of Japanese ideals. Mamiya Michio combined

XII

traditional timbres with 12-tone compositional technique in a koto quartet. Mayuzumi Toshirō has produced many clever eclectic results in such works as his *Nirvana Symphony* (1958); Buddhist sutra texts mix with a combination of choral writing in the style of Igor Stravinsky, orchestral tone clusters, and sweeping vocal lines derived from Japanese Buddhist chant style.

20th-century musical internationalism

It has often been felt that no true combination of Japanese and Western music would be possible until there was some composer who was equally knowledgeable in both Western and Japanese traditional styles. Until recent times such a musical, aesthetic barricade seemed unbroken. But the international musical styles of the last third of the 20th century seem to make culturally transcendental eclecticism a viable medium for those composers with enough talent and insight to control the infinite idioms available to them. In Japan, Takemitsu Toru seems a likely candidate for such an accolade. His music is totally contemporary and never directly "orientale," yet some of his senses of timing, texture, and structure are characteristically Japanese.

In modern Japan all styles of music are available, from the traditional to the most avant-garde. Fully professional performances of Kabuki music are matched by complete Beethoven symphonic series. Huge choruses singing polemics of every type and mass bands of children bowing violins in the widely imitated method of instruction developed by Suzuki compete for audiences with intimate recitals of Heike *biwa* music and hundreds of other events. Research in Japanese traditional music has flourished among native scholars as well as among an increasing number of foreign devotees. National, private, and academic organizations have been founded for the collection, study, and publication of material dealing with all aspects of Japanese musical life; and the commercial record companies periodically enter into prestige contests that result in stunning, well-documented albums concerning nearly every type of Japanese music from Buddhist chant and folk songs to electronic music and prewar popular tunes.

From the outline of Japanese musical culture given above it should be evident that old traditions can still be heard along with the newer ones. For the most part, the older forms probably do not sound the same today as they did in their heyday. Such changes in traditions are inevitable, however, and are common to music in most other world cultures, including the Western. For example, present day *gagaku* performances are undoubtedly different from those of 1,000 years ago, but Mozart symphonies as well do not sound the same as they did in the 18th century. Now modern technology has made it possible to "freeze" a given performance of some musical event through a recording. But each musician in each generation may choose as he desires to add fresh flavour to such earlier items or leave them "pure." Part of the charm and fascination of Japanese music is that it still offers so many stylistic listening and studying choices to anyone curious or energetic enough to want to know them better. A major point of this entire article is that none of the various styles of Oriental music is any more mystical or incomprehensible than is Bach or Beethoven. Each tradition is simply different. All of them are also logical and—perhaps of greater importance—they are beautiful to those who learn their special forms of musical language.

BIBLIOGRAPHY

Music of China: For Western-language sources see FREDRIC LIEBERMAN, *Chinese Music: An Annotated Bibliography* (1970). Excerpts from basic historical Chinese texts are in *Chung-kuo ku-tai yin-yueh shih-liao chi-yao* (1962). See also RULAN CHAO PIAN, *Song Dynasty Musical Sources and Their Interpretation* (1967); ROBERT H. VAN GULIK, *The Lore of the Chinese Lute*, rev. ed. (1969); L.E.R. PICKEN, "The Music of Far Eastern Asia," in *The New Oxford History of Music*, vol. 1 (1954); FRITZ KUTTNER, "The Music of China," *Ethnomusicology*, 8:121–127 (1964); and "A Musicological Interpretation of the Twelve Liis in China's Traditional Tone System," *ibid.*, 9:22–38 (1965); and DANIEL S.P. YANG, "Censorship: 8 Model Works," *The Drama Review*, vol. 15, no. 3 (1971).

Music of Korea: The major sources for Korean music study are in Korean or other Asian languages. LEE HYE-KU, the major scholar, has provided a few items in English such as *Korean Classical Music Instruments* (n.d.) and *Topics in Korean Music* (1967). A mimeographed *Glossary of Korean Music* by CHANG SA HOON was printed by the Korean Radio Management Bureau, Ministry for State Council Administration (1961). The Korean Broadcasting Corporation (KBC) has produced several excellent LP records with English notes of value. See also DONALD SUR, "Notes on the Yongsan Hoesang Source," in *Essays in Ethnomusicology* (1969); and ALAN HEYMAN, "Pansori," in *Essays on Asian Music and Theater* (1971) and his "Comparative, Critical Study of Korean Traditional Music Today," *Korea Journal*, vol. 2, no. 9 (1962); and BANG-SONG SONG, *An Annotated Bibliography of Korean Music* (1971).

Music of Japan: General introductions are FRANCIS T. PIGGOTT, *The Music and Musical Instruments of Japan* (1909, reprinted 1971); WILLIAM P. MALM, *Japanese Music and Musical Instruments* (1959); and SHIGEO KISHIBE, *The Traditional Music of Japan* (1966, Nippon Victor record album JL 32–4). Studies of special topics are HANS ECKARDT, *Das Kokonchomonshu* (1956); TOYOTAKA KOMIYA, *Japanese Music and Drama in the Meiji Era* (1956); WILLIAM P. MALM, "The Rhythmic Orientation of Two Drums in the Japanese Nō Drama," *Ethnomusicology*, vol. 2, no. 3 (1958), and *Nagauta: The Heart of Kabuki Music* (1963); ETA HARICH-SCHNEIDER, *Roei: The Medieval Court Songs of Japan* (1965) and *A History of Japanese Music* (1972); ELIZABETH MAY, *The Influence of the Meiji Period on Japanese Children's Music* (1963); and WILLEM ADRIAANSZ, "A Japanese Procrustean Bed: A Study of the Development of *Danmono*," *Journal of the American Musicological Society*, vol. 23, no. 1 (1970).

(W.P.M.)

Music, Recording of

Because music evaporates as soon as it is produced, men, seeking permanence in life's ephemera, have long sought ways to record and reproduce it.

Man's efforts to capture the fleeting sounds of music have followed two basic methods: that of symbols and that of signals. The former—musical notation—matured earlier, and in one form or another it virtually monopolized the recording of music for centuries; the latter had to await the emergence of technology for its development. In notation, symbols are written down as a message to a trained performing musician who understands them and reinterprets them into sound. Signals, on the other hand—being direct physical impressions of, and potential stimuli to, sounds—bypass the performer in their reproduction and, in some electronic compositions, even in their recording. This essay will concern itself solely with the latter, nonsymbolic, method. Information on the former method will be found in the article MUSICAL NOTATION. For a longer discussion of the technical aspects of the reproduction of sound, the reader is referred to SOUND RECORDING AND REPRODUCING.

Types of reproduction. The physical reproduction of music has been accomplished in three major ways, which can be designated the mechanical, the acoustical, and the electrical. In the mechanical, an automatic instrument, such as the barrel organ, plays music that has been built, or programmed, into the mechanism by the designer; the resulting sound is that of the apparatus. In the acoustical and electrical methods of reproduction, sonic vibrations themselves are captured in performance and reproduced—by purely mechanical means in the acoustical method and by the use of vacuum tubes, transistors, and other such devices in the electrical. In both cases, the resulting sound is expected to be that of the independent performance.

Mechanical, acoustical, and electrical methods

Until the end of the 19th century, music was reproduced primarily by means of the mechanical method. There are reports of other methods, probably based on the action of wind or forced air, dating as far back as about 1500 BC, in the 18th dynasty of ancient Egypt, when a colossal statue of the god Memnon at Thebes made some sort of sounds to greet his mother, the goddess of the dawn. (Toppled by an earthquake in the year 27, the statue seems to have lost this ability upon reconstruction.) Friar Roger Bacon is reported to have invented some sort of talking head in the Middle Ages, and Josef Faber created in Vienna in 1860 a

talking man with ivory reeds for vocal cords, a rubber tongue and lips, and with a keyboard that altered the mouth cavity to control word formation. The most common technique, however, called for a human hand or clockwork to turn a cylinder embedded with pins that would strike or otherwise operate some sound-producing apparatus, such as the metal teeth of a music-box comb; the hammers, quills, or pipes of a keyboard instrument; or the clappers of a set of bells. Automatic carillons are known from the 1300s, automatic harpsichords and organs, from the 1500s. King Henry VIII of England owned an automatic virginal; his daughter Queen Elizabeth I in 1593 sent the sultan of Turkey an elaborate musical clock. Every six hours it played a tune on 16 chimes, followed by a two-trumpet tantara, then by an organ tune and performance by "a holly bushe full of birds and thrushes, which at the end of the musick did singe and shake theire winges." In the 19th century, Queen Victoria of Great Britain owned a bustle that would play "God Save the Queen" when she sat down.

Some of the most illustrious composers in the history of music wrote for mechanical devices. Haydn wrote tunes for musical (pipe organ) clocks; Mozart wrote several pieces for mechanical organ; and Beethoven wrote his *Wellington's Victory* (or *Battle Symphony*) for the panharmonicon, a full mechanical orchestra invented by Johann Nepomuk Maelzel (Mälzel), a German musician who specialized in mechanical musical devices of this type.

At the end of the 19th century, two inventions superseded the barrel-and-pin mechanism. One was the player piano, which used a perforated cardboard roll to control a stream of air that activated the piano's hammers. This had the advantage of enabling pianists to record their performances for future playback, and many virtuosos and composers took advantage of this device, among them Ignacy Jan Paderewski, Edvard Grieg, Claude Debussy, and Sergey Rachmaninoff. Some composers, including Igor Stravinsky and Paul Hindemith, wrote music especially for the piano roll, using devices such as combinations of as many as 30 notes played simultaneously; while impossible for two hands, such chords could readily be played by the perforated paper.

The player piano

The second invention, which was to make obsolete all previous music-reproducing apparatus (except in toys and cuckoo clocks), was the phonograph.

The impact of recording. *On composition.* In 1967 a survey of hundreds of American composers indicated that they were almost unanimous in regarding the recordings of their works as being more important than either printed publication or live performances. Through recordings, composers gained not only an easy familiarity with the music of others but also a new medium for their own works.

The contemporary American composer and teacher Milton Babbitt, in a conversation in 1965 with the Canadian pianist Glenn Gould (who maintained his own reputation largely by means of records and broadcasts, rather than by concert performances), said:

> We have all been affected as composers, as teachers, as musicians by recordings to an extent that cannot possibly be calculated as yet. . . . I don't think one can possibly exaggerate the extent to which the climate of music today is determined by the fact that the total Webern is available on records, that the total Schoenberg is becoming available.

The use of the record as a medium had superficial beginnings as early as 1904 in Ruggero Leoncavallo's song "Mattinata," specifically written for the record according to the label. Later, in 1925, Igor Stravinsky composed a piano piece, *Serenade in A Major,* expressly for the record medium, though it is also perfectly capable of being performed live. Ottorino Respighi's *Pines of Rome* (1924) incorporates a recording of a nightingale's song in its third movement. Much more important use of recording as a medium occurred toward midcentury, in works fundamentally relying on recorded tape, such as Edgard Varèse's *Poème électronique,* an 11-channel tape played through 425 speakers at the 1958 Brussels World Fair, and Morton Subotnick's *Silver Apples of the Moon* (1967), an electronic work playable only as a recording.

On teaching. In music education the phonograph was early adopted as a tool in teaching both serious students and laymen. Teachers who could not adequately illustrate musical examples at the piano found in records a means of demonstration. They could also bring entire orchestras into the classrooms by means of the phonograph.

In 1930 the *Columbia History of Music by Ear and Eye,* a phonographic survey that became popular in music history classes, enabled many students—as well as many of their teachers—to hear for the first time such instruments as viols, lutes, virginals, clavichords, and harpsichords, together with the then little-known music written for them. A half dozen years later another educational recorded project, *L'Anthologie sonore,* added impetus to this specialized field. By the 1960s, the Baroque music of the 17th and 18th centuries—as well as the earlier music of the Renaissance and medieval era—increasingly was recorded in performances using the instruments for which it was written. Such music found a wide audience beyond educational institutions; this audience was developed in large part by the phonograph.

Recorded histories of music

By the 1970s many conservatories, colleges, universities, and even some secondary schools had constructed recording studios to enable students to analyze their own performances or to rehear their own compositions.

On criticism. Records also enabled music critics to expand their knowledge, and for the better critics their perspective, of music and performance practices. Unfortunately, a record collection also allows reviewers to write on superficial differences between performances with very little expenditure of intellectual energy. New music has suffered especially from the resulting loss of the ability of many critics to expostulate on music for their readers.

On concerts. The impact of recordings on the concert hall has also been enormous, both for classical and for popular performances. Performers today can hardly hope to attract a concert audience if they have not produced distinguished recordings; usually, their audiences, both at home and abroad, consist of persons who know the performers' work through recordings. In the popular music field especially, many performers cannot compete in live appearances with recordings in which they depend heavily on technical aid. There are some who feel that the phonograph may cause the demise of the concert hall, which, if it survives at all, will do so for social rather than musical reasons. A possible indication of this trend is the disappearance of independent nonacademic nonprofessional classical music magazines in America; instead, there are record magazines. Their name changes are significant. *The American Record Guide,* established in 1935 as *The American Music Lover*; *High Fidelity,* established in 1951 (in 1965 it incorporated the last major concert–opera trade magazine, *Musical America,* into a combined edition); and *Stereo Review,* established in 1958 as *Hi Fi/Music Review.* The record magazine is not a peculiarly American phenomenon; England has *The Gramophone* and *Records and Recording*; France, *Diapason Microsillon* and *Harmonie*; Germany, *HiFi-Stereophonie* and *Fono Forum*; Italy, *Discoteca*; Holland, *Luister*; Belgium, *La Revue des Disques*; Sweden, *Musikrevy;* and Japan, *Record Geijutsu,* among many others.

On musicology. The entire field of comparative musicology—*i.e.,* the study of the relationships between Western, non-Western, and primitive music—depends upon disc and tape recordings. Although the discipline may be traced to the 18th century, it did not emerge from a primitive state until it acquired phonographic tools. Primitive music is generally transmitted orally rather than through a written tradition, and as such its performance practices—certainly in rhythm and intonation—cannot be accurately transcribed into Western notation. Since World War II, anthropologists and musicologists have visited the most remote parts of the world with tape machines to record aboriginal music before it was either tainted or wiped out by Western civilization. The most recent studies have been conducted as a race against time or more specifically against the transistor radio, a ubiquitous commodity that is homogenizing the world's musics.

The role of the producer. Although the record producer has at times become an equal partner with the musicians in creating the recorded performance of classical music, in the popular field he is frequently in total command. Here, in fact, the sounds produced by the musicians may simply be the raw material for the producer to work with; artificial sounds, overlays of sound upon sound, electronically introduced reverberation, multichannel effects with directional interplay and moving instruments, all may serve as vital ingredients of the recording. Paradoxically, as technological advancement brought the recording beyond the mere imitation of live performances, popular musicians began to bring complex electronic equipment into the concert halls to imitate the sounds of their recordings.

In productions of classical music, serious thought is given to whether the recording should faithfully capture the performance as heard from the optimum position in the concert hall or studio or whether the recording setup should be used to "enhance" the performance. Few question any longer the common practice of correcting actual mistakes. Ever since magnetic tape made detailed editing possible, extra takes have been made of sections in which musical problems are evident. The best taping of each section is spliced into a master tape. Even in recordings made during an actual concert, performers sometimes return to the hall afterward to emend any blemishes. The improvements in recorded performances made possible by tape splicing, however, often mislead audiences into anticipating the same perfection in live performances. Also, although tape editing facilitates the excision of poor passages that, while acceptable in the heat of a concert, would become irritating upon repeated hearings, it also has been said to hamper the continuity of the performance. It is unlikely, however, that a listener can spot the rare movement that has required no splicing from the majority that have. This alleged lack of continuity, however, was much worse when music had to be recorded in five-minute segments, for recordings at 78 revolutions per minute.

Microphone placement

Microphone placement has been perhaps the major criterion in separating the "natural" or "re-creative" from the "creative" technique of large-scale classical recordings. In a natural setup, microphones are placed in the optimum positions in the hall—often directly over the conductor—in order to re-create the concert-hall or opera-house effect. In the natural arrangement, the conductor is responsible for instrumental and vocal balances.

Some record-producing companies prefer to put microphones closer to the performers—the so-called close-miking technique. Here the record producer—generally with the final approval of the conductor or leader—is responsible for balances, for bringing out particular instrumental or vocal lines, in other words, for participating in the interpretation. Studio-made popular recordings —other than those of a lush semiclassical nature—have generally used the close-miking technique; in some cases, each performer in a small musical group is assigned his own microphone. In a close-miked symphonic recording session, 18 microphones may be used: three for violins; one for cellos and basses (sometimes one for each); one each for woodwinds, brasses, timpani, snare drum and triangle, bass drum and cymbals, celesta or harp, and soloist; and from three to six for a chorus. Several separate recordings, or "tracks," each comprising the inputs of several microphones, generally are made at the same time, and the producer must balance the strength of these various inputs during the recording session. Until about 1960, two-track machines were ordinarily used; by 1970 eight-track recorders were in use, allowing much more subtle mixing of channels during the editing sessions subsequent to the actual recording. For popular music sessions, 16-track recorders are sometimes used. For stereophony, all the recording tracks must be edited and mixed to make the final two channels.

The record producer also determines the degree of separation between those two channels, and during a dramatic recording—an opera, for example—he may function as stage director in guiding the performers around the aural stage. In quadraphony (quadriphony), which has four channels, the controversy between natural and close-miked recordings persists; in classical music, when the two rear channels are used mainly for hall ambience, the arguments centre on the placement of the two front channels. Some companies, however, have begun to use the four channels as equal partners even in the classics. Columbia Records pioneered in this technique by placing the conductors Leonard Bernstein or Pierre Boulez in the middle of the orchestra, which is seated for optimum quadraphonic array rather than for optimum concert-hall effect. The introduction of quadraphonic discs in the early '70s found several systems competing for prominence, most notably Columbia's SQ disc, Japan Victor Company's CD-4 disc (RCA's Quadradisc in the U.S.), and Sansui's RM (also called QS) disc.

The development of musical recording. In 1877 the U.S. inventor Thomas Alva Edison heard "Mary had a little lamb" emanate from a machine into which he had just spoken the ditty. It was the first time a recording of the human voice had been reproduced, and the event signalled the birth of the phonograph. There was, to be sure, a long gestation period.

Edison sent representatives, machines, and cylinders to Europe almost as soon as he had invented the phonograph, and between 1888 and 1894 recordings were made by such notables as Alfred Tennyson, Robert Browning, and even Johannes Brahms, who played a Hungarian rhapsody. The first "celebrity" recording, however, was made in Edison's West Orange, N.J., laboratories when the Polish pianist Josef Hofmann, then a 12-year-old prodigy, paid a visit to Edison's studio in 1888. The German musician Hans von Bülow followed shortly after with a recording of a Chopin mazurka on the piano.

The first "celebrity" recordings

In 1894, Charles and Émile Pathé built a small phonograph factory in a suburb of Paris and began to record singers as eminent as Mary Garden. Within a decade their catalog boasted some 12,000 items, and their name became almost synonymous with the cylinder phonograph in Europe. Meanwhile, Emile Berliner, a German immigrant living in Washington, D.C., had filed a patent in 1887 for a "Gramophone," using a disc rather than a cylinder, and he began manufacturing Gramophones and discs in 1894. The disc had the commercial advantage of being more easily manufactured than the cylinders. One of his representatives established a branch in London, the Gramophone Company (in 1898), a branch in Berlin, Deutsche Grammophon A.G., and one in France, the Compagnie Français du Gramophone, while Berliner's brother set up a disc-pressing facility in Hamburg. Most of Europe's recording industry thus was started by Berliner's representatives, and in the United States the small Berliner organization was to turn into the giant Victor company.

By the beginning of the 20th century, recording industries had been established in Germany, Austria, Russia, and Spain. Much of the managerial and technical talent, not to mention equipment, had been imported from America. (By 1970, the positions would be reversed, with Europe in command of a major portion of the American market.)

During the 1890s, recordings had become popular primarily through coin-in-the-slot phonographs in public places. Talent was incidental to the novelty of the apparatus; most of the recordings were of whistlers, bands, comic numbers, ditties, ethnic routines, and the like. In the first years of the 20th century, Victor and its affiliates raised cultural expectations with its Red Seal series (Red Label in Europe), particularly with discs made, beginning in 1902, by the Italian tenor Enrico Caruso. By 1910, the vast majority of record sales—some estimates are as high as 85 percent—were classical.

The Red Label had been initiated in 1901 in Russia with some of the first ten-inch disc recordings made, and the basso Fyodor Chaliapin was among the first artists to record on the new Russian Red Label.

In 1902, Victor and another major label, Columbia, decided to help the development of the new industry by pooling their patents. Victor was thereby legally able to record on wax (which would then be electroplated) for

the first time, and the new wax discs were then used in recording Caruso in Milan. Caruso's discs were a major catalyst in transforming the amusing gadget of a phonograph into a respected cultural phenomenon. That same year the new series received London-made recordings by stars of the Covent Garden opera house, primarily through the efforts of the Gramophone Company's music director, Landon Ronald, a bona fide serious musician and conductor who was able to convince his colleagues of the musical worth of the Gramophone. One instrumentalist also appeared in the new Red Label series, the Czech violinist Jan Kubelík.

In America, Columbia followed suit in 1903 with its ten-inch Grand Opera Records, recording the Metropolitan Opera stars. Shortly after, Victor began its own celebrity recording sessions of opera stars and others on 3½-minute 12-inch discs. Victor also made many of its associate European companies Red Label recordings—which included Mary Garden singing music by Debussy with the composer at the piano—available in the U.S. on its Red Seal series. Columbia soon dropped its opera series when the recordings did not sell as well as songs and marches, but Victor saw an institutional value in the celebrity recordings. The prestige of the Red Seal rubbed off on Victor's other products: "Victrola" became, in the popular mind, almost a generic term for the (disc) phonograph, and the company practically monopolized the quality-minded market for many years. Indeed, the total Western Hemisphere record market became virtually monopolized by Victor and Columbia, while their London affiliates controlled the rest of the world. The first major break did not come until World War I, when ties were severed with the Deutsche Grammophon of Germany, which emerged after the war as the independent Deutsche Grammophon Gesellschaft (DGG).

Between 1907 and 1910 Columbia tried to approach Victor's cultural prominence by releasing records from Europe, later by reinstating its own recording sessions with operatic singers. Columbia also began issuing double-sided discs, as had already been done in Europe. Victor did not do so until 1923.

During the early days of recording, both the cylinder and the disc were produced acoustically rather than electronically. A singer would sing into a horn, and his accompanist behind him played a piano placed on a platform so that the rear of the instrument—with the back removed—would also be level with the horn. With the development of a sound box to be placed on violins and violas, small orchestras could be used as accompaniment, but bassoons were required to play the cello part and a tuba the double bass part. It was an event worthy of a London newspaper announcement in 1904 when the Czech violinist Jan Kubelík made two records with his own Stradivarius, rather than on a violin with the sound box. When symphonic recordings came to be made, the wind and brass instruments still played or doubled the parts written for the lower strings, which could not be reproduced adequately. Although acoustical recordings were improved by the 1920s, the problems were not overcome until the introduction of the microphone and the consequent electrical recording process around 1925.

Birth of a mass medium

In the decade 1910–20, the phonograph became a truly mass medium for popular music, and recordings of large-scale orchestral works and other classical instrumental music proliferated. The rise of the popular record coincided in the U.S. with the new ragtime, popularized by Scott Joplin's rags at the turn of the century and sensationalized by Irving Berlin's "Alexander's Ragtime Band," written in 1910, which swept the country the following year. It stimulated an unprecedented dance craze at a time when the phonograph was becoming increasingly available. As the fad spread to millions who had never danced before, phonographs were sold to people who had never owned records before. Between 1914 and 1919, phonograph sales increased more than fivefold. In 1917, Victor issued the first jazz recordings, by the Original Dixieland Jass (sic) Band, but few major jazz releases appeared before the 1920s.

The first large-scale symphonic recording, Tchaikovsky's *Nutcracker Suite*, was issued in 1909 in England. The first attempt to record a concerto came the following year, also in London, when Wilhelm Backhaus recorded a cut version of the first movement of Grieg's *Piano Concerto*. In 1913, the first complete symphonies, Beethoven's *Fifth* and *Sixth*, were recorded in Germany under an anonymous conductor; later that year, a celebrity conductor, Arthur Nikisch, led a full-scale symphony for the first time, again the Beethoven *Fifth*. Solo instrumentalists vied with the opera singers for the record-buying public's affection, mainly by recording tidbits. In 1917, Victor began to record with a combination that was to prove its star classical music attraction for decades: Leopold Stokowski and the Philadelphia Orchestra.

The serious recording of serious music became a growing phenomenon as the phonograph matured during the early 1920s. Tidbits and orchestral snatches gave way to a spate of uncut symphonies, sonatas, quartets, and concertos; the music itself came to mean as much as its star performers, and the electrical recording process, from 1925 on, raised the quality of the recordings as well. But the mid-'20s also saw the rise of radio, which provided free music, and this new factor, plus the worldwide economic depression of the '30s, threw the phonograph industry into serious decline.

The companies realigned themselves. In 1929 the Radio Corporation of America (RCA) bought Victor. Later that year, Edison left the phonograph business he had created. In 1931 the English Columbia Graphophone Company divested itself of its near-defunct American progenitor and joined the Gramophone Company to form Electric and Musical Industries, Ltd. (EMI), bringing into the merger nearly every important European firm except Deutsche Grammophon Gesellschaft (DGG) and its export label, Polydor. American Columbia was revived by its purchase, in 1938, by the Columbia Broadcasting System.

During the 1930s, as the American companies relied mainly on dance records in jukeboxes (*e.g.*, coin-operated devices in public dance halls, taverns, and the like) to satisfy a dwindled market, Europe supplied a slow but steady trickle of classical recordings. In 1931 the HMV label in Great Britain began its "Society" issues: a limited public was asked to subscribe in advance to then esoteric releases—the complete Beethoven piano sonatas played by Artur Schnabel or Pablo Casals performing the Bach unaccompanied cello suites. A new British company, Decca, organized in 1929, also began to issue serious recordings. In the United States, Columbia began to record a number of distinguished orchestras, including those of New York, Chicago, Cleveland, and Minneapolis. RCA retained its leadership, however, with the Philadelphia Orchestra under Eugene Ormandy, the Boston Symphony Orchestra under Serge Koussevitzky, and—perhaps the greatest orchestral combination ever assembled—the NBC Symphony under Arturo Toscanini, as well as with the violinist Jascha Heifetz and the pianist Vladimir Horowitz.

British Decca had a far reaching role to play after World War II when its "ffrr" full frequency range recording became internationally known. The frequency range of discs had been dramatically extended, and Ernest Ansermet's recording of Stravinsky's *Petrushka* in the new process was to awaken the unsuspecting ears of many record collectors in 1946 to the future high fidelity, or "hi-fi," possibilities of the phonograph.

Two other late-1940s developments combined with the extended frequency range to produce a radical change in the development of recordings: magnetic recording and the long-playing (LP) record. In 1948 Columbia Records demonstrated 12-inch unbreakable vinyl discs that could play about 25 minutes of music per side. The standard shellac disc had revolved at 78 revolutions per minute, and a 12-inch disc had to be changed, automatically or manually, every five minutes, thus breaking up the continuity of longer works; the 12-inch LP, revolving at 33⅓ revolutions per minute, could hold the average symphony, sonata, or quartet on a single side. And the vinyl discs had quieter surfaces than the shellac. Victor soon countered with its own microgroove records: seven-inch vinyl

discs at 45 revolutions per minute. Each contained approximately as much music as a 12-inch 78-revolution-per-minute disc, but the package was smaller. By 1950, a pattern had been set: 12-inch LPs for classical works and popular albums, 45s for individual popular songs. Extended-play 45s also were developed and successfully merchandized. The LP opened up an entire new market —not only newcomers but older record collectors who could see the advantage of the new technology and were willing to repurchase their collections as LPs. The 78-revolution-per-minute shellac disc followed the cylinder into oblivion.

Tape had a major impact on recording starting in the late 1940s: anyone with a good recorder and microphone could become a record producer. Small companies sprang up in areas of music ignored by the giants: the esoteric and the avant-garde, the music of the periods before and after the highly popular Romantic classics of the 19th century. Chamber music, as well as Baroque works of the 18th century and earlier, which required paying fewer musicians than an entire symphony orchestra, flooded record stores and resulted in an unprecedented Baroque revival among music lovers. All-Vivaldi concerts were sold out, and Bach became a best seller. Orchestral recordings of less familiar works—produced in a Europe that had been ravaged by World War II, where musicians' fees were minimal—crested the flood. New companies recorded for the first time many symphonies, quartets, masses, little-known operas, and many other once esoteric works, some of which were now available in competing versions. The more popular standard works, the symphonies of Beethoven, Brahms, and Tchaikovsky, became duplicated by the dozens. By the mid-1950s it seemed that practically the entire worthwhile musical output of Western civilization—and much from Asia and Africa—had been made available to the average home.

The stereo revolution

For a few owners of some deluxe tape recorders, a new listening experience was available by 1956: stereophonic tape recordings. Within two years, stereo discs made their commercial appearance; every major U.S. company began issuing stereo discs by the end of 1958.

Now a new flood of records hit the market: notably popular were those that displayed the spectacular effects possible with stereo. It was again Decca/London that convinced the serious music lover of the musical benefits of stereo with the release in 1959 of Wagner's *Rheingold*, conducted by Georg Solti, a pioneer work in the "creative" school of classical record production. Within a decade, two complete recordings of Wagner's *Ring* cycle, comprising four complete operas on 19 discs, were available. Again, as with the advent of the LP, a technical advance spurred the record industry into recording an even greater repertoire than was available previously. By the late 1960s, most American record companies had discontinued their monaural recordings, except for "historical reissues."

The 1950s saw a rearrangement of record company alliances, as Europe began its strong invasion of America. By the 1970s not only did Europe own a sizable chunk of the American record industry, but it had taken over the recording, for the first time, of many of the most prestigious U.S. orchestras.

Another far-reaching phenomenon of the 1950s was Elvis Presley, a popular U.S. singer who inspired a new, militantly youth-oriented style of music: rock and roll. It generated a multitude of solo singers and groups as well as a teenage and subteen culture of rabid record buyers. The success of the British rock group the Beatles helped stimulate record sales in the 1960s to an all-time high. The sales of classical records, however, represented a declining portion of the total. It seemed that most people who wanted the standard classics had already bought them and that few new standard works of any length were being written. Rather than new classical recordings, re-releases were issued in new packages (*e.g.*, "Debussy's Greatest Hits") and in the new medium of tape. During the mid-1960s, two small and conveniently packaged tape formats began a steady rise to popularity: the continuous-loop one-reel cartridge and the two-reel cassette.

Cartridges, cassettes, and video-tape

Each obviated the need for threading tape in order to play it. The cartridge first achieved consumer acceptance as an automobile accessory and was designed primarily as a playback-only format; the cassette configuration was first introduced in an inexpensive portable recorder player.

Cassettes had the advantage over continuous-loop cartridges in being rewindable and thus easier to control for selective "spotting" and for amateur recording. For non-selective music or music in which it is not necessary to start at the beginning—background music, for instance —the continuous-loop cartridge had the advantage of not having to be rewound at all. A new potential appeared in recording music with the addition of television. Videotape cartridges and cassettes were studied by most major record producing companies, and a number of artists experimented with video productions, including the U.S. orchestral conductor Leonard Bernstein, who recorded the Verdi *Requiem* for a video cassette, and the Italian opera and film director Franco Zeffirelli, who directed a video production of Beethoven's *Missa Solemnis* at St. Peter's in Rome. Opera appeared to be the most likely musical genre for video, if the size of the television screen could be made commensurate with the grandeur of the productions. The grafting of images onto the great tradition of Western music will probably require genius comparable to that of the great composers themselves.

BIBLIOGRAPHY. ROLAND GELATT, *The Fabulous Phonograph: From Edison to Stereo*, rev. ed. (1966), the best written, general history of the phonograph, particularly good on the complex corporate developments; OLIVER READ and W.L. WELCH, *From Tin Foil to Stereo* (1959), a scholarly and detailed phonographic history with much elucidation of technical matters, including a reproduction of the original Edison patent and a "genealogical" chart of companies; F.W. WILE, *Emile Berliner: Maker of the Microphone* (1926), an adulatory, but worthwhile biography; FRED GAISBERG, *The Music Goes Round* (1942), an autobiographical account of the history of the phonograph by a man associated with Berliner from the earliest days (a valuable source despite many inaccuracies); MATTHEW JOSEPHSON, *Edison* (1959), a thoroughly researched biography of the inventor of the phonograph; JOSEPH BATTEN, *Joe Batten's Book: The Story of Sound Recording* (1956), another personal documentation of early phonographic history, not as far-ranging as Gaisberg's but with much unduplicated material.

(L.M.M.)

Music, Theatrical

Music as an art of the theatre has its roots in primitive ritual and ceremony and its branches in every modern means of theatrical presentation. Its functions are as varied as the forms require and range from being the primary reason for performance, as in opera, to mere noise, filling a vacuum in imagination for some screen and stage presentations. The more significant and universal aspects of theatrical music are considered in this article; opera is the subject of a separate article.

Theatre music is all music composed to govern, enhance, or support a theatrical conception. Music composed for theatrical purposes obeys different laws than does the music for concert performance or conventional opera. Whereas in opera the music dictates the form in which the dramatic visual imagery is presented and governs its development, in other kinds of theatre the music is, at best, an equal partner among its principal elements. In concert, of course, the music is the sole factor that determines the experience.

In the West, the concept of music as an intellectual experience for its own sake emerged only in the second half of the 18th century. Theatrical music is variously related to something other than itself, whether as an enrichment of words (as in operetta), a factor in structure and mood (ballet), or an intensification of situation and feeling (as in incidental music for plays and films).

In some instances music is dominant, in some it is subservient, and in operetta or stage musical the emphasis alternates between speech, song, and dance. In opera and spoken drama, in which words are wholly sung or spoken, a convention once set is consistently sustained and thereby creates its own kind of reality. The constant change of focus in operetta and musical, from music to

speech and back again, emphasizes the artificiality of the illusion they seek to create.

The classical mainstream of theatre music in the West extends from the mid-17th century to the 1930s, and the instances of drama and music meeting on an equal level of imagination are relatively few. More frequently great music was lavished on weak or corrupt theatre, or great drama was embellished with indifferent music. From the early 20th century new dramatic developments were seldom directly matched in music. A German–Italian composer, Ferruccio Busoni, wrote in 1906:

> The greater part of modern theatre music suffers from the mistake of seeking to repeat the scenes passing on the stage, instead of fulfilling its own proper mission of interpreting the attitudes of the persons represented.

The German composer Kurt Weill's score for *Der Silbersee* (1933; *The Silver Sea*) was the last major musical contribution to a serious play requiring a full orchestra and chorus. Thereafter, for economic reasons, the dramatic theatre had to equip itself with small-group music or prerecorded tapes. The orchestra and chorus became the prerogative of stage musicals and films. The more these were commercially debased the more they came to rely heavily on the clichés of 19th-century music, to the exclusion of newer musical developments.

Producers of stage musicals, the choreographers of dance, and the directors of drama need to be wary of the properties of music. It is more demanding of attention than is often thought, and its use should ideally be confined to circumstances where it can provide something that none of the other theatrical elements can offer. The more its qualities are understood and respected, the better it can be guided to an effective theatrical purpose.

Types of theatrical music

MUSIC FOR BALLET

The partnership of choreography and music

During the 20th century the element of music in all forms of ballet has changed and developed its significance to an unprecedented extent. It has acquired the status of an equal partner with the choreography, where once it was entirely the servant of the ballet master. In the 19th century he asked little more than that the music should decorate his ballets prettily and give rhythmic support to the dancers. His modern successors have been made aware that the highest level of balletic achievement now requires the music and choreography to become extensions of each other, to be heard and seen on equal levels of perception.

The finest modern choreographers are not content simply to ride the surface of their chosen music, whether it is specially written, borrowed, or adapted. They seek to exploit the relationship of eye and ear, recognizing that the effect of any danced step can be changed by the stress of the musical rhythm, the degree of loud or soft in its dynamics, the nature of its harmonic character, and the expressive quality of its instrumental timbre. All these factors can make a positive contribution to the balletic image, and they give coherence to the sequence of movement as they merge into its continuous momentum.

Until the end of the 19th century each new ballet customarily had music specially composed for it, but it was rare for any composer of distinction to write ballet music, unless it was part of an opera. There were exceptions—such as Beethoven's score for *Die Geschöpfe des Prometheus* (1801; *The Creatures of Prometheus*), originally a ballet by an Italian choreographer, Salvatore Viganò—but the great era of Romantic ballet from about 1830 was seldom enhanced by music of intrinsic worth. When ballet gained a musical interest the difference was soon apparent; it has been well said that Léo Delibes gave ballet music a heart, Peter Tchaikovsky gave it a soul, and Igor Stravinsky made an honest woman of it.

Stravinsky's influence on music for ballet

Stravinsky was the dominating figure for nearly half a century (beginning in 1910) in composing music for ballet. He gained international acclaim with the first products of his collaboration with the Ballets Russes of the Russian impresario Sergey Diaghilev: *The Firebird* (1910), *Petrushka* (1911), and *The Rite of Spring* (1913). The first two continue to be performed in their original choreography by Michel Fokine, also a Russian, each with a narrative basis illustrated in music notable for its expressive colour and harmonic innovations. *The Rite of Spring* provoked one of the most notorious scandals in theatre history when Vaslav Nijinsky's original ballet reduced its first Paris audience to verbal insult and physical assault; its rhythmic audacity has since remained a recurring challenge to other choreographers.

Largely as a result of the standard set by Diaghilev's flair and artistic success, most leading composers in the 20th century have contributed something to the art of dance. Diaghilev directly commissioned two outstanding examples in the French composer Maurice Ravel's *Daphnis et Chloé* (1912), which the composer defined as a "poème choréographique," and *The Three-cornered Hat* (1919) by the Spanish composer Manuel de Falla. Distinctive original scores for ballet continued usually to be the outcome of specific commissions. Composers do not yet normally think in terms of dance (as they do in terms of song), although in Great Britain the composer Peter Maxwell Davies has incorporated an integral role for solo dancer in his otherwise instrumental work, *Vesalii icones* (1969).

Contrary to what is still sometimes thought by musicians and a section of the public, music for ballet is not necessarily written to be "interpreted" in dance. Stravinsky has emphasized:

> Choreography must realise its own form, one independent of the musical form though measured to the musical unit. Its construction will be based on whatever correspondences the choreographer may invent, but it must not seek merely to duplicate the line and beat of the music.

Stravinsky's collaboration with Balanchine

The composer was writing from the experience of a long collaboration with a Russian-born choreographer, George Balanchine, mainly for what is now the New York City Ballet. Their partnership began in 1928 with *Apollo*, reached a peak in 1957 with *Agon*, and has been to the 20th century what the collaboration of Tchaikovsky with choreographers Marius Petipa from France and Lev Ivanov from Russia was to the 19th.

In the Soviet Union a post-revolutionary equivalent to the classical three-act narrative ballet has continued to be in demand. This form has been furnished by many composers of ballets that have not been performed beyond the U.S.S.R., but it has been chiefly distinguished by the work of Sergey Prokofiev. His full-length scores for *Romeo and Juliet* (1940), *Cinderella* (1948), and *The Stone Flower* (posthumously staged in 1954) have variously succeeded in reconciling an older classical form to new expressive demands. Its offshoots in western Europe have included, in Great Britain, Benjamin Britten's music for *The Prince of the Pagodas* (staged by the Royal Ballet, 1957) and, in Germany, Hans Werner Henze's music for *Ondine* (Royal Ballet, 1958).

The stronger emphasis in western Europe and the United States is currently on one-act ballets, varying from about 15 minutes to more than one hour's duration. Composers continue to show some reluctance to write music specifically for dancing, partly because they are seldom closely involved with the art, and also because they have relatively less control over the finished performance than in opera and in most other forms of musical theatre. Apart from works already mentioned, however, the ballet repertory has been enriched by such scores as the French composer Erik Satie's *Parade* for Diaghilev in 1917; the British composer Vaughan Williams' *Job* (1931) and the English composer Sir Arthur Bliss's *Checkmate* (1937) for the British ballet mistress Ninette de Valois, in Britain; and the U.S. composer Aaron Copland's *Appalachian Spring* (1944) and the American composer Samuel Barber's *The Serpent Heart* (1946; later revised as *Cave of the Heart*, 1947) for the ballet mistress Martha Graham, in the United States.

Creation of modern ballets to existing scores

Nevertheless, the greater proportion of new ballets in the West during the last half-century have been created to pre-existing music, and it is evident that choreographers have felt a greater freedom to experiment visually in the use of such music. The practice dates principally from the first orchestration (by the Russian composer Aleksandr

Glazunov, 1894) of an arbitrary suite of piano music by Frédéric Chopin to which Fokine created *Chopiniana* (1908)—a title retained by Soviet ballet companies for what Diaghilev renamed *Les Sylphides* (1909). More than 60 years later, another arbitrary suite by Chopin, although retained in its piano form, proved to be no less fruitful for the U.S. choreographer Jerome Robbins in *Dances at a Gathering* (1969).

Almost every category of music—from medieval music to advanced electronics and from symphonic compositions to the simplest pop tunes—has now been used for choreographic purposes. The function of music in a dance context has varied almost as much—staining the ballet's surroundings as a kind of aural decor at one extreme, as with the soundscapes of John Cage for Merce Cunningham's dance works in the United States, to the mutual absorption of music and choreography phrase by phrase in such works as British choreographer Sir Frederick Ashton's *Symphonic Variations*, French composer César Franck's music (Royal Ballet, 1946), and the Balanchine–Stravinsky *Movements* (New York City Ballet, 1963).

Sometimes the same music has been used for several different ballets within a short time by various choreographers. An outstanding example is the *Sinfonia* by an Italian composer, Luciano Berio, written originally in 1968 as a concert work. By the end of 1971 it had been taken over for at least eight separate ballets in almost as many countries in western Europe alone.

Another trend in ballet in the second half of the 20th century has been to make ballets apparently more concerned with musical associations than with human personality. Instances include Balanchine's *Agon* and *Movements*, already mentioned, and the British choreographer Kenneth MacMillan's *The Song of the Earth* (1965) to the song-symphony by the Austrian composer Gustav Mahler. The dancers seem required to assume the "personality," or expressive character, of the musical instruments they parallel, as if the choreographers were moving toward a form of "ideal" dance once postulated by a French poet, Stéphane Mallarmé, who envisaged music and dance not only as equals but also equally devoid of human personality.

Ballet concerned with musical associations

In the view of most ballet critics, the antidote lies in the continuing appeal of narrative dance-dramas with their illustrative music, although the success or otherwise of any ballets that engage pre-existing music is basically governed by a single crucial principle: the level of choreographic imagination should never be less than that of the music. A ballet can be better than its music, but it can never afford to be worse. There will always be ballets dependent on music for no more than expressive colour and supporting rhythm, but the "perfect analogous concord between what we see and what we hear," recommended as the ideal nearly 150 years ago by Carlo Blasis, a great Italian teacher, still remains the most desirable aspiration for every kind of ballet music.

STAGE MUSICALS

When, in the 1930s and 1940s, dancing became an integral element in a genre governed chiefly by song—instead of being merely a diversion—the "musical" established itself as the legitimate theatrical heir to "musical comedy" and a form of popular theatre art that dominated the third quarter of the 20th century. It has lately been challenged by the newer "rock musical," using a variation of the common musical vernacular and techniques related more to the recording studio than to the theatre, the effect of which is not yet determined. Meanwhile, what originally started as a democratic counterpart to aristocratic opera reached its fruition as the theatrical association of sentiment with illusion.

The sentiment is usually dispensed by the narrative; the illusion is created by the music. The most potent narratives in stage musicals have often been adaptations of classical drama and literature—for example, *Romeo and Juliet* transformed into *West Side Story*; *The Taming of The Shrew* into *Kiss me, Kate*; *Don Quixote* into *Man of La Mancha*; and *Oliver Twist* into *Oliver*—or the many variations on the *Cinderella–Pygmalion* legend by which rags are transformed into riches (from *The Shop Girl* in 1894 to *My Fair Lady* in 1956). A distinction was at one time drawn between the frivolous musical comedy and the "musical play," denoting a dramatically serious or even tragic narrative, but both are now equally defined as musicals.

Their specifically musical character is born from a marriage of convenience between first and second generation descendants of European operetta and music-hall variety, on the one hand, and American jazz and American music hall, on the other—plus the romantic balladry of both continents. An English musicologist, Wilfred Mellers, asserts that although most successful stage songs contain subtleties unappreciated by the non-musical listener, they all reflect "an illusion that we can live on the surface of our emotions" and that "the world of musical comedy never gets beyond, or wants to get beyond, this illusion."

The first musical comedy so to be called was *A Gaiety Girl*, staged in 1893 by George Edwardes at the Gaiety Theatre, London. A romantic farce, adorned by the songs of Sidney Jones, it was successfully exported to New York in the same year. John Hollingshead (Edwardes' predecessor at the Gaiety Theatre) wrote in 1903:

First musical comedy

> The invention or discovery of musical comedy was a happy inspiration of Mr. George Edwardes's. It provided a new form of entertainment for playgoers who go to a theatre for amusement and recreation, which was more elastic in plot or story than the old burlesque. . . [It] exhibited a little of the old burletta and vaudeville, most of the best elements of farce, a dash of the French revue. . .and much that would not have been out of place in Parisian *opéra-bouffe*.

Some 50 years of development in musical theatre is reflected in the contrast between the foregoing remarks and the following comment in 1952 by Jack Burton, U.S. theatre historian, on *Oklahoma!* (1943), an epoch-making musical by Richard Rodgers and Oscar Hammerstein:

> This phenomenal production set a new pattern in which every line, every song, every dance routine is an indispensable part of a closely-knit whole. It was a show that had dramatic substance and never ran off the plot track, and so real, so simple, so engrossing was its story that its narrative could be safely entrusted to other than big-name stars.

The years embraced by these comments set up a dominant axis of theatrical exchange between New York and London. Success in this field is governed more by economic than artistic considerations, with longer and longer runs of each production becoming necessary to recover increasingly heavy expenditures before profits can be made. The primary requirements for the composer are, therefore, quick assimilation combined with durability. His music should also be easily adaptable to other media such as motion pictures and phonograph or tape reproduction, whereby it becomes a commonplace of experience to a mass audience far beyond the reach of the original theatrical context.

Although musical theatre of this kind has developed toward a closer integration of music and story, its primary feature has remained the individual song. Lehman Engel, a leading conductor of stage musicals in the United States, has defined five types of song basic to the stage musical: ballad—usually but not exclusively romantic in feeling; rhythm song—varied in emotional character but primarily propelled by a prominent musical beat; comedy song—enhancing verbal humour and divided into "short joke" and "long joke"; charm song—generally delicate, optimistic, and lightly rhythmic; and musical scene—in which a song may form part of a continuous dramatic episode. Engel further asserts that the successful impact of any song in the first instance is generally governed by the following considerations: the tempo, the mood of the scene, the song's position in relation to the whole production, the inherent value of the song itself, and the relative importance of the character who delivers it. It will be noted that integral musical quality is subordinate on this scale, although it is specifically the musical appeal that establishes success in the first place, disseminates that success through other media, and may later lead to revi-

val in the country of origin and to reproduction in other countries.

Success of *My Fair Lady*

Probably the most successful specimen to date is *My Fair Lady*, with music by Frederick Loewe, Vienneseborn U.S. composer. This musical had first runs of 2,717 performances in New York (from 1956) and 2,281 in London (from 1958), and it has since been staged in translation in most European countries and in the Soviet Union. It is rare for the English-language dominance of the musical-comedy genre to be breached by other countries, as France did with *Irma la Douce* (1956) or pre-Nazi Germany did with *Die Dreigroschenoper* (*Threepenny Opera*, 1928) and *Im weissen Rossl* (1930; *White Horse Inn*); but in most European countries except Britain the line between musical and operetta (see below) is less distinctly marked.

In Italy such lighter forms of musical theatre are submerged in an already popular taste for the broadest range of opera, while in Spain they manifest themselves in the category of zarzuela (discussed below in conjunction with operetta). Differences of idiom are often more the outcome of theatrical or other conditions in their respective countries than of theatrical or musical distinctions in the work itself. In the Soviet Union, according to Andrey Olkhovsky, a Russian-born musicologist,

> the numerous attempts which have been made to create a Soviet repertory have led to no results. At best the plots of comedies are based on episodes from Soviet life, but musically they are still imitations of the pre-Revolution operettas.

Stage musical comedies in the Western sense have produced their own original talents among composers—notably Jerome Kern, Irving Berlin, George Gershwin, Cole Porter, and Richard Rodgers in the United States; and Noel Coward, Vivian Ellis, Ivor Novello, Lionel Bart, and Sandy Wilson in Great Britain. They have also had occasional recourse to adaptations from classical composers, including Franz Schubert and Edvard Grieg, who are dramatically characterized respectively in *Lilac Time* (originally *Das Dreimäderlhaus;* 1916) and *The Song of Norway* (1944); Georges Bizet, whose music became the basis of *Carmen Jones* (1943); and Aleksandr Borodin, for *Kismet* (1953).

Musicals ought to be adaptable to varied instrumentation, because theatre orchestras can vary considerably in size and composition from place to place. Paradoxically, the looser form of the new rock musical is propelled by a much more rigid instrumentation derived from the ensemble used in pop-music recording, itself determined by studio techniques. In *Jesus Christ Superstar* (1971) the covering of the orchestra pit, the permanent amplification of instruments, and the use of voices entirely dependent on microphones amounts to a replacement of the illusion of theatre in any traditional sense with the actuality of modern recording studio made visible.

The arrival of the rock musical

MUSIC FOR MOTION PICTURES AND TELEVISION

Motion pictures. Many successful stage musicals have become additionally popular through the medium of motion pictures, but music as a basic element in film making has only gained recognition since midcentury as something more than a means to heighten local colour or intensify emotional expression. In the early silent films, all kinds of music were recorded, classified, and adapted to fit different moods (Beethoven overtures for cowboy–Indian chases, for instance). Several talented hacks also wrote short descriptive pieces. A few bigger films, such as *The Birth of a Nation*, had special scores fitted to them. Since the 1960s it has turned a full circle of the wheel back to extensive musical quotation from classical resources for similar ends but in different ways.

Russian film makers first gave serious consideration to the contribution music could make. V.I. Pudovkin, a Russian musicologist, defined a theory and practice of film music in the early 1930s, advocating a close and contrapuntal relationship between sound and sight. The Russian film director Sergey Eisenstein described his careful collaboration with Prokofiev in making *Alexander Nevsky* (1938). His perceptive observations on the potential link between cinematic rhythm and musical rhythm suggested a technique that has influenced others, such as the British composer Sir William Walton in his score for *Hamlet*.

Hanns Eisler's theories of film music

Hanns Eisler, a German-born composer, formed his own theories of film music, based on empirical experience composing in this medium. His published findings recommended: short musical forms in a film context; the composer's conscious awareness of the film's realistic sound element (the "where" and "when" of its location); and music that could suggest an objective, universal character for the film's emotions, rather than being introspective on its own account. Eisler also supported Pudovkin in maintaining that film music should create its own sense of line independently of, although related to, the film narrative.

Film music has travelled through five broad phases: an initial borrowing from existing conventional sources; the use of musical-cliché catalogs, which enabled any musician of modest ability to assemble an emotional or dramatic sequence and which served as the basis for most later background music in films; the active interest of major composers in writing original scores; a subsequent reaction of either near silence or advanced techniques of electronic sound generation and transformation; and the borrowing from classical sources for new purposes.

In the Soviet Union and France, in particular, the regular participation of leading composers in film music has been the rule rather than the exception. The opposite holds true in the United States and Great Britain, although most composers of distinction in both countries between the mid-1930s and the mid-1950s made some original contributions to motion-picture medium. In the United States these included Aaron Copland, Marc Blitzstein, and Virgil Thomson, who received the first Pulitzer Prize awarded to film music—for *Louisiana Story* (1948); in Great Britain, Ralph Vaughan Williams, Benjamin Britten, Sir William Walton, Alan Rawsthorne, and Richard Rodney Bennett.

Isolated experiments to marry films to specific classical music have been made from time to time. These include the Austrian director Max Reinhardt's film of Shakespeare's *A Midsummer Night's Dream* (1935), which varied the cinematic structure to incorporate the complete suite of Mendelssohn's incidental music originally written for the play. The U.S. film maker Walt Disney's celebrated *Fantasia* (1940) adapted the technique of the animated cartoon to illustrate a sequence of musical classics, outraging some people because the visual relationships were held to be irrelevant to the music but continuing into the 1970s to entertain audiences, especially young ones, on an international scale.

When the making of motion pictures became an industry more than a craft, a situation developed in which music was recognized as desirable by the manufacturers, who nevertheless made no claim to "understand" music in the way they would be expected to understand motion-picture techniques. A parallel attitude among musicians unfortunately came to regard the skilled provision of a score tailored to the demands of script and camera as a spiritually impoverished relation of aspiring symphonic or operatic works.

Renewed use of concert music

In the late 1960s original film composition tended to decline in favour of renewed and often extensive borrowing from existing concert music. The slow movement from a Mozart piano concerto served to express the passage of time as well as for emotional mood in *Elvira Madigan*. The U.S. film director Stanley Kubrick's *2001: A Space Odyssey* (1968), led an audience's imagination outward into space by a transition from the diatonic (using the natural scale of five tones and two semitones) C major of the introduction, to Richard Strauss's tone-poem *Also sprach Zarathustra*, to the polytonal Kyrie from the *Requiem* by the contemporary Hungarian-born composer György Ligeti. In the Italian film director Luchino Visconti's *Death in Venice*, four repetitions of a long passage from the Adagietto movement of the Austrian composer Gustav Mahler's *Symphony No. 5* achieved a different expressive purpose in association with the visual scene each time it was heard.

Examples like these achieved an even greater impact than many original scores because they formed an organ-

ic part of the film maker's conception. Conversely, even scores distinguished by their own merits—such as Copland's for *Of Mice and Men* (1939), or Walton's for *Henry V* (1944)—were added only after the film itself was more or less finished. The major hope for the future of this medium lies in what Vaughan Williams urged: the composition of original scores but prepared in conjunction with the director from the film's inception. Film music might then be turned to more constructive instead of merely decorative purposes, without the dramatic license of the British film director Ken Russell's treatment of extensive passages from Tchaikovsky's music in his film *The Music Lovers* (1971).

Television. The screen medium's first law, that the visual element must come first, has been intensified by television. On the home screen, the experience of music performed for its own sake customarily operates under a double disadvantage. First, it runs the risk of being swamped by its visual presentation, which may range in character from the matching of nonmusical images in varying degrees of relevance to the technique of using close-ups of musicians in action. Secondly, it suffers the continuing handicap of inadequate reproduction by the average television receiver.

Limited repertory of television music

Apart from rare exceptions—such as an occasional "television opera," a dance-film, or Stravinsky's mixed media *The Flood* (1962)—original music to television is chiefly confined to the provision of theme passages or supporting music hopefully intended to enhance verbal or dramatic presentation. Like the cinema pianist who played for silent films, television music has a limited repertory of conventional gestures. Even when these are given a more contemporary harmonic or instrumental garb, they remain basically governed by the 19th-century mode of musical thought, to which it is assumed that mass audiences will most easily react.

Programs about rather than of music have obtained a modicum of television success. While the occasional theme quotation has perhaps introduced a famous musical classic to millions who would not otherwise have heard it, the "workshop" program, showing how music and musicians go about their business, has broken down barriers of technique and exposed the raw materials of music in a way that has probably helped to foster a wider interest in the finished product. Television cannot otherwise be accepted as a musical medium, until sets have a higher standard of musical reproduction.

INCIDENTAL MUSIC FOR THE THEATRE

Incidental music in the theatre, whatever its idiom or degree of stylistic emancipation, justifies itself through its exclusive concern with a specific play or theatrical presentation. Its three main uses involve songs, intensified dramatic effect, and interlude filling, and these have been clearly defined in Western theatre since Renaissance drama freed itself from the church in the 16th century.

A major modification of its character since the 1920s has been brought about largely by the wider use of mechanical techniques of amplification and recorded music. This has encouraged short musical fragments rather than fully composed pieces, except where the latter are specifically called for. Examples of this technique were heard during the late 1960s and early 1970s in the musical productions of the National Theatre and the Royal Shakespeare Company in Great Britain. Such carefully planned but more informal use of incidental music has replaced the elaborate suites customary in the 19th and early 20th centuries, which presupposed the complete performance of each piece and required the stage drama to be produced so that the music could be accommodated in it (a characteristic example is Mendelssohn's music for *A Midsummer Night's Dream* at Potsdam, Germany, in 1843).

Informal incidental music

Economic conditions are now the principal factor governing the provision of incidental music. Theatres not concerned with opera or ballet can no longer afford to hire musicians for a pit band. Trade-union restrictions in the musical profession also limit the public use of recorded music in many countries. The trend is consequently for spoken drama either to dispense with music, to restrict it to one or two musicians with a singer, or to make increasingly fruitful use of electronic sounds on prerecorded tape.

Some famous 19th-century suites of incidental music were brought into being by conditions that favoured resident musicians at court theatres in Germany or lavish orchestral resources in Tsarist Russia. The music has since acquired independent status on its own merits and has become part of the classical concert repertory, such as the Mendelssohn just mentioned and Beethoven's music for Goethe's *Egmont* (1810); Schubert's for the German playwright Helmina von Chézy's *Rosamunde* (1823); Schumann's for Lord Byron's *Manfred* (1852); and Grieg's for Henrik Ibsen's *Peer Gynt* (1876).

Earlier theatre music was often, for long periods of time, governed by measures of censorship and legal restrictions that varied greatly from one country to another. In England, for instance, the monopoly of spoken drama was vested by King Charles II in the Theatres Royal in the 17th century, and this continued until 1843. Other theatres that opened during this time—including those catering to a new working class—were licensed only on condition that plays included five musical items in each act or a "musical accompaniment." The latter condition was sometimes held to be satisfied by no more than a chord struck at intervals on a piano during the performance. Such conditions inevitably brought about a profusion of inferior music, which in turn gave rise to the traditions of music hall and burlesque.

The renaissance of secular theatre, from the 17th century onward, led in Italy to the evolution of the predominantly musical forms of opera and oratorio, in France to the court ballet, and in England and Spain to the cultivation of spoken plays with incidental music. The unparalleled achievements of secular drama in the latter two countries are the roots of present-day musical theatre and help to explain why opera failed to flourish in competition. The drama in England expanded into the allied form of masque and involved the participation of such composers of distinction as Henry Purcell.

The masque

The masque in relation to its own period might well have been defined as "mixed media." This term has now come to stand for theatre presentations in a line of descent from Stravinsky's *Histoire du Soldat* (1918; *The Soldier's Tale*), which combines speech, song, mime, dance, and instrumental music with pictorial design. These elements have since been supplemented by the mechanical techniques of film or photographic projection, and of electronic sounds, in almost infinite permutations, together with a free form of expression no longer necessarily shaped by a narrative content.

Most of these manifestations incorporated two different kinds of musical contribution. One has been defined by a 20th-century German composer, Bernd Alois Zimmermann:

> All elements of the theatre of movement, including film, sound, speech, electronic music, must be mobilized into one great time-space structure, whose arrangement will be constituted by music as the most general form of temporal order.

Zimmermann's ideas were embodied in his opera, *Die Soldaten* (The Soldiers). The alternative is described by another composer, John Cage, as "Single sounds or groups of sounds which are not supported by harmonies but resound within a space of silence" and are added more or less at random to the other elements. It remains to be discovered whether the future of theatre music in drama or mixed media lies primarily with the highly organized patterns of interaction postulated by Zimmermann or with Cage's arbitrary combination of simple, disparate activities into a complex whole.

OPERETTA AND ALLIED FORMS

It seems unlikely that any future exists for operetta except as part of a "museum" repertory, because the contemporary musical has taken its place in musical theatre. Operetta in the usual sense of a work of lesser musical pretensions than opera, with spoken dialogue linking the

musical episodes, was a direct descendant of comic opera, overlapping this category, on one hand, and early musical comedy, on the other. It was born as a democratic expression of popular wit and social satire, flourished for almost exactly a century (*c.* 1840–1940), and died from a surfeit of sentiment.

Offenbach. The satirical, romantic operetta emerged primarily in Paris in the mid-19th century with the French composer Jacques Offenbach; two of his works are still widely staged: *Orphée aux enfers* (1858; *Orpheus in the Underworld*) and *La Belle Hélène* (1864; *Beautiful Helen*). The character of Offenbach's operettas established several musical precedents, including the burlesque of Italian opera, the romantic ballad in $\frac{3}{8}$ or $\frac{6}{8}$ metre, and the drinking song and the *ensemble de perplexité* ("ensemble of confusion"). In England, Arthur Sullivan followed in Offenbach's wake with his fruitful partnership with the author W.S. Gilbert, bequeathing a commentary on aspects of Victorian society through music of popular and enduring, if somewhat relentless, charm.

Viennese operetta. Charm is the main ingredient of the more sentimental Viennese operetta, and it usually submerges the rarer shaft of social comment. The younger Johann Strauss made operetta an international entertainment by an expert blend of charm and craft, and his *Die Fledermaus* (1874; *The Bat*) remains a classic of its kind. A second generation in this tradition was chiefly distinguished by Franz Lehár, whose *Die lustige Witwe* (1905; *The Merry Widow*) represents the genre at its peak of romantic elegance, demonstrating a style and craftsmanship that seems in serious danger of being lost altogether.

Johann Strauss's contribution to operetta

Such operettas remain current in today's musical theatre mainly as an indulgence of musical and emotional nostalgia. Their popular style enabled them to take root and flourish far from their native territories, including transplantation to the United States. The indigenous tradition of the U.S. stage musical, already mentioned, first had to compete with European-style operetta. That the latter keeps a tenacious hold on popular affections is demonstrated by figures listing Rudolf Friml's *Rose Marie* (1924) and Sigmund Romberg's *The Desert Song* (1926) as the most frequently performed works in U.S. musical theatre, in terms of both amateur and professional performances.

Zarzuela. Spain was a prominent exception to the wide dissemination of operetta, preferring instead the flourishing native variety of zarzuela. This form customarily incorporates regional songs and dances, sometimes with traditional rather than original music. It continues to some extent as a staple fare in Spanish musical theatre, although the general contemporary trend toward a more universal style of musical expression has meant that the younger Spanish composer has shown much less interest in the zarzuela form as an outlet for his musical imagination.

The Romantic zarzuela has little resemblance to the aristocratic and courtly character of its 17th-century namesake and emerged with French and Viennese operetta during the 19th century. It divided into two forms—the *zarzuela grande* in three acts, equivalent to romantic operetta, and the *género chico* in one act, invariably comic, usually satirical, employing the broadest musical vernacular, and verging on revue. In the former category, the names of Francisco Barbieri, Amadeo Vives, and Federico Moreno Torroba are probably the most significant representatives of their respective generations.

Zarzuelas in South America

The Spanish character and language of the Romantic zarzuelas made them exportable to Central and South America, where they became a model for the limited indigenous musical theatre. The South American centres are otherwise dominated by imports or imitations of the Italian opera tradition or used as transit bases for the latest North American stage musical on its theatrical circumnavigation. Certain examples of the Spanish zarzuela, such as Moreno Torroba's *Luisa Fernanda* (1932), have achieved popular success in Latin American countries, where local contributions to the genre have notably been made by Juan Bautista Massa in Argentina, Andrés Martínez Montoya in Colombia, Luis Delgadillo in Nicaragua, and Teodoro Valcárcel in Peru.

ORIENTAL MUSICAL THEATRE

Theatre music generally serves different purposes in non-Western idioms—usually adding dimension and perspective to song and dance, indicating symbolic associations, suggesting mood, and even inducing a desired response in an audience. Most Oriental streams of music divide between popular music of a folk character and a more sophisticated style for a cultured elite. The distinction is often less clearly defined than in Western music and is now not so firmly maintained in the wake of recent political and social changes.

China. The classical Peking opera (*ching-hsi*) in China is a form of musical theatre in which music is one among several elements rather than a governing factor, as in Western opera. The vocal writing alternates between styles broadly equivalent to recitative and song, distinguished by a forced high falsetto tone required from the male singers. A less stylized variety is the all-female *yüeh ch'ü*, in which natural singing voices perform musical plays in realistic and decorative scenery, and the Manchurian *P'ing Hsi*, which has developed into an operetta-like equivalent, with traditions and subjects derived from strolling players and folk legends.

Since 1964 the performance of classical Peking opera in Communist China has been mainly restricted to festival occasions (although state-sponsored schools continue to train performers especially for it). More emphasis has been put on entertainments closer to Western musicals, involving contemporary dialogue, everyday dress, and less stylized music. As a popular form of musical theatre it has been turned to political and social advantage with a new and adapted repertory of dramatic ballets and musical plays, bearing such titles as *Taking Tiger Mountain by Strategy* and *The Red Detachment of Women.*

Political overtones of modern Chinese opera

Japan. Music is as much a regular part of theatre performance in Japan as it is in China. The highly formal tradition of Japanese Nō drama incorporates music as an integral feature, usually performed by flute, a variety of string instrument called the samisen, drums, and singers. The music varies in content and character with the subject of the play and obeys detailed melodic rules—especially in the central dance episode designed to reveal the spirit of the play's principal character. A less formal counterpart, the Kabuki theatre, has almost as impressive an ancestry as the Nō and continues to be widely performed, with music used to indicate period, place, time, or mood and often functioning by phrase association like the principle of leading motives in Richard Wagner.

India. Japanese theatre also incorporates music dramas of Indian origin, and the Indian theatre tradition is a full combination of poetry, music, dance, and symbolism. The music is often interpolated rather than specially composed and is likely to be drawn from the repertory of widely known songs without aiming at a high classical standard. The close association of music with drama in Indian culture has been carried over into Indian film, which cannot hope to enjoy wide success among its modern audiences unless it is liberally embellished with songs and other forms of music.

The history of theatrical music

FORMATIVE PERIOD

What is thought to be the oldest document of musical history depicts a man wearing an animal mask, manipulating what is possibly a form of musical bow, and dancing in the wake of a herd of reindeer. This is a prehistoric cave painting dating from the Stone Age, discovered at Ariège in France. Masks are tangible signs of that transfer of personality on which every form of theatre is based and in which song and dance have participated since the dawn of communication and animated ritual. Music in dramatic entertainment reached early peaks of development in European and Oriental cultures in, respectively, the ancient Greece of Homer and, some centuries later, the Chinese classical drama.

Early Greek and Roman theatrical music

Descriptive evidence of the earliest Greek theatre indicates that music, mostly sung by a chorus, was essential but not continuous. At drama festivals the poet wrote his own music (as well as being actor, producer, and choreographer), probably based on some kind of traditional repeated formula. Later Greek theatre, after the fall of Athens (404 BC), initiated both the repertory system and a category of musicians trained more highly than the populace. Amateur and professional became separated for the first time, and increasing sophistication brought about its counterpart in popular pantomime expressed in song and dance, often satirical or bawdy in character.

The Roman musical theatre derived directly from the Greek, ousting a short-lived native form with Etruscan actors who also danced to pipe music. Latin versions of the Greek theatre with music were supplemented by a Roman variant of the pantomime as a dramatic solo dance with chorus and orchestra. It implied some prior knowledge on the audience's part of the subject and the dance vocabulary. Amphitheatre shows of gladiatorial contests were regularly accompanied by music, sometimes involving up to 100 horn blowers and 200 pipers, as well as such extra devices as water organs.

About the time the Roman theatre flourished, an Oriental equivalent emerged in China from ritual ceremonies that came to be repeated for their entertainment value. The puppet theatre was a significant intermediate stage in this process, and the forms evolved into different styles of entertainment for courtier and commoner. Strings, flute, and handbells accompanied the songs and dances in upper-class entertainments; a form of mouth organ replaced the bells in shows for the common people. By the time of the Sung dynasty (AD 960–1279), from which the earliest written music survives, a type of musical variety theatre, the *tzarjiuh*, was widely popular.

The Chinese classical opera tradition has already been mentioned as a modern form of musical theatre. It first developed during the Yüan dynasty (1279–1368) and reached its peak of style and classical form in the Ming period (1368–1644). Its evolution was accompanied by a less formal counterpart based on the dramatization of folk songs linked by a thin narrative plot (*Chueichang*). The full-scale opera and its regional variants remained the most significant form of Oriental musical theatre until the modern post-revolutionary times, but throughout the Far East the indigenous forms of music have always played a prominent part in theatrical presentations.

In Europe the vestiges of Greco-Roman culture were submerged by the early Christian Church. By the 6th century the church had suppressed drama and adapted pagan rituals to its own purposes. A small flame of musical theatre was left burning only in the form of religious ceremonial (for example, in the mass). Festive religious celebrations eventually expanded into the liturgical music drama that slowly developed from about the 10th century. This brought in its wake the equally religious "mysteries" and miracle plays of the Middle Ages in Europe, which were performed in the vernacular instead of in Latin, had a strong musical element, and, in due course, a secular counterpart.

Musical elements of the miracle plays

In a pattern that was to repeat itself after the birth of opera 200 years later, the secular theatre in the Middle Ages established itself either as lighthearted interludes in serious moralities or as deliberate parody tolerated by the church as a safety valve to consistent piety. The annual Feast of Fools in 15th-century Paris, for instance, incorporated an obscene parody of the mass performed in song and dance within the church. By the year 1400 numerous comedies and farces had appeared, usually performed on festive occasions in aristocratic houses or on open stages in municipal squares.

These plays often employed musical forces comparable to those of the religious plays and used them for similar purposes. Choirboys from the church sometimes took part, but surviving texts suggest that there was little choral music as such. The individual actors incorporated parts of songs chanted monophonically to embellish or heighten the dramatic effect, and dancing to specific instrumental music also had a regular place in the entertainment. Professional musicians might be hired and might also be required to act; the constituent parts of the entertainment varied widely from place to place.

The fact that, except for songs, documents of the period contain almost no music directly linked with the theatre is thought to indicate that very little original instrumental music was written for theatrical purposes at this time. Whatever was suitable for weddings, banquets, and other feasts perhaps served a theatrical purpose just as well. Musicians probably had little or no acquaintance with musical notation and played pieces from their regular repertory. These seem to have included arrangements of vocal melodies as well as dance tunes, among which the play texts most frequently identify *basses-dances* and branles.

THE RENAISSANCE AND BAROQUE PERIODS

The French court ballet

When Catherine de Médici married King Henry II of France in 1533, she brought from Italy a taste for entertainments in which dancing was prominent. Her encouragement established the court ballet (*ballet de cour*) as the foundation of classical ballet, the source of a new theatrical identity for music and a precursor of French opera. As a unified blend of poetry, music, and movement, the court ballet dates from the performance of the *Ballet comique de la reine* at a court wedding in 1581. The form comprised an optional number of scenes in mime and dance, prefaced by explanatory verses that were either spoken or sung; the scenes were accompanied by solo and choral songs with lute and instrumental ensemble pieces for strings.

In about 1605 a more mannered style of singing had become customary, and by 1620 the court ballet was more a vehicle for display than drama. Unified dramatic plots were restored by the poet Isaac de Benserade in midcentury, and Jean-Baptiste Lully, who entered the service of King Louis XIV in 1652, endowed the music with fresh distinction. He favoured dramatic musical expression with the use of larger choral and orchestral forms and formulated such dances as the minuet, first danced by Louis as "Le Roi soleil" (the sun king), the gavotte, rigaudon, bourée, passepied, and loure—each with its particular rhythmic metre.

Lully collaborated with the playwright Molière in a famous succession of comedy-ballets, of which *Le Bourgeois Gentilhomme* (1670) is probably the best known. Thereafter the character of Lully's work became essentially operatic, and music in the French theatre was left to function in a more subsidiary role. The dramatist Pierre Corneille, for instance, wrote "I have employed music only to satisfy the ear while the eyes are occupied with looking at the machines." In another category, the pastoral comedies derived from Italian models were liberally embellished with songs loosely strung together and alternating with spoken verses.

The English masque

The French court ballet exerted an influence on the English masque, which took its name and some of its early character from the medieval Italian *mascherate* (masquerades) in carnival entertainment. At a time of distinctive English literary achievement, the masque reached a high artistic level, performed mainly as an aristocratic entertainment. It combined instrumental and vocal music, mixed with dancing and acting, in the representation of mythological and allegorical subjects. At the same time, the first public theatres in London came into being. They date from 1576 (61 years before the first public Venetian opera house); the earliest public plays are known to have incorporated some form of music.

Cues for music recur throughout Shakespeare's plays, usually for simple songs or dances. Music between the acts of public plays was customary by 1600, with the audience often calling for the tunes they wanted played. Instrumental music was employed for supernatural effects and to heighten dramatic tension; it was usually performed behind, at the side of, or even under the stage. After the suppression of playhouses during the Civil War, the restoration of the English monarchy with King Charles II brought an even richer flowering of theatre music, led by Henry Purcell.

Besides those works of Purcell that are nearer to masque than drama, such as the Shakespeare adaptations for *The Fairy Queen* (1692) and *The Tempest* (1695), he composed suites of incidental music for more than 40 plays. These generally comprised overtures, "act-tunes" (interludes), dances, and songs. Rather than growing out of the verbal drama, however, they are often so arbitrarily interpolated into it that only the quality of the music can justify most of them. After Purcell's death, English theatrical music ceased to contribute significantly to the theatre, but Thomas Arne, who wrote numerous masques and ballad operas such as *Love in a Village* (1762), was very popular in the mid-18th century, and his simplicity of expression has a certain direct appeal.

The flourishing tradition of Spanish Renaissance drama precluded much opera from taking root in Spain, but the music for plays had generally less distinction than the English equivalent. The early zarzuela (not to be confused with the later Romantic version already mentioned) originated in the 17th century as a court entertainment. It was the Spanish counterpart to the court ballet and acquired a strong Italian influence on its musical character. Spanish music and musicians travelled to the Western Hemisphere with the early explorers, and by the late 17th century the Peruvian capital of Lima had become musically important. The composer José Diaz worked there and wrote much incidental music to the plays of Calderón de la Barca.

The Italian *intermedio*

Renaissance theatre in Italy bred the *intermedio,* which consisted of songs and instrumental music added before or after the acts of a play. The words of the songs were generally relevant to the action of the drama, and this development—together with more extended musical settings in pastoral plays—became the direct precursor of Italian opera. As a new form of "drama in music" which rapidly acquired serious artistic pretensions, the opera inevitably brought in its train a less aristocratic variety of musical theatre variously termed opera buffa (comic opera), vaudeville, ballad opera, singspiel (literally, song-play), or *tonadilla,* always performed in the vernacular of its audience and often in dialect.

These depicted current events instead of historical or mythological subjects, involved elements of parody and social satire, and usually depended on modest musical resources for economic reasons. They began as interludes performed between the acts of serious opera and comprise an essential link in the history of musical theatre. By about 1700 the scenes (usually two) had acquired a linked plot, and by 1740 they were performed apart from the opera as a genuinely popular entertainment. Eventually they responded to the demands of the rising middle classes by raising their own standards into the category of comic opera.

CLASSICAL DEVELOPMENTS

The Italian commedia dell'arte entertainment of strolling players in mainly improvised comedy had left its mark on French fairground theatre, although the performers were expelled from France in 1697 for having ventured their satire too close to court topics. Ten years later French satirical comedies were also banned, whereupon the resourceful performers found a new way round by employing monologue, mime, and music. They thereby developed a new form of popular entertainment to contrast with the aristocratic opéra-ballets, which were soon to be dominated by the spectacular productions with Jean-Philippe Rameau as composer, and in 1713 two theatrical managements in Paris were given license to perform "Le nouvel Opéra-comique."

Opéra comique was a contraction of *opéra rendu comique* ("opera made comic"), signifying parody and satire at the expense primarily of serious opera. The entertainment soon came to veer either toward *comédie vaudeville,* mostly made up of bawdy satire or simply songs of disparaging social comment, or to the alternative *comédie à ariette,* involving a generally more decorous musical parody at the expense of Italian styles. The *Guerre des Bouffons* ("war of the comedians") between partisans of French and Italian theatrical styles was eventually resolved by the emergence of the opéra bouffe (literally, "comic opera")—the French variety of operetta. It is usually dated to the Paris production in 1753 of *Les Troqueurs, Iroqueurs,* based on a fable by Jean de La Fontaine and having original music by a court violinist, Antoine Dauvergne.

Ballet was declining about this time from courtly heroics to simple diversion unrelated to any dramatic point. Apart from the opera-ballets of Rameau, little significant music was composed for the dance until the German composer Christoph Willibald Gluck, who initiated the major reform of serious opera during the century, first turned his attention to a move for balletic reform involving a more dramatic style. In 1761 Gluck composed the music for *Don Juan,* a ballet by Gasparo Angiolini, the Italian ballet master at Vienna, who maintained that dancing should be self-expressive without recourse to verbal explanations.

Gluck's contribution to ballet music

Gluck's vividly descriptive score contains 31 musical pieces, alternating between formal dances and narrative drama. It made the music a foreground element in the ballet instead of a background accompaniment and could have brought about—if its example had been followed up—a revitalization of ballet music almost as significant as the operatic reform Gluck launched in *Orfeo ed Euridice* a year later. In the ballet, the fight scene near the beginning and the dance of the Furies at the end (itself later incorporated into *Orfeo*) have a concentrated intensity of musical expression, and the graveyard scene has a degree of imaginative orchestration that is unsurpassed in any other music at that date.

For reasons that belong more properly to the history of ballet, Gluck's influence on its future course was less fruitful than in opera. Perhaps his example, nevertheless, encouraged the participation of such composers as Mozart with *Les Petits Riens* (1778; "Sweet Nothings") and Beethoven in *The Creatures of Prometheus* (1801). It was, otherwise, an era theatrically dominated by opera of various kinds, so that there was at first little call for music in relation to spoken drama. Haydn, however, composed some music (1796) for an early German translation of Alexander Bicknell's *The Patriot King, or Alfred and Elvida,* and Mozart contributed a suite of superior choral and orchestral music for *Thamos, König in Ägypten* (1773; *Thamos, King of Egypt*), which was never used for the play in his lifetime but which has survived where the play has long been forgotten.

Eighteenth-century opera nourished musical developments in the theatre chiefly through establishing regular orchestras of some quality. Outside of cities with more than one theatre, such as Vienna, Prague, Paris, and London, the numerous court theatres needed to keep a resident and costly orchestra reasonably occupied, with interest divided between opera and drama. It therefore became customary either to commission incidental music for existing plays whenever possible (especially the historical classics, as more and more of them were translated from one language to another) or to commission new plays that would incorporate ample provision for orchestral and sometimes choral music.

ROMANTIC EXPANSION

Examples of commissioned incidental music have previously been cited in Beethoven's music for *Egmont,* which belongs to the first category just mentioned, and Schubert's for *Rosamunde,* which is in the second. The practice spread as a matter of rivalry and prestige to cities without a court but which maintained a municipal theatre (for example, Hamburg and Leipzig) and to other countries with a thriving theatrical interest and ample funds. In Russia, for instance, Mikhail Glinka composed music to *Prince Kholmsky* (1840), an otherwise obscure drama by Count Kukolnik, and the Shakespeare repertory brought the collaboration of such composers as Mily Balakirev (*King Lear,* 1861) and Tchaikovsky (*Hamlet,* 1891).

France and England, having different systems of patronage, produced different results during the 19th century. English theatre music was confined for most of its course

to a taste for crude melodrama and burlesque at a low level, apart from a sporadic interest in mostly imported opera and ballet. Arthur Sullivan, however, provided incidental music for Shakespeare plays as well as cultivating, in his collaboration with the author W.S. Gilbert, a native variety of operetta derived from the French model. France fared somewhat better with the popularity of opéra bouffe, and the birth of romantic ballet in Paris also kindled a new kind of theatre interest, even though its musical quality was usually secondary.

Gilbert and Sullivan operetta

Music for Romantic ballet developed in two directions. From the time of the French composer Adolphe Adam's score for *Giselle* (1841), ballet composers made rudimentary attempts to express mood and scene, to create dramatic tension, and to characterize personality in music. The general level was somewhat raised by the French composer Léo Delibes in his music for *Coppélia* (1870) and more especially for *Sylvia* (1876); the latter was a score that Tchaikovsky came to know and admire. The second feature in many ballet scores of the period was the attempt to compose suitably flavoured music to match the new growth of interest in national and regional dances, which were regularly incorporated into the ballets.

Music in 19th-century ballet reached its peak of achievement with Tchaikovsky, whose instinct for the theatre was probably stronger than his talent for the subtleties of symphonic argument. He treated the art of ballet as worthy of real musical imagination and told a colleague who adversely likened some of his *Symphony No. 4* to ballet music: "I cannot understand why the term should be associated with something reprehensible. There is such a thing as good ballet music." Tchaikovsky demonstrated its possibilities in three original scores for ballet that enjoy continuing universal popularity in the theatre: *Swan Lake* (first performed 1877); *The Sleeping Beauty* (1890); and *The Nutcracker* (1892).

Swan Lake achieved lasting success only after the composer's death—a fact which accounts for the recurring problems of the relationship of music to choreography, because some of the original musical sequence was changed for the 1895 production at St. Petersburg (now Leningrad), from which most current versions of the ballet are descended. The two later scores benefitted from detailed choreographic specifications. According to Stravinsky, *The Sleeping Beauty* is "the convincing example of Tchaikovsky's great creative power," and *The Nutcracker* in its theatre context has a narrative vividness much beyond the limited charm of the concert suite that Tchaikovsky arranged from it.

Adolphe Adam's contribution to the development of ballet music had its parallel in the sphere of romantic operetta. By incorporating a measure of frivolous vaudeville into the otherwise conventional comedy of *Le Châlet* (1834), Adam stimulated a popular taste for what became the mainstream of operetta. Its source was in Paris, and it flowed in turn principally to Vienna, to London and thence to North America, submerging the German singspiel and its Scandinavian offshoots but leaving the Spanish zarzuela to cultivate its own regional idiom.

Adam's enterprise in opening a theatre in 1847 to stage his own works and those of other young composers disdained by the operatic establishment in Paris was brought to a premature end by the political uprising a year later, but Offenbach was poised to take advantage of the subsequent situation. He opened his Bouffes-Parisiens theatre in 1855, whence travelled such immediate hits as *Orpheus in the Underworld, La Belle Hélène, La Grande-Duchesse de Gérolstein*, and *La Périchole* over almost all of Europe. The Parisian operetta was principally continued to the end of the century by Charles Lecocq and André Messager.

Offenbach's influence in Europe

Offenbach meanwhile had paid several visits to Vienna from 1858, when Franz von Suppé was quick to turn the French model to local advantage, notably with *Die schöne Galatea* (1865; *The Beautiful Galatea*). The younger Johann Strauss was eventually persuaded to follow this trend by turning the Viennese craze for his waltzes, polkas, and other social dances to theatrical purpose. With the aid of an unusually good libretto, Strauss created the supreme example of Viennese operetta in *Die Fledermaus*. The Viennese tradition was continued in turn by Franz Lehár and Emmerich Kálmán.

Offenbach's influence extended southward to Bohemia, where the composer Bedřich Smetana compared his first song-and-dialogue version of *The Bartered Bride* (1866) to an Offenbach operetta, and Antonín Dvořák composed an outstanding but little-known example in *The Peasant a Rogue* (1878). Otherwise, no specialist or specially distinguished composer of operetta emerged in southeastern Europe, nor in Poland or Russia, where there were only occasional and tenuous contacts with forms of musical theatre other than full-scale opera and ballet for the moneyed classes.

The northward spread of the Offenbach model reached England, where it influenced the character of Sullivan's first operetta, *Cox and Box* (1867). It led in due course to the success of *H.M.S. Pinafore* (1878), in collaboration with Gilbert, and the subsequent line of "Savoy operettas" (their collective nickname derived from the London theatre where they were first performed by the D'Oyly Carte Opera Company, Ltd.). England made no other contributions of comparable musical interest to the operetta repertory until it was overtaken in due course by the trend to musical comedy in the 20th century.

English operetta represented by Gilbert and Sullivan nevertheless put down fresh roots in North America after the D'Oyly Carte company first travelled there in 1879, in the wake of a pirated version of *H.M.S. Pinafore*. The impact of Sullivan's music in New York and Boston was comparable to that of Offenbach in Vienna. Instead of stimulating an American equivalent, it first opened the way to 20 years or so of European imports to the American stage. The composers Reginald De Koven and Victor Herbert later established a short-lived local counterpart to European operetta, before it was overtaken by the indigenous idiom of American musical theatre noted earlier.

The language of ragtime and early jazz, with its rhythmic syncopation and varying degrees of harmonic innovation within a common musical vernacular, brought the first new element to the idiom of musical theatre (in musical comedy) since the emergence of national folk characteristics in 19th-century Europe. As the trends already described succeeded one another during the 20th century, in and out of fashion, and the musical theatre tried to reconcile the nostalgia for its past heritage with the need to experiment in search of a viable future, the immediate present can be seen to represent only one turn of a larger wheel across seven or eight centuries. The religious rock musical of the contemporary musical scene is but a variant of the mysteries and miracle plays that initiated all our modern forms of musical theatre when Western civilization first groped its way out of the Dark Ages.

BIBLIOGRAPHY. MARK LUBBOCK, *The Complete Book of Light Opera*, with an American section by DAVID EWEN (1962), gives synopses and background dates of numerous European and American operettas and musicals. GERVASE HUGHES, *Composers of Operetta* (1962), is a reliable critical and historical account of classical operetta. STANLEY GREEN, *The World of Musical Comedy*, rev. ed. (1969), gives a comprehensive account of the American musical in terms of its composers; and an informed discussion of its character and constituents is in LEHMAN ENGEL, *The American Musical Theatre* (1967). IVOR GUEST, *The Romantic Ballet in Paris* (1966), contains much little known background on music of its period; and NATALIA ROSLAVLEVA, *Era of the Russian Ballet Music* (1966), performs a similar but less well-documented service for its own subject. ROGER FISKE, *Ballet Music* (1958), is the only musically informed analysis to date of certain major ballet classics. HANNS EISLER, *Composing for the Films* (1947), remains a useful account of applied research. The wider historical aspects of theatrical music as a whole are included in the various volumes of the *New Oxford History of Music*, projected 11 vol. (1957–); and are comprehensively summarized in the excellent one-volume *Man and His Music* by ALEC HARMAN and WILFRID MELLERS (1962).

(N.Go.)

Music, Western

All ancient civilizations entered historical times with a flourishing musical culture. That the earliest writers explained it in terms of legend and myth is evidence of the remote beginnings of the art of sound. Among the speculations about its origin, the more plausible are that it began as a primitive form of communication, that it grew out of a device to expedite communal labour, or that it originated as a powerful adjunct to religious ceremonies. While such theories must necessarily remain speculative, it is clear, despite the prehistoric musical artifacts found in central Europe, that the cradle of Western music was the fertile crescent cupping the eastern end of the Mediterranean Sea. There the Mesopotamian, Egyptian, and Hebrew nations, among others, evolved political, social cultures that were absorbed by the conquering Greeks and, in turn, by the Romans, who transported a relatively sophisticated art form back to western Europe.

In all of these early cultures the social functions of music were essentially the same since their climate, geographic location, cultural pace, and mutual influences produced many more social similarities than differences. Magico-religious origins The primary function of music was apparently religious, ranging from heightening the effect of "magic" to ennobling liturgies. The other musical occasions depicted in both pictures and written accounts were equally functional: stirring incitements to military zeal, soothing accompaniments to communal or solitary labour, heightening aids to dramatic spectacles, and enlivening of social gatherings involving singing and dancing. In every case musical sounds were an adjunct either to bodily movement (dance, march, game, or work) or to song. Many centuries were to pass before pleasure in euphonious sound became an end in itself.

The establishment of Western musical traditions

ROOTS IN ANTIQUITY

Ancient Near East and Egypt. The inhabitants of the Mesopotamian region around the Tigris and Euphrates rivers (the Sumerians, the Babylonians, and the Assyrians) flourished from *c.* 3500 to *c.* 500 BC. Their pictures and the few surviving artifacts indicate that they had instruments of every basic type—idiophones, whose sound is made by resonating as a whole; aerophones, which resonate a column of blown air; chordophones, with strings to be plucked or struck; and membranophones, made of stretched skins over a resonating body. An undecipherable hymn engraved in stone, dating from *c.* 800 BC, is evidence that a primitive system of musical notation had been developed.

The Egyptians, entering historical times about 500 years later than the Mesopotamians, enjoyed all of the same types of activities and instruments, as may be deduced from numerous written references to music as well as seen on many artifacts, especially the pictures preserved on pottery utensils.

The musical culture of the Hebrew peoples, recorded from about 2000 BC and documented primarily in the Old Testament, was more directly influential in the West because of its adoption and adaptation into the Christian liturgy. Because of the prohibition of Jewish religious law against the making of "graven images," there are very few surviving artifacts or pictures. Among the established practices of the temple service still current in the synagogue are the extensive use of the shofar (a ritualistic ram's-horn trumpet) and the singing of passages from the Pentateuch (the first five books of the Old Testament), prayers, and songs of praise.

Ancient Greece. Of the eastern Mediterranean cultures, it was undoubtedly that of the Greeks that furnished the most direct link with musical development in western Europe, by way of the Romans, who défeated them, but adopted much of Greek culture intact. Entering historical times relatively late, *c.* 1000 BC, the Greeks soon dominated their neighbours and absorbed many elements of earlier cultures, which they modified and combined into an enlightened and sophisticated civilization. The two basic Greek religious cults—one devoted to Apollo, the other to Dionysus—became the prototypes for the two aesthetic poles, classical and romantic, that have contended throughout Western cultural history. The Apollonian and Dionysiac expression Apollonians were characterized by objectivity of expression, simplicity, and clarity, and their favoured instrument was the kithara, a type of lyre. The Dionysians, on the other hand, preferred the reed-blown aulos and were identified by their subjectivity, emotional abandon, and sensuality.

The prevailing doctrine of ethos, as explained by ancient Greek philosophers such as Plato and Aristotle, was based on the belief that music has a direct effect upon the soul and actions of mankind. As a result, the Greek political and social systems were intertwined with music, which had a primary role in the dramas of Aeschylus, Sophocles, Euripides, and Aristophanes. And the Grecian educational system was focussed upon *musica* and *gymnastica*, the former referring to all cultural and intellectual studies, as distinguished from those related to physical training.

To support its fundamental role in society, an intricate scientific rationale of music evolved, encompassing tuning, instruments, modes (melodic formulas based on certain scales), and rhythms. The 6th-century-BC philosopher and mathematician Pythagoras was the first to record the vibratory ratios that established the series of notes still used in Western music. From the total gamut of notes used were derived the various modes bearing the Origin of musical modes names of Grecian tribes—Dorian, Phrygian, Lydian, etc. The rhythmic system, deriving from poetry, was based on long–short relationships rather than strong–weak accentual metre. After Pythagoras, Aristoxenus was the major historian and theoretician of Greek music.

Ancient Rome. When the musical culture of the eastern Mediterranean was transplanted into the western Mediterranean by the returning Roman legions, it was inevitably modified by local tastes and traditions. In most cases, the resulting practices were more limited than their models. The diatonic (seven-note) scale, for example, became the standard, displacing the chromatic and enharmonic structures of the Grecian system. Of particular consequence was the new concept of metre as a series of equal durations, with emphasis being determined by accent (stress) rather than by duration.

An inventory of the musical heritage transplanted from the ancient East (particularly Greece) to Rome reveals the rich treasure inherited: an acoustical theory that accounted for the identification and classification of tones; a concept of tonal organization resulting in the system of modes; principles of rhythmic organization; basic principles of instrument construction; a system of notation that conveyed all necessary indications of pitch and duration; and a large repertory of melodies to serve as models for further composition.

THE MIDDLE AGES

Monophonic liturgical chant. With the decline of the Roman Empire, the institution destined to perpetuate and expand the musical heritage of antiquity was the Christian Church, but it was not a unified process. Many of the cultural centres of the Western church developed distinctive characteristics while sharing the common heritage of the Hebrew liturgy and Greek culture. In Milan, for example, metrical hymnody, as distinguished from the earlier practice of unmetred psalmody, was cultivated, particularly under the influence of the 4th-century Bishop Ambrose, who first attempted to codify the growing repertory of chants. This body of Milanese church music, therefore, came to be called Ambrosian chant. Somewhat later a unique style and repertory known as Mozarabic chant evolved in Spain, and in France the Gallican style prevailed.

But the mainstream of church music was the type of chant practiced in Rome. Beginning in the late 6th century, according to tradition, with Pope Gregory I, the vast Gregorian chant number of traditional melodies that became the foundation for the later development of Western art music were codified and organized. A systematic organization of tonal materials also was gradually accomplished, resulting in

the eight church modes. Each melody was assigned a specific function in the services of the liturgical year—some for the mass and some for the divine offices such as Matins, Vespers, and Compline. After a period of assimilation, the Gregorian chant repertory began a process of expansion in the 9th century, when the practice of troping originated. A trope is either a text or a melodic section added to a pre-existing melody or a combination of text and music incorporated into existing liturgical music. It is not surprising that church musicians, after years of singing traditional chants, should want to express themselves by adding words to vocalized melodies. Perhaps the motive was more functional: the added syllables would make the long textless passages easier to remember. Tuotilo (died 915), a monk of St. Gall in Switzerland, is credited with the invention of tropes. Notker Balbulus (died 912) is notable for his association with the sequence, a long hymn that originated as a trope added to the final syllable of the Alleluia of the mass.

Development of polyphony. At the same time that the Gregorian repertory was being expanded by the interpolation of tropes and sequences, it was being further enriched by a revolutionary concept destined to give a new direction to the art of sound for hundreds of years. This concept was polyphony, or the simultaneous sounding of two or more melodic lines. The practice emerged gradually during the Dark Ages, and the lack of definite knowledge regarding its origin has brought forward several plausible theories: it resulted from singers with different natural vocal ranges singing at their most comfortable pitch levels; it was a practice of organists adopted by singers; or it came about when the repetition of a melody at a different pitch level was sung simultaneously with the original statement of the melody. Whatever motivated this dramatic departure from traditional monophony (music consisting of a single voice part), it was an established practice when it was described in *Musica enchiriadis* (*c.* 900), a manual for singers and one of the major musical documents of the Middle Ages. To a given plainsong, or *vox principalis*, a second voice (*vox organalis*) could be added at the interval (distance between notes) of a fourth or fifth (four or five steps) below. Music so performed was known as organum. While it may be assumed that the first attempts at polyphony involved only parallel motion at a set interval, the *Musica enchiriadis* describes and gives examples of two-part singing in similar (but not exactly parallel) and contrary movement—evidence that a considerable process of evolution had already taken place.

The beginnings of polyphony

The next major source of information was the *Micrologus*, written in the early 11th century by the Italian monk and musical theorist Guido of Arezzo. This work documented principles that were crucial to the further development of polyphony. Rhythmic independence was added to melodic independence, and the added voice might sing two or more tones to one in the original plainsong. During the half century after Guido's death, developments came more rapidly as the plainsong chant became the lower rather than the upper voice. After the emancipation of the organal part, *vox organalis*, its ultimate freedom was reached in the organums of the monastery of Saint-Martial in Limoges, France, where the plainsong part was reduced to the role of sustaining each tone while the organal part indulged in free melismata (groups of notes sung to a single syllable), either improvised or composed. This new style was called *organum purum*.

The Notre-Dame school. Early in the 12th century the centre of musical activity shifted to the church of Notre-Dame in Paris, where the French composer Léonin recorded in the *Magnus Liber Organi* ("Great Book of Organum") a collection of two-part organums for the entire church year. A generation later his successor, Pérotin, edited and revised the *Magnus Liber*, incorporating the rhythmic patterns already well-known in secular music and adding more than one part to the cantus firmus (the "given" or pre-existing plainsong melody). When metre was applied to the original plainsong as well as to the *vox organalis*, the resulting form was called a clausula. Then, when words were provided for the added part or parts, a clausula became a motet. At first the words given to the motet were a commentary in Latin on the text of the original plainsong tenor (the voice part "holding" the cantus firmus; from Latin *tenere*, "to hold"). Later in the 13th century the added words were in French and secular in nature. Finally, each added part was given its own text, resulting in the classic Paris motet: a three-part composition consisting of a portion of plainchant (tenor) overlaid with two faster moving parts, each with its own secular text in French. At the same time another polyphonic form, the conductus, was flourishing. It differed from a motet in that its basic part was not plainsong and that all parts sang the same Latin text in note-against-note style. The conductus gradually disappeared with the rise of the motet, which apparently served both liturgical and secular functions.

The beginnings of the motet

Ars Nova. When the influential treatise *Ars Nova* ("New Art") by the composer Philippe de Vitry appeared early in the 14th century, the preceding epoch acquired its designation of Ars Antiqua (Old Art), for it was only in retrospect that the rapid developments of the century and a half from *c.* 1150 to *c.* 1300 could appear as antiquated. De Vitry recorded the innovations of his day, particularly in the areas of metre and harmony. While 13th-century music had been organized around the triple "modal" rhythms derived from secular music and a harmonic vocabulary based on "perfect" consonances (unison, fourth, fifth, octave), the New Art of the 14th century used duple as well as triple divisions of the basic pulse and brought about a taste for harmonious intervals of thirds and sixths.

The musical centre of 14th-century Italy was Florence, where a blind organist, Francisco Landini, and his predecessors and contemporaries Giovanni da Cascia, Jacopo da Bologna, and Lorenzo and Ghirardello da Firenze were the leading composers of several new forms: madrigals (contrapuntal compositions for several voices), ballatas (similar to the French virelai), and caccias (three-voice songs using melodic imitation).

Monophonic secular song. Secular music undoubtedly flourished during the early Middle Ages, but, aside from sporadic references, the earliest accounts of such music in the Western world described the music of the goliards, itinerant minor clerics and students who, from the 7th century on, roamed the land singing and playing topical songs dealing with love, war, famine, and other issues of the day. The emergence in France of a fully developed secular-musical tradition about the beginning of the 12th century is evidence that the art had been evolving continuously before that time. Partially motivated by the attitude of chivalry engendered by the Crusades, a new life-style began among the nobility of southern France. Calling themselves troubadours, they circulated among the leading courts of the region, devoting themselves to writing and singing poetry in the vernacular. The movement flourished in Provence during the 12th and 13th centuries. About the middle of the 12th century, noblemen of northern France, most notably Adam de La Halle, took up the pastime, calling themselves trouvères. In Germany a similar group known as Minnesingers, represented by Walther von der Vogelweide, began their activities about 1150 and continued for almost a century after their French counterparts had ceased composing. Late in the 13th century the burgher class in Germany began imitating the aristocratic Minnesingers. Calling themselves Meistersingers, they flourished for more than 500 years, organizing themselves into fraternities and following strict rules of poetry, music, and performance. The most famous of them, Hans Sachs, was immortalized in the 19th century in Richard Wagner's opera *Die Meistersinger von Nürnberg*. Relatively little is known of similar activities in Italy, Spain, and England. Closely associated with the entertainments of the aristocratic dilettantes were the professional musicians of the peasant class called jongleurs and minstrels in France, *Gaukler* in Germany, scops and gleemen in England.

Troubadours and Meistersingers

The musical style established by the troubadours—which was monophonic, of limited range, and sectional in

structure—was adopted by each of the succeeding groups. Of particular significance in view of its influence on polyphonic music was the metric system, based on six rhythmic modes. Supposedly derived from Greek poetic metres such as trochaic (long–short) and iambic (short–long), these modes brought about a prevailing triple metre in French music; German poetry produced duple as well as triple metre. A great variety of formal patterns evolved, in which musical and poetic structure were closely related. The most characteristic was the ballade, called *Bar* form in Germany, with an AAB structure. This type, along with the rondeau (song for solo voice with choral refrain) and the similar virelai (an analogue of the Italian ballata), was destined to become a favoured form of composers of polyphony such as Guillaume de Machaut, the universally acknowledged master of French music of the Ars Nova period. Machaut also continued the composition of motets, organizing them around recurrent rhythmic patterns (isorhythm), a major structural technique of the age. The beginnings of an independent instrumental repertory during the 13th century are represented by the estampie, a monophonic dance form almost identical in style to the vocal secular music.

THE RENAISSANCE PERIOD

The term Renaissance, in spite of its various connotations, is difficult to apply to music. Borrowed from the visual arts and literature, the term is meaningful primarily as a chronological designation. Some historians date the beginning of the musical Renaissance at about 1400, some with the rise of imitative counterpoint, about 1450. Others relate it to the musical association with Humanistic poetry at the beginning of the 16th century, and still others reserve the term for the conscious attempt to recreate and imitate supposedly classical models that took place about 1600.

The court of Burgundy. No one line of demarcation is completely satisfactory, but, adhering to commonly accepted usage, one may conveniently accept as the beginning of the musical Renaissance the flourishing and secularization of music at the beginning of the 15th century, particularly at the court of Burgundy. Certainly many manifestations of a cultural renaissance were evident at the time: interest in preserving artifacts and literature of classical antiquity, the waning authority and influence of the church, the waxing Humanism, the burgeoning of urban centres and universities, and the growing economic affluence of the states of western Europe.

Rivalry between church and state in musical patronage

As one manifestation of their cultivation of elegant living, the aristocracy of both church and state vied with one another in maintaining resident musicians who could serve both chapel and banqueting hall. The frequent interchange of these musicians accounts for the rapid dissemination of new musical techniques and tastes. Partly because of economic advantages, Burgundy and its capital, Dijon, became the centre of European activity in music as well as the intellectual and artistic focus of northern Europe during the first half of the 15th century. Comprising most of eastern France and the Low Countries, the courts of Philip the Good and Charles the Bold attracted the leading musicians of western Europe. Prime among them was Guillaume Dufay, who had spent some time in Rome and Florence before settling in Cambrai about 1440. An important contemporary of Dufay was Gilles Binchois, who served at Dijon from about 1430 until 1460. The alliance of Burgundy with England accounted for the presence on the Continent of the English composer John Dunstable, who had a profound influence on Dufay. While the contributions of the English to the mainstream of continental music are sparsely documented, the differences in style between Dufay and his predecessor Machaut are partially accounted for by the new techniques and, especially, the richer harmonies adopted by the Burgundian composers from their English allies.

New religious musical forms. The social circumstances of the age determined that composers would devote their efforts to the mass, the motet, and the chanson (secular French song). During the first half of the 15th century, the mass became established as a unified polyphonic setting of the five main parts of the Ordinary of the mass (Kyrie, Gloria, Credo, Sanctus, Agnus Dei), with each movement based on either the relevant portion of plainsong or, reflecting the dawning Renaissance, a secular song such as the popular "L'Homme armé" ("The Armed Man") and "Se la face ay pale." Still reflecting medieval practices, the pre-existing melody (or cantus firmus) was usually in the tenor (or lowest) part and in long, sustained tones, while the upper parts provided free elaboration. Dufay's nine complete settings of the mass, compared with Machaut's single setting, give a clear indication of the growing importance of the mass as a musical form. The motet became simply a setting of a Latin text from Scriptures or the liturgy in the prevailing polyphonic style of the time. It was no longer necessarily anchored to a plainsong tenor; the composer could give free reign to his invention, although some did, of course, resort to older techniques.

Importance of the mass

Secular music. It was in secular music that giant strides took place. While their chansons continued the tradition of rondeaux, virelais, and ballades, Dufay and his contemporaries added free forms divorced from the ordered patterns of the Ars Antiqua and Ars Nova periods.

Among the distinctive features of Burgundian musical style was the prevailing three-part texture, with melodic and rhythmic interest centred in the top part. Because it was so typical of secular songs, this texture is commonly referred to as "ballade style" whether it appears in mass, motet, or chanson. Its possible stylistic implication is that a solo voice sang the upper melody, accompanied by instruments playing the lower parts, although no documents remain to establish exactly how the music was performed. There was probably no standard performing medium: all parts may have been sung; some or all may have been doubled by instruments; or there may have been one vocal part supported by instrumental accompaniment.

The Franco-Flemish school. A watershed in the history of music occurred about the middle of the 15th century. The fall of Constantinople in 1453 and the end of the Hundred Years' War at about the same time increased commerce from the East and affluence in the West. Most significant musically was the pervasive influence of musicians from the Low Countries, whose domination of the musical scene during the last half of the 15th century is reflected in the period designations the Netherlands school and the Franco-Flemish school. These musicians travelled and resided throughout Europe in response to their great demand at princely courts, including those of the Medici family in Florence and the Sforzas in Milan. Further dissemination of knowledge resulted from the invention and development of printing.

Travelling musicians at courts

The leading composers, whose patrons were now members of the civil aristocracy as well as princes of the church, were Jean d'Okeghem, Jakob Obrecht, and, especially, Josquin des Prez. D'Okeghem, born and trained in Flanders, spent most of his life in the service of the kings of France and was recognized by his contemporaries as the "Prince of Music." Obrecht remained near his birthplace in the Netherlands, going occasionally to Italy in the retinue of Duca Ercole I (Duke Hercules I) of Ferrara. More typical of the peripatetic Netherlanders was the career of Josquin, the most influential and most celebrated composer of the period. After his training at St. Quentin, he served the Sforza family in Milan, the papal choir in Rome, Ercole I, and King Louis XII of France before returning to his native Flanders in 1516. These three composers and several of their contemporaries hastened the development of the musical techniques that became the basis of 16th-century practice and influenced succeeding musical developments.

Rather than the three parts typical of most Burgundian music, four parts became standard for vocal polyphony in the late 15th century. The fourth part was added below the tenor, increasing the total range and resulting in greater breadth of sound. The presence of the four parts also allowed for contrasts of texture such as the "duet style" so characteristic of Josquin, when the two upper parts might sing a passage alone and be echoed by the two

lower parts alone. The emergence of the technique of imitation (one voice repeating recognizably a figure heard first in another voice) as the chief form-generating principle brought about more equality of parts. At the same time "familiar style," in which all parts move together in chords, provided a means of textural contrast. The great variety of rhythmic techniques that evolved during the 14th and early 15th centuries made possible a wide range of expression—from quiet tranquillity for sacred music to lively and spirited secular music. Knowledge of the musical practices comes not only from the thousands of surviving compositions but from informative treatises such as the 12 by the composer Johannes Tinctoris (1436–1511), one of which, *Terminorum musicae diffinitorium* (*c.* 1475), is the earliest printed dictionary of musical terms.

Wider range of expression through rhythmic developments

The chief forms of vocal music continued to be the mass, the motet, and the chanson, to which must be added other national types that developed during the 15th century—the villancico (secular poetry set for voice and lute or for three or four voices) in Spain and the frottola (a simple, chordal setting in three or four parts of an Italian text) in Italy. The emergence of the frottola in northern Italy led to the development of the Renaissance madrigal, which impelled that country to musical supremacy in Europe.

Instrumental music. At the same time, an independent instrumental idiom was evolving. While instruments had been in common usage throughout the Middle Ages, their function was primarily to double or to substitute for voices in vocal polyphonic music or to provide music for dancing. Techniques unsuitable for voices were doubtless part of an instrumentalist's musical vocabulary, but most such music was improvised rather than being written. Although there are a few sources of instrumental music dating from the 13th and 14th centuries, the earliest relatively extensive documentation comes from the 15th century, particularly from German sources, such as the *Buxheimer Orgelbuch* and Konrad Paumann's *Fundamentum organisandi* (*Fundamentals of Organ Playing*). The compositions in both collections are of two basic types, arrangements of vocal works and keyboard pieces entitled *Praeambulum* (*Prelude*).

During the course of the 16th century, instrumental music burgeoned rapidly, along with the continually developing idiomatically instrumental techniques, such as strongly accented rhythms, rapid repeated tones and figures, angular melodic lines involving wide intervallic skips, wide ranges, long, sustained tones and phrases, and much melodic ornamentation.

Musical forms. Dance forms, a continuation of a tradition unbroken since the beginnings of recorded music history, were most characteristically composed in pairs, although single dances as well as embryonic suites of three or more dances appeared. The pairs usually consisted of pieces in contrasting tempo and metre that often were unified by sharing a common melody. Common dance pairs included the pavane and galliard, the allemande and courante, and the basse danse and tourdion.

Preludes continued as a major form of organ music and were joined by the fantasia, the *intonazione*, and the toccata in a category frequently referred to as "free forms" because of the inconsistency and unpredictability of their structure and musical content—sections in imitative counterpoint, sections of sustained chords, sections in virtuoso figuration. If a distinction must be made, it might be said in very general terms that the fantasia tended to be more contrapuntal while the toccata (or "touch piece") featured passages designed to demonstrate the performer's agility, although the designations were freely interchangeable. To the same category belong the descriptive pieces such as *The King's Hunt*, which featured naive musical representations of natural sounds.

Fantasias and toccatas

The ricercar and the canzona, generally referred to as fugal forms because of their relationship to the principle of the fugue (that of melodic imitation), arose out of the growing understanding of and dependence on imitation as a unifying structural technique. Although these designations were applied to a great variety of pieces—some identical in style to the fantasia or prelude—the classic ricercar of the 16th century was virtually an instrumental motet, slow and churchlike in character and consisting of a number of sections, each utilizing imitation. The canzona followed the same structural principle but was a lively counterpart to the chanson, with the sections sometimes in contrasting tempo and metre. Cantus firmus compositions were based upon pre-existing melody. During the 16th century most were designed for liturgical usage but were based upon both secular melodies and plainsong. In most cases the cantus firmus was sounded in long, sustained tones while the other part or parts added decorative contrapuntal lines. The organ mass, in which the choir and the organ alternated lines of the liturgical text, was a popular practice.

Variations also often used a pre-existing melody but differed from cantus firmus compositions in that the melody was much shorter and was repeated a number of times, each time with different accompanying parts. The two basic types during the Renaissance were the plain, or melodic, variations and the ground. In the former, the chosen melody usually appeared in the top part and was varied in each repetition with ornamentation and melodic figuration or with changing accompaniments. The ground, or ground bass, was a simple melodic pattern sounded in the lowest part, which served as a foundation for imaginative figuration in the upper parts.

Solo and ensemble instruments. The four major vehicles for instrumental music of the period were the lute, the organ, stringed keyboard instruments, and instrumental ensembles. Most popular by far was the lute, which could produce the major elements of instrumental style except for long, sustained tones. Noteworthy composers of lute music included Luis Milán in Spain, Arnold Schlick in Germany, and John Dowland in England. The organ, because of its close association with liturgical music, continued to be an important instrument, and its literature includes all of the formal types except dances. Among the leading organ composers were the Germans Paumann, Schlick, and Paul Hofhaimer, the Italians Claudio Merulo and Andrea and Giovanni Gabrieli, the Spaniard Antonio de Cabezón, and the Englishman John Bull.

The two basic classes of stringed keyboard instruments were the harpsichord (virginals, spinet, clavecin, clavicembalo), with quill-plucked strings, and the clavichord, with strings struck by thin metal tongues. Keyboard instruments were highly capable of idiomatically instrumental effects and flourished, particularly in England, from the last half of the 16th century onward, thanks to the composers William Byrd, John Bull, and Orlando Gibbons. A major manuscript source of the keyboard works of these masters is the famous *Fitzwilliam Virginal Book* of the 17th century.

Instrumental ensembles of the Renaissance were not standardized, although consorts (groups) of viols, of woodwind instruments such as recorders and shawms (loud oboes), or of brass-family instruments such as the cornet and sackbut (early trombone) were common. More common, however, were mixed consorts of various types of instruments, depending on the players available. All types of instrumental forms were performed by ensembles except for the prelude and the toccata, which were essentially keyboard works. Representative composers included the Gabrielis and Gibbons.

Vocal music in the 16th century. At the beginning of the 16th century the style of vocal music was generally uniform because of the pervading influence of Netherlanders during the preceding half century. That uniformity persisted well into the late Renaissance but was gradually superseded by emerging national differences, new forms, and the increasing importance of Italy as a musical centre during the last half of the 16th century.

The rapid accumulation of new musical techniques and resources produced a wide vocabulary of artistic expression, and the invention of music printing helped the rapid dispersal of new techniques. In an age in which music was an essential social grace, composers wrote more secular music, in which fewer technical restrictions were in force

and experimentation and novelty were applauded. Advances were particularly apparent in venturesome harmonies as chromaticism (the use of notes not belonging to the mode of the composition) sounded the death knell of the modal system.

Harmonic experiment

Liturgical practice dictated that the mass and the motet remain the chief forms of sacred vocal music. Compared with secular music, their style was conservative, but inevitably some of the newer secular techniques crept in and figured effectively in the music of the Counter-Reformation within the Roman Catholic Church.

Four distinct types of mass settings were established during the century. Two types were continuations of earlier practice: the tenor mass, in which the same cantus firmus served for all five portions of the Ordinary of the mass, and the plainsong mass, in which the cantus firmus (usually a corresponding section of plainsong) differed for each portion. Reflecting the more liberal attitudes of the Renaissance were the free mass, with no borrowed materials, and the parody mass, in which the entire polyphonic web was freely adapted from a motet or a secular composition. In all cases when a cantus firmus was used, the pre-existent melody might appear in its original form or in paraphrased version, with tones added, omitted, or altered. As a result of the upheaval in the church caused by the Reformation, new forms derived from established models appeared in Protestant worship: the German Lutheran chorale (hymn tune, arranged from plainsong or a secular melody), the chorale motet, English anthems (Anglican form of motet) and services, and the psalm tunes in Calvinist areas.

Italy. While not young in a chronological sense, the musical life of Italy was reborn at the beginning of the 16th century after a century of relative dormancy. The frottola remained the prevailing secular form in northern Italy for the first three decades of the century.

When the humanistic poets, seeking a more refined expression, and the Netherlanders and composers trained by them, applying a more sophisticated musical technique, turned their efforts to the frottola, the result was the madrigal. The name was borrowed from the 14th-century form, but there was no resemblance in poetic or musical structure. Compared to the frottola, the earliest Renaissance madrigals, dating from about 1530, were characterized by quiet and restrained expression. Usually written for three or four voices, they were mostly homophonic (melody supported by chords) with occasional bits of imitation. Among the early madrigal composers were several Flemish composers resident in Italy, among them Adriaan Willaert, Jacques Arcadelt, and Philippe Verdelot. About 1560 the normal number of parts increased to five or six, and the texture became more consistently polyphonic. At the same time, more attention was given to expressive settings of the text, notably in the madrigals of Cipriano de Rore, Philippe de Monte, and the Gabrielis. During the last two decades of the century and continuing until the middle of the 17th century, the musical style of the madrigal changed appreciably. The late madrigals were of a very dramatic nature, featuring colouristic effects, vivid word-painting, and extensive chromaticism. Their declamatory character dictated a return to a more homophonic style. Noteworthy among the many composers of the late madrigal were Luca Marenzio, Don Carlo Gesualdo, and Claudio Monteverdi.

Birth of the madrigal

During the course of the century simpler secular forms, such as the villanella, the canzonetta, and the balletto, appeared in Italy, largely as a reaction against the refinement, complication, and sophistication of the madrigal. They reverted to the chordal style of the frottola, often with intentionally parodistic lyrics. The balletto was particularly distinguished by a refrain of nonsense syllables such as "fa la la."

England. Most of the Italian forms, along with their designations, were adopted by Elizabethan England during the last half of the 16th century. Most leading English composers, from William Byrd and Thomas Morley to John Wilbye, Thomas Weelkes, and Orlando Gibbons, contributed to the vast treasury of English secular music. Morley is particularly important as the editor of the most significant collection of English madrigals, the *Triumphes of Oriana,* published in 1603 and dedicated to Queen Elizabeth I (Oriana). These pieces correspond in style roughly to the middle-period Italian madrigal. English counterparts of the canzonetta and balletto were the canzonet and ballett. A late-16th-century innovation in both Italy and England was the ayre (air), a simple, chordal setting especially suitable for a solo voice with a lute or a consort of instruments playing the other parts. John Dowland and Thomas Campion were notable composers of ayres.

France. The French counterpart of Italian and English madrigals was the polyphonic chanson, a continuation of the chief medieval and early Renaissance form of secular music. Revitalized by composers such as Josquin, Clément Janequin, and Claudin de Sermisy, the chanson developed several distinctive features: a clearly delineated sectional structure with some repetition of sections, much vivid programmatic writing, and occasional use of irregular metric organization. The irregular metric structure, called *musique mesurée,* was used for maintaining faithfully the accentuation of the poetry and reflects the traditional primacy of textual over musical considerations in French music.

The chanson

Germany. The lied, or song, continued its 15th-century role as the chief secular form in Germanic areas, but it did not develop to the same extent as the madrigal and the chanson. Throughout the Renaissance it was relatively conservative in its adherence to the cantus firmus principle and its tendency toward chordal over contrapuntal texture. Following Heinrich Isaac in the 15th century, the major 16th-century lieder composers were Ludwig Senfl, Hans Leo Hassler, and Johann Hermann Schein. To all national schools of the 16th century must be added the name of the Flemish composer Orlando di Lasso, who wrote in French, Italian, or German, depending on his current employment. The Spanish villancico was a flourishing popular form, but there was no Iberian equivalent to the madrigal, the chanson, or the lied.

The tonal era and after: 1600 to the present

The beginning of the 17th century was one of the most dramatic turning points in the history of music, even more so than the beginning of the Ars Nova and almost as revolutionary as the beginning of the 20th century. The winds of change had been felt several decades earlier, and the establishment of the new style required several decades after the turn of the century, but the year 1600 saw the performance of several works destined to change the course of music.

THE BAROQUE ERA

Originally used in a derogatory sense of referring to something bizarre, degenerate, and abnormal, the term Baroque gradually acquired a positive connotation for the grandiose, dramatic, energetic spirit in art that prevailed during the period from about 1600 to about 1750. The new spirit required a vastly expanded musical vocabulary, and a rapid evolution of new techniques occurred, particularly in vocal music. Two distinct musical styles were recognized. One, the *prima prattica* (or *stile antico*), was the universal style of the 16th century, the culmination of two centuries of adherence to Flemish models. The other, called *seconda prattica,* or *stile moderno,* referred to the new theatrical style emanating from Italy. The expanded vocabulary allowed for a clearer distinction between sacred and secular music as well as between vocal and instrumental idioms, and national differences became more pronounced. The tonal organization of music evolved also, as the medieval modes that had previously served as the basis of melody and harmony were gradually replaced, during the 17th century, by the system of tonality dominating Western music until about 1900: a system based on contrasting keys, or sets of interrelated notes and chords deriving from a major or minor scale. Viewing the period as a whole, two additional innovations most clearly distinguish it from the preceding Renaissance: concertato, or the contrast, combination, and alternation of voices and instruments, and basso

The spirit of the Baroque

continuo (thorough bass, figured bass), an accompaniment consisting of a low-pitched instrument, such as a violoncello or a bassoon, combined with a keyboard instrument or lute capable of harmonic elaboration.

Opera. Most typical of the emerging style were the dramatic productions of the Camerata, a group in Florence who were dedicated to recreating and imitating the musical ideals and practices of classical antiquity —in a sense, the musical manifestation of the Renaissance. Their guiding philosophy was the pre-eminence of textual over musical considerations; their belief was that the function of music was to heighten the dramatic impact of words. The musical result was monody: originally recitative (solo singing reflecting speech rhythms), later also arioso (more lyric than recitative) and aria (more elaborate song), accompanied by a basso continuo that could provide an innocuous background to a solo voice. Among the major figures in this revolutionary movement were Giulio Caccini and Jacopo Peri, both of whom composed operas based on the legend of Orpheus and Eurydice. Caccini also provided the name for the new movement with his publication of *Le nuove musiche*, a collection of solo songs with continuo accompaniment. The ideas and techniques conceived by the Camerata spread rapidly over Italy and, subsequently, all over Europe.

Early Italian operatic schools. During the 1620s and 1630s the centre of operatic activity shifted from Florence to Rome, where several distinctive features developed: a chorus was used extensively, dancing was incorporated into the dramatic spectacle, and an overture in the style of a canzona became the accepted norm. A flourishing operatic activity developed a decade later in Venice, where the first public opera house was opened in 1637. Public taste began to influence operatic composition, and, as a result, several innovations, such as the extensive use of popular tunes, spectacular staging, and short, fanfare-like overtures, were introduced. The audience's desire for tuneful songs also contributed to the clear distinction between recitative and aria, which began with the Venetian school. Foremost among contemporary composers was Monteverdi, who had known of the activities of the Florentine Camerata while serving as musical director to the Gonzaga family in nearby Mantua. He adopted the new style for his later madrigals and wrote two operas, *Orfeo* (1607) and *L'Arianna* (1608), before moving to Venice in 1613. Francesco Cavalli and Antonio Cesti became the leading Venetian operatic composers after Monteverdi's death in 1643.

17th-century developments in Rome

Neapolitan opera. The last major operatic centre to develop in Italy began its activities in the 1670s in Naples. Neapolitan opera seria, or serious opera, with characters from classical history or mythology, dominated Europe for a century. It was essentially a series of recitatives and arias, the latter mostly of the da capo type (ABA, the A section given improvised embellishment on its repetition) characterized by florid virtuosic singing. Other features were, first, the distinction between recitativo secco (dry recitative), accompanied by the continuo, and recitativo accompagnato, or stromentato, accompanied by the orchestra, and, second, the establishment of the Italian overture. Called a sinfonia, the overture in three parts (fast–slow–fast) evolved into the symphony during the 18th century. Alessandro Scarlatti was the most influential of the early Neapolitan operatic composers.

France. During the same period, opera was introduced at courtly functions outside Italy. After Luigi Rossi's *Orfeo* was performed in Paris in 1647, the Italian form was gradually merged with the major French dramatic form, the ballet; the importance of dancing in French operas thereafter is not surprising. Another distinguishing feature was the French overture (a slow movement, a fast movement, and, occasionally, a return to the opening slow section), which, like the Italian overture, later had an independent life. The masters of French opera during the Baroque period were Jean-Baptiste Lully and his successor Jean-Philippe Rameau. Because of the social and political upheaval of the Thirty Years' War, there was less operatic activity in Germany than in France, and the activity that did occur was more completely dominated by the Italian style. Hamburg, Munich, Dresden, and Vienna were the major centres, with Reinhard Keiser and Georg Philipp Telemann as the most prolific composers.

The joining of the French and Italian forms

England. The situation in England resembled that in France, since the English also had a flourishing musico-dramatic form, the masque, which gradually merged with Italian opera. Henry Purcell and John Blow were the chief composers of opera in English before Italian domination of serious opera became almost complete during the 18th century.

Cantata and oratorio. The leading Neapolitan opera composers also helped to establish the Baroque successor to the madrigal—the cantata—which originated as a secular form for solo voice with instrumental accompaniment. Giacomo Carissimi standardized the form as a short drama in verse consisting of two or more arias with their preceding recitatives. The cantata was introduced into France by one of Carissimi's students, Marc-Antoine Charpentier; Louis Nicolas Clérambault continued the tradition in the late Baroque period. With the fading stylistic distinction between sacred and secular music, the cantata was quickly converted to church purposes, particularly in Germany, where it became the chief decorative service music for the Lutheran Church. Dietrich Buxtehude and Johann Kuhnau were two of the leading composers of such church cantatas.

While the new concertato techniques were being applied to established forms of church music, such as the mass, service, motet, anthem, and chorale, new forms emerged that were clear departures from Renaissance styles and types. The oratorio and settings of the Passion story developed simultaneously with opera and on almost identical lines, consisting of recitatives, arias, vocal ensembles, instrumental interludes, and choruses. Emilio de' Cavalieri was the "founder" of the oratorio with his *La rappresentazione di anima e di corpo* (*The Representation of the Soul and the Body*). Produced in Rome in 1600, this work, unlike true oratorio, used actors and costumes. Carissimi and Alessandro Scarlatti were the chief Italian Baroque composers of oratorio, and Heinrich Schütz, a pupil of both Giovanni Gabrieli and Monteverdi in Venice, was the leading 17th-century German composer in this field.

Instrumental music. The new techniques of *Le nuove musiche* were to be heard in music for instruments, especially now that they participated in genres formerly written for unaccompanied voices (*e.g.*, the motet). The forms and mediums of instrumental music remained essentially the same but with considerably different emphasis. The lute, for example, lost status quickly with the rise of the harpsichord as the most common instrument for continuo accompaniment of dramatic productions. The organ, as the traditional church instrument, retained its position and assimilated the evolving forms.

Continuo instruments

Modification and expansion of older forms. Dance pairs of the Renaissance grew, about the middle of the 17th century, into dance suites consisting basically of four dances: allemande, courante, saraband, and gigue, with optional dances such as the gavotte, the bourrée, and the minuet sometimes inserted before the final movement. Variation forms—the chaconne (in which a set of harmonies or a bass theme is continuously repeated), the passacaglia (in which the theme is repeated but not necessarily in the bass), along with the ground bass and variations on well-known melodies—continued to be popular. Free forms also continued in the patterns of their Renaissance antecedents, while growing in dimension and inventiveness. The toccata, prelude, and fantasia were expanded into multisectional forms using the three basic instrumental textures—imitative counterpoint, chordal homophony, and virtuosic passage work—in combination, alternation, and contrast. The Renaissance fugal forms, chiefly the canzona and the ricercar, gradually evolved into the late Baroque fugue, and cantus firmus compositions continued to flourish as a result of their liturgical function.

The sonata and concerto. The major new categories of instrumental music during the Baroque period were the sonata and the concerto. Originally applied to instrumen-

tal ensemble pieces derived from the canzona, the term sonata became the designation for a form that was to dominate instrumental music from the mid-18th until the 20th century. In its keyboard manifestation, it was a binary (two-part) structure similar to a dance-suite movement. For small ensemble, it evolved into a series of independent movements (usually in a slow–fast–slow–fast arrangement) called a *sonata da chiesa* ("church sonata") or a dance suite called a *sonata da camera* ("chamber sonata"). Especially prominent was the trio sonata, for two violins (or flutes or oboes) and cello with continuo. Eventually, similar forms were adopted for orchestra (sinfonia or concerto), for orchestra with a small group of featured instruments (concerto grosso), or for a solo instrument with orchestra (solo concerto). The fundamental principle of the concerto was that of contrast of instrumental groups and musical textures.

Throughout the period, keyboard music flourished, notably in the hands of Jan Pieterszoon Sweelinck in the Netherlands, Johann Pachelbel and Johann Froberger in Germany, Girolamo Frescobaldi in Italy, and Domenico Scarlatti in Spain; in France the chief exponents included Rameau and François Couperin.

Instrumental ensemble music, both chamber and orchestral, was dominated by Italians, chiefly from Bologna, the Bolognese school producing such composers as Arcangelo Corelli, Antonio Vivaldi, and Giuseppe Tartini. Purcell in England and Couperin and Jean-Marie Leclair in France are representative of the many composers in other nations who were influenced by Italian models of instrumental ensemble music.

The late Baroque. The Baroque era reached its zenith in the work of Johann Sebastian Bach (1685–1750) and George Frideric Handel (1685–1759). Both were born in the same part of Germany; both were reared in the Lutheran Church; and both were primarily organists; but because of different environmental circumstances each became a master of different musical forms. Handel, because of his conditioning in Italy, was primarily a dramatic composer, writing opera, oratorio, and secular cantatas, mostly after he reached England. He also wrote quite extensively for orchestra and instrumental ensemble. Bach, by contrast, was influenced by his lifelong employment in the church and by his dedication as a teacher; his works thus include Passions, cantatas for church services, liturgical organ pieces, and harpsichord compositions, many instructional in purpose.

The primacy of Bach and Handel

In the works of both Handel and Bach changes in technique reached a culmination with the clear establishment of the tonal system, allowing for modulation from one key centre to another, primarily as a device for formal organization. Rich, chromatic harmonic language was both reason and result of such a change. The fusion of contrapuntal technique with homophonic style resulted in a distinctive hybrid texture that employed figured bass (homophony) as a foundation for two or more independent melodic lines (polyphony).

THE CLASSICAL PERIOD

The term classical

As in the case of the Renaissance, difficulties with terminology again arise with the label classical. Does it refer to a period of time, a distinctive musical style, an aesthetic attitude, an ideal standard, or an established norm? Again, the term was borrowed from the visual arts of the same epoch and is awkward when applied to music in that there were no known models from classical antiquity for composers to imitate. A full understanding of the term depends on a clear conception of the term romantic, for the two stand at opposite poles. Each represents a set of artistic ideals that has been in opposition to the other since both were recognized by early Grecian writers. As has been noted, the ancient Greek followers of Apollo established the ideal of classicism, whereas the cult of Dionysus produced the prototype of romanticism. A mixture of the two qualities has prevailed throughout recorded history, with first one and then the other in the ascendancy. Thus, there have been many "classic" and many "romantic" eras, but the labels have come to refer most specifically to the last half of the 18th century and the 19th century, respectively, because those periods represent most vividly the two tempers.

The social and political scene during the late 18th century was hardly a setting for a quiet, composed "classical" age in view of the prevailing revolutionary spirit and colonial rivalry. The revolutionary movement did have a direct effect on music in that "music for the masses" became a new ideal—music directly appealing to a large number of unsophisticated people who had previously been excluded from courtly entertainments.

Precursors of the Classical style. *The Rococo style galant.* As the pendulum swung from the predominantly romantic Baroque period toward the Classical period, there was an inevitable overlapping of the old and the new. While Bach was composing his intricate and erudite polyphony, his sons were reflecting a new ideal, the Rococo. Fostered by the court of the French king Louis XV, whose life-style was far less formal than that of his illustrious great grandfather, the Rococo ideal was artistic expression dedicated to elegance, frivolity, and gracefulness; a work of art must be delicate, playful, entertaining, and immediately appealing. The result was often artificial and unrealistic, but it succeeded in capturing the discreetly sentimental and hedonistic attitudes of the times. Powdered wigs, lace cuffs, and perfumed handkerchiefs for both sexes were other manifestations of the same playful spirit that produced music in the *style galant.*

The empfindsamer Stil. The German counterpart of the essentially French Rococo was the *empfindsamer Stil,* or "sentimental style," which flourished in the 1750s and 1760s. Its leading exponent was one of J.S. Bach's sons, Carl Philip Emanuel Bach, who served for a time at the court of Frederick the Great in Berlin. The distinguishing feature of this German reaction against Baroque profundity was its concern with emotional feeling in the music itself, on the part of the performers and, hopefully, in the reaction of the audience. The French obsession with lightness, gracefulness, and decoration was countered by the German determination to affect sensibilities that were often more attuned to tears than to laughter. A late and less reserved manifestation of *Empfindsamkeit* was the *Sturm und Drang* ("storm and stress") movement in the arts during the 1770s and 1780s. The inclination toward the more intense personal expression of that movement was a harbinger of the coming Romantic period.

Consolidation of the Classical style. The fundamental changes in musical style that distinguished Classical from Baroque were inspired by Rococo ideals and refined and stabilized by the Classicists, particularly Joseph Haydn, Wolfgang Amadeus Mozart, Christoph Willibald Gluck, and the young Ludwig van Beethoven.

Stylistic elements. For the first time in the history of music, instrumental music became more important than vocal music. The orchestra and chamber groups, such as the string quartet, trio, and quintet, and the piano trio became standardized and replaced the heterogeneous trio sonata and other ensembles of the Baroque period. The basic duple and triple organization of metre remained unchanged, but rhythmic patterns tending toward more regularity and simplicity became the rule, producing the "tyranny of the bar line" that was to prevail for more than a century.

Central role of instrumental music

Melody was inclined to be more motivistic, tuneful, and epigrammatic, in contrast to the extended, figurative style of many Baroque melodies. Harmony was second only to melody as a focal element. Harmonic patterns that clearly established the tonal centre were the rule of the day.

As a reaction against the intricate polyphony of the later Baroque period, homophonic texture dominated by melody became the norm, but the accompanying patterns were different from those of the early Baroque, when monody supported by sustained chords was the prevailing style. In the late 18th century, figurations such as the Alberti bass (form of accompanying figure consisting of broken chords) and rhythmically enlivened repeated chords formed the typical textural patterns. Counterpoint was retained in some forms, however, and regained status particularly in development sections of works in sonata form.

Formal structure, a definitive aspect of classical style, was characterized by simplicity and clarity. Sectional forms (created by contrast and repetition of thematic materials, tonalities, and textures), variations, and the new principle of development (fragmentation, expansion, and modification of themes) were the established norms. Phrases of musical material became shorter and more clearly demarcated as well as more balanced and regular. A new concept of dynamic contrast also contributed to formal clarity. Shading from loud to soft or vice versa provided a dramatic means of building toward an expressive climax. Orchestration and instrumentation were closely allied to dynamic variation, and much more colour contrast and variety appeared in orchestral music, even though the ensemble was more standardized than formerly.

Sonata form. The pattern that served as the structural basis for most instrumental music of the classical period was the sonata. A large-scale work in several movements, it evolved from several Baroque predecessors, chiefly the Italian overture, the *sonata da chiesa,* and the concerto grosso. Depending on the medium of performance for which it was intended, it would be called, for example, a symphony, a concerto, a string quartet, a sextet, a trio. The designation sonata was reserved for a solo instrument or for an instrument accompanied by harpsichord or piano. Originally in three movements, the sonata became standardized as a four-movement form when a minuet was incorporated in the following sequence: (1) a serious allegro, (2) a slow, lyrical movement (andante or adagio), (3) a minuet and trio, and (4) a brilliant, vivacious finale. The internal structure of the first movement was so uniform that it acquired the designation sonata-allegro form; that is, the form employed in the allegro movement of a sonata, consisting basically of exposition, development, and recapitulation. The slower second movement is less structurally predictable. It is frequently a sectional form (for example, ABA, AABA, ABCA) or a set of variations. It may, even though in a slow tempo, be a sonata-allegro form, illustrating again the inconsistency of musical terminology. The third movement, usually omitted in the concerto and sometimes in other forms, is either a graceful minuet or a scherzo, a lively rhythmic form derived from the minuet. The structural pattern of the minuet had been fixed when it was established as the official court dance by Louis XIV in the mid-17th century. The last movement is frequently a rondo form, in which the principal theme recurs regularly between subordinate themes.

Instrumental music. *The symphony.* The most important and influential manifestation of the sonata form was that played by an orchestra—the symphony. During the 17th century the term sinfonia had been used for various kinds of instrumental music. "Sonata" was equally ambiguous. Late in the century, the designation sinfonia began to be confined to the Italian opera overture—a three-movement arrangement, fast–slow–fast. By the mid-18th century, opera overtures were being played independently in concerts. The insertion of the minuet between the last two movements resulted in the prototype of the Classical symphony.

During the waning Baroque period, vigorous advocates of the burgeoning Rococo and *Empfindsamkeit* ideals were active in Milan, Vienna, and Mannheim. In Milan, Giovanni Battista Sammartini began writing his symphonies, some 25 of them, in the 1730s. While employing the continuo of his models, Corelli and Giuseppe Torelli, the bithematic plan for his opening movements foreshadowed the exposition of the Classical symphonies. At about the same time, young composers in Vienna were experimenting with the new genre, thus laying the foundation for the later Viennese masters. The most famous and probably the most influential group was active in Mannheim in the court orchestra of Karl Theodor, the elector of the Palatinate. Their activity began in the 1740s, when Johann Stamitz became leader of the orchestra. His experiments with dynamic techniques—crescendo (increasing in loudness), diminuendo (decreasing in loudness), sforzando (special emphasis)—with homophonic textures featuring the first violins in virtuoso passages and with tremolo and other dramatic effects, became the hallmarks of the Mannheim style and served as models for his son Karl Stamitz and for composers in Vienna. Thanks to the fortuitous presence of certain instrumentalists as well as to benevolent patronage, the basic ensemble of the modern symphony orchestra was gradually established: violins, violas, violoncellos, and double basses; two flutes, two oboes, and two bassoons; two French horns and two tympani. Trumpets were added for festive occasions.

The Mannheim school

The concerto. Unlike the symphony, which had its origins in other forms, the Classical concerto grew directly out of the Baroque solo concerto and resembles it in that it is based on exchange of musical material between solo instrument or instruments and orchestra. While directly derived from the ritornello principle of the Baroque concerto (that of a recurrent musical passage when the soloists are silent), the internal structure of the first movement assimilated the developmental principle of sonata-allegro form. Pietro Locatelli and Giuseppe Tartini are especially notable for their numerous late-Baroque violin concerti.

Chamber music. While music for small instrumental ensembles had flourished for over 200 years previously, the late 18th century witnessed the establishment of chamber music in the modern sense of the term: music in sonata form for a small group of instruments with one player for each part. Replacing the trio sonata of the Baroque period, the most popular classical ensemble was a group of four stringed instruments—two violins, a viola, and a violoncello. Both the group itself and a sonata written for the group were called a string quartet. Among other popular ensembles were the string trio (violin, viola, and violoncello; or two violins and a violoncello) and the piano trio (violin, violoncello, and piano).

The keyboard sonata. The solo keyboard sonata was one of the most vital forms of the period, partly because of the great increase in amateur performers resulting from the newly affluent middle class. The sonatas of Domenico Paradisi, of J.S. Bach's sons, and of Haydn and Mozart reflect the evolution from the one-movement, binary form of the Baroque period to the standard classical three-movement form. A four-movement form did not become popular until the time of Beethoven. A celebrated contemporary of Mozart, Muzio Clementi, composed more than 60 sonatas for the piano alone and half again as many for piano and violin or flute and strongly influenced the style of piano writing.

Clementi's 60 sonatas

Other instrumental forms. While the sonata was unquestionably the most important form of instrumental music during the period, several other types were cultivated. For orchestra and chamber ensemble, a suitelike work called variously divertimento, serenade, cassation, or notturno was popular for light entertainment, differing from the more serious symphonies, concerti, and sonatas (which were intended for attentive listening) in that the ensemble of instruments was inconsistent, unpredictable, and often unspecified. The number, types, and arrangements of movements were equally flexible, ranging from three to ten or more, some in dance forms and others in forms suitable for a sonata. While nonsonata forms for solo instruments (particularly keyboard) occasionally bear these designations, the most popular smaller solo forms were sets of variations, individual dances or marches, fantasies, and small pieces that would have been appropriate as movements of sonatas. For some reason, composition for the organ dwindled drastically after the death of J.S. Bach, in 1750.

Vocal music. *Opera.* There was less distinction between Baroque and Classical opera than between instrumental styles of the two periods because opera, with musical interest centred on a solo voice, had been largely melodic–homophonic since its inception. Another reason for the continuity of operatic style throughout the 18th century was the universal domination by the Neapolitan opera seria. Even in Paris, where the Lully–Rameau tradition maintained its vitality, there was an Italian opera theatre. While there was some effective reform of certain

aspects of Neapolitan style that had become decadent and some nationalistic reaction in the field of comic opera, nothing in the nature of serious opera challenged Neapolitan supremacy. As a result, the late 18th century was a period of great vitality in operatic composition.

Domination of opera seria

The distinguishing characteristics of Neapolitan opera seria reveal why it is little known and rarely heard today. It was a very conventionalized form, with artificial and overcomplex plots. There were usually six main characters representing three of each sex, with some of the male and female parts sung by castrati (emasculated male sopranos and contraltos). Each character was allotted a standardized number of arias in fairly standardized succession. Obviously, with such constant interruption of the action, dramatic truth received little if any consideration. The singers and the arias were the focus of the entire production, with little of musical interest in the parlando recitatives (*i.e.*, using speech rhythms), little use of chorus, and little function for the orchestra aside from providing a subordinate accompaniment.

Objections to the decadence and artificiality of the Neapolitan style, which had begun to appear as early as the 1720s, would have been fruitless had not a champion appeared to put suggestions and theories for reform into actual practice. Culminating the movement for reform was Christoph Willibald Gluck, who began his career in the 1740s by writing about 20 operas in the prevailing style. Then, beginning with *Orfeo ed Euridice* in 1762, he attempted to enhance both the dramatic and musical components of opera. Superfluous virtuosity and vocal display were drastically curtailed if not eliminated by providing music that reflected the emotional or dramatic situation. As a result of Gluck's reforms, opera moved toward a classical simplicity of style of which his and Mozart's works were the culmination.

Comic opera

A second challenge to established Neapolitan opera was emerging through comic opera in which the subject matter was light, sentimental, often topical, and satirical, reflecting both the social changes of the period and ridicule of serious opera. The music was engagingly tuneful, easy to perform and to comprehend. Comic opera had appeared during the 17th century but began its independent existence during the first half of the 18th century in Italy, where it was called the opera buffa. The French opéra comique evolved during the same period and was given new impetus by the *guerre des bouffons* ("war of the buffoons") of the early 1750s, when support of the Italian opera buffa company then performing in Paris exceeded that of the French heroic opera of Rameau. In England, ballad opera, beginning with *The Beggar's Opera* in 1728, followed a course of development similar in both period and style to that of the opéra comique. German singspiel grew out of translations and imitations of English ballad opera. Models that centred on Vienna adhered to the Italian style and culminated in Mozart's *The Abduction from the Seraglio* and *The Magic Flute*. Yet Mozart also brought the old Italian style to its zenith in *Le Nozze di Figaro (The Marriage of Figaro)*, *Così fan tutte (Thus Do They All)*, and *Don Giovanni*.

Other vocal music. Aside from opera, secular vocal music was composed for solo voice and chorus. But the production of solo songs and cantatas in other countries could not compare with the growing interest in the German lied, which flourished under C.P.E. Bach and later composers. The most extensive development of secular part-songs took place in England, where numerous catches and glees were written.

Large-scale sacred choral music of the period was strongly influenced by the prevailing operatic style. Except for the text, some passages from oratorios and passions are indistinguishable from an operatic excerpt. But the Handelian tradition combined with the Neapolitan style and culminated in Haydn's two noble oratorios, *The Creation* and *The Seasons*. Liturgical music, such as masses, motets, litanies, psalms, and canticles, also demonstrated that the same composers were writing for both church and theatre. In many instances the style was uniform for the two types, although the chorus naturally played a much greater role in church music.

THE ROMANTIC PERIOD

The beginning of the 19th century witnessed a change of both musical style and aesthetic attitude that has become identified as Romantic. The term romantic originated in German literature of the late 18th century, illustrating once again the overlapping of classical and romantic attitudes and ideals. The Franco-Swiss writer Mme de Staël articulated the new ideals of the movement in 1813 as original, modern, national, popular, derived from the soil, religion, and prevailing social institutions. Obviously, some of these proclaimed romantic ideals and purposes were the same as those of the 18th-century Classicists.

Establishment of the Romantic idiom. In defining classicism, it was suggested that the distinctive elements of musical romanticism embrace emotionalism, subjectivity, individualism, nationalism, and a preference for a certain type of subject matter. Emotionalism is reflected in the revelation of intensely personal feelings, indeed, the sentimentality that pervades 19th-century music. Subjectivity replaces the formalism of the Classical period, in that impulse and inspiration play a major role in the motivation of both composer and performer, and the listener's response is expected to be more sensory than intellectual. Closely related to subjectivity is the individualism reflected in the highly self-centred expression of composers of the period, as well as by the conviction that a composer's most personal thoughts and feelings are the ultimate artistic message. In contrast to the universality of musical style that prevailed during the 18th century, much 19th-century music is identifiable in terms of national origin. Nationalism—the consciousness of the distinctive features of a nation and the intent to reveal, emphasize, and glorify those features—played a prominent part in Romantic music, partly as a result of social and political developments. The subject matter favoured by Romantic composers is most apparent in vocal music, where words can convey the explicit theme, but instrumental music was also affected by the Romantic attraction to national identification and to remoteness, strangeness, and fantasy, particularly to the fantastic aspects of medieval tales and legends.

Personal aspect of romanticism

Stylistic traits. During the 19th century, musical techniques and materials were rapidly enriched by new resources, all devoted to the ideal of emotional or dramatic expressiveness. The orchestra, the piano, the solo voice with piano accompaniment, and the opera were the four predominating mediums; chamber and choral music occupied a less central position. Duple and triple divisions of the measure remained the basis of metre, but there were occasional experiments with metric irregularity, and rhythm was recognized as one of the most effective agents of expressiveness in music. Strong rhythmic energy, frequently produced by dotted patterns, provided a vigorous force that could be enhanced by faster and faster tempi. Flaccid patterns in slower tempi provided for the requisite pensive or sorrowful moods.

Melodic style was determined on one hand by the vocal ideal of song, with long, lyric lines. On the other hand, the new idiomatic possibilities of instruments were being exploited. In either case, expressiveness was the governing ideal. Harmonic and tonal elements were gradually expanded during the century, with more chromaticism, enriched sonorities (seventh and ninth chords), and more nonharmonic tones resulting in a more flexible tonal scheme. Tonic and dominant chords (those based on the first and fifth notes of the key), were no longer the secure poles of tonal movement; frequent and remote modulations (changes from one key to another) contributed to the restlessness of key centres. Because musical interest was centred in melody and harmony, texture remained prevailingly homophonic, though counterpoint played a prominent role in developmental sections. Fugues and other imitative forms occurred as a result of studied archaism or for special effect, but the composers' preoccupation with direct and immediate expression led them to neglect the traditional polyphonic forms, with their inherent traditions and restraints.

Increasing importance of modulation

The Romantic abhorrence of formalism has frequently been exaggerated for the purposes of distinguishing be-

tween Classical and Romantic attitudes. Established patterns such as the sonata-allegro and rondo forms were subjected to many modifications and extensions, but musical coherence demands a judicious balance of unity and variety, so most compositions of the 19th century are still fairly clear-cut sectional, variational, or developmental forms. The distinctive new features are largely in the area of emphasis and dimension—symphonies lasting over an hour in contrast to the 20- to 30-minute standard of the Classical period. Instrumental colour and variety, as another aspect of expressiveness, was made possible by a greatly enlarged orchestra and figured prominently in the new sound ideal.

Beethoven as a transitional figure. The Janus-like figure who marked the transition from the Classical to the Romantic style was Beethoven, the first composer whose personality and character made a purposeful impact on the types and style of music he composed. Inspired by the revolutionary forces prevailing at the time, he declared himself a free artistic agent, with neither allegiance nor responsibility to any patron. His early works reflected the 18th-century acceptance of providing music on demand, and he applied his craftsmanship to supplying compositions in hope of financial reward. But in his later works, from about 1820 on, he declared his personal independence and wrote only what his imagination and inspiration dictated, thus establishing individuality, subjectivity, and emotional expression as the standard for Romantic composers. Yet the body of music he produced reflects the tastes of the 18th rather than the 19th century, in that he was attracted more by the absolute forms of instrumental music than by the dramatic and lyrical forms cultivated by the Romanticists. Symphonies, chamber music (particularly string quartets), and piano pieces (including 32 sonatas) far outweigh his one opera, one oratorio, one major mass, and assorted songs and part-songs. His lack of interest in dramatic vocal music reflects the classical side of his nature, though the expressive changes apparent in his instrumental works are evidence of his being the springboard to the Romantic epoch.

Opera. The opera remained a flourishing medium throughout the 19th century, and Italian opera continued as the dominant type during the first half of the century in the hands of Gioacchino Rossini, Vincenzo Bellini, and Gaetano Donizetti. The reforms instigated by Gluck were discernible, but enough of the genre's indigenous Italianate character remained to distinguish it from other national types. The man who, more than any other, personifies Italian opera of the period is Giuseppe Verdi (1813–1901), whose works are still among the most performed. Late in the century, the tendency toward even more realistic and topical subject matter produced the *verismo* ("realism") school of Ruggiero Leoncavallo, Pietro Mascagni, and Giacomo Puccini.

Meanwhile, German opera developed into the epitome of romantic subject matter and expressiveness, beginning in 1821 with the performance of *Der Freischütz* (*The Freeshooter*), by Carl Maria von Weber. Plots based on tales from Teutonic mythology and medieval legend that emphasized the mystical aspects of nature were a distinctive feature of Germanic operas and distinguished them from the more mundane Italianate plots. Richard Wagner (1813–83) crystallized the German Romantic ideal into the music drama, in which all aspects of the production—drama, music, design, performance—were intended to fuse into a manifestation of pure artistic expression in which no one element predominated over the others, as singing still tended to do in Italian opera. There was no development in Germany after Wagner comparable with the post-Verdi *verismo* group in Italy. Wagner's innovations—once again a readjustment of dramatic versus musical forces in musical theatre—were the point of departure for most German opera since his time, from Richard Strauss to the present.

Wagner's music drama

In Paris, the operatic centre of the world from late in the 18th century until well into the 19th, native composers were quick to sense the Gluckian changes in Italian opera as well as the new directions in Germany. Beginning with Rossini's *Guillaume Tell* in 1829 and crystallizing in the operas of Giacomo Meyerbeer in the 1830s and 1840s, French grand opera emerged as the most opulent and grandiose musico-dramatic spectacle of the first half of the century. During the later 19th century, opéra comique and grand opera merged to produce the prevailing French lyric opera. At the same time, opéra comique branched off in another direction to produce operettas, which developed into the musical comedies of the 20th century. Indigenous opera appeared in other regions, especially in Russia, Bohemia, and Scandinavia, as a result of nationalistic fervour.

Orchestral music. Reaching both a culmination and a turning point in the nine symphonies of Beethoven, orchestral music developed in two directions during the 19th century. On the one hand were composers who, because of their training and temperament, adhered primarily to Classical forms and ideals of absolute music. On the other hand were the composers seeking new realms of dramatic content, colour, and expressiveness. Even for the more conservative group, both the forms and the orchestra itself were greatly expanded during the century, but the total output of works was much smaller than in the Classical period. Romantic musical vocabulary replaced the Classical language in symphonies, of course, and programmatic content (*i.e.*, an extramusical image or story) was a frequent element.

New orchestral forms. The more progressive composers cultivated new musical types that represented the tastes and ideals of the Romantic period—the concert overture, the symphonic poem (later called tone poem), the symphonic suite, and symphonic variations. The concert overture, a direct development of overtures to dramatic works, was an attempt to reconcile the old classical demands for form with Romantic desire for programmatic content. It was usually a sonata-allegro form with picturesque themes designed to suggest (with the help of the title) characters, locations, or dramatic situations. Felix Mendelssohn's *Hebrides* overture and Brahms's *Tragic Overture* (completed 1880) are representatives of the genre. The symphonic poem, foreshadowed in Hector Berlioz' *Symphonie fantastique* (completed 1830), was originated by Franz Liszt at midcentury as an orchestral work, usually in one movement, based on an extramusical idea such as a poem or a narrative. The futility of attempting to depict explicit events and attitudes in purely musical terms resulted in the demise of the form early in the 20th century after the many tone poems of Richard Strauss. The symphonic suite was one of three distinct types: (1) an outgrowth of 18th-century dance suites, divertimentos, or serenades, (2) the extension of the symphonic poem into a composite work of several movements of related programmatic nature, or (3) a group of selections from a dramatic work such as a ballet. Rimsky-Korsakov's *Scheherazade* represents the second type, and Tchaikovsky's *Nutcracker Suite* is typical of the third. While variations had appeared as movements of symphonies and concerti since the mid-18th century, they became an independent orchestral form during the last quarter of the 19th century. César Franck's *Variations symphoniques* (1885) is a good example of the type. Popular orchestral pieces, such as the waltzes of Johann Strauss, also flourished.

Programmatic and picturesque works

The mainstream of composers whose orchestral music reflected most clearly their allegiance to classical forms and models—though conditioned by Romanticism, of course—is represented by Beethoven, Franz Schubert, Robert Schumann, Johannes Brahms, and Anton Bruckner. The more overtly Romantic contingent was centred around Berlioz, Liszt, Gustav Mahler, and Richard Strauss.

The movement toward national expression. A third group, chiefly nationalists who were reacting against Germanic domination of instrumental music as well as reflecting the sociopolitical developments of the era, combined features of both conservative and progressive camps, to which they added national characteristics. While there were manifestations of the movement in countries such as Hungary, Poland, Spain, and England, the most productive and outstanding of those who sought

to reflect national distinctiveness were the Russian "Five" —César Cui, Mily Balakirev, Aleksandr Borodin, Modest Mussorgsky, and Nikolay Rimsky-Korsakov—the Bohemians Bedřich Smetana and Antonín Dvořák, and the Scandinavians Edvard Grieg, Carl Nielsen, and Jean Sibelius. French composers such as Camille Saint-Saëns, César Franck, and Vincent d'Indy were motivated by the same impulse of independence, but they could hardly be categorized as nationalists in the same sense as their eastern European colleagues, since there was less distinctive folk material from which to draw colourful materials and since they had been in the mainstream of musical development all along.

Piano music. One of the most popular media of the Romantic era, thanks to the rapid technical development of the instrument, was piano music. Another reason for the popularity of the piano was the growing demand for recreation and entertainment on the part of the newly affluent middle class. In tune with the taste of the times, small pieces of distinctive expressive character (hence, "character pieces") were the most popular type, either as single pieces or as parts of composite works. Stylized dances continued to be popular, but nationalistic types such as the polonaise and mazurka and the novel waltz replaced the staid minuets of the previous era. Sonatas continued to occupy serious composers, and sets of variations continued to flourish. The virtuosity of the violinist Niccolò Paganini and his contemporaries led to many studies, or études, designed to exhibit the performer's dexterity as well as the invention of the composer. Chopin, Schumann, Liszt, and Brahms were the major composers of piano music after Beethoven, but practically all composers of the time contributed to the literature.

Songs. The vocal counterpart of the keyboard character piece was the solo song with piano accompaniment. With the rise of the German romantic poetry of Goethe, Schiller, Heine, and others, about the beginning of the 19th century, the German lied ("song") flourished. After 1850, composers of other nations, especially France and Russia, also produced a song literature of universal appeal. A pioneer and certainly the most prolific composer of lieder was Schubert, who in his short life wrote more than 600 songs. His chief successors, in chronological order, were Karl Loewe, Mendelssohn, Schumann, Brahms, Hugo Wolf, and Richard Strauss.

Schubert's 600 lieder

Chamber and choral music. The great Viennese tradition of chamber music reached its zenith in the works of Beethoven and with the death of Schubert came temporarily to a close. The conciseness, unity, and balance that were basic to the Classic ideal were incompatible with the essence of musical Romanticism. When writing for instruments, the typical Romantic composer was inclined toward the colouristic effects and expressive possibilities of the orchestra. Chamber music continued to be written and performed, of course, but nowhere was it one of the primary interests of composers as it had been during the 18th century. Predictably, the more conservative composers, such as Mendelssohn, Schumann, and Brahms, were the chief composers of chamber music.

While the same musical vocabulary and style had served both church and opera house since the rise of dramatic music, the 19th century witnessed a separation of musical idioms according to function—sacred or secular. Music for use in church was generally conservative, especially after the "rediscovery" of Palestrina and systematic research into the reform of Gregorian chant. On the other hand, cantatas and popular part-songs produced for the many amateur choral societies incorporated as many of the new techniques as could be managed by the singers. There was some fusion of the two idioms in oratorios and in settings of liturgical texts for the concert hall or for special occasions. The requiem mass, with its vividly dramatic content, was attractive to Romantic composers, and Berlioz's and Verdi's settings remain as emotionally telling today as most operas of the period.

MODERN PERIOD

Diversity of styles. The striking changes in musical style that occurred about 1900 were a turning point in the history of Western music comparable to the dramatic transformation of the early 14th and early 17th centuries. But never before had the change been so rapid, and never before had there been such a diversity of resulting styles. The last decades of the 19th century witnessed what might be termed the diffusion of Romanticism, when significant departures from the current musical vocabulary appeared in the works of some nationalist composers and especially in the Impressionistic style represented in France by Claude Debussy and Maurice Ravel. The amorphous rhythmic patterns, the whole-tone scale, the concept of free relationship of adjacent harmonies, and the kaleidoscopic textures of musical Impressionism were a musical manifestation of the aesthetic movement current in painting and literature.

The experimental works of Arnold Schoenberg and Igor Stravinsky about 1910 heralded a new epoch in music. Schoenberg was the pioneer when his adoption of the ideals of the Expressionist movement—like Impressionism an aesthetic development shared by other art forms —resulted in his discarding traditional concepts of consonance and dissonance as well as tonality and led into the development of atonality and 12-tone technique (in which all 12 tones of the octave are serialized, or given an ordered relationship). Stravinsky's revolutionary style, variously labelled "dynamism," "barbarism," or "primitivism," concentrated on metric imbalance and percussive dissonance and introduced a decade of extreme experimentation that coincided with World War I, a period of major social and political upheaval.

Stravinsky's innovations and experiments in Neoclassicism

In contrast with Schoenberg's and Stravinsky's experiments during the second decade of the century, another line of demarcation appeared about 1920 with a general return to the aesthetic ideals of the late 18th century. Under the leadership of Stravinsky, Paul Hindemith, Béla Bartók, and Sergey Prokofiev, among others, most prominent composers entered a Neoclassical period characterized by restraint of emotional content; simplification of materials, structures, and textures; a greater attention to craftsmanship; and a revival of concern for linear counterpoint rather than instrumental or harmonic colour. Baroque emphasis on counterpoint and Classical formalism were now clothed in 20th-century melodic, harmonic, rhythmic, tonal, and orchestral idioms. The Expressionist followers of Schoenberg, most notably Alban Berg and Anton von Webern, remained aloof from the Neoclassical movement and continued in their preoccupation with serial techniques.

Neoclassicism continued as the dominant trend throughout the period from about 1920 until World War II, while the experimental techniques introduced during the revolutionary second decade of the century were gradually refined, modified, and assimilated into the accepted musical vocabulary. At the same time, experimentation continued alongside a tenacious conservatism that echoed Romantic ideals and styles. Nationalism also continued to flourish, reaching a level in some countries never achieved during the 19th century.

After World War II the two leading artistic attitudes tended to merge when the followers of Webern carried serial composition to such a rigorous extreme that its craftsmanship and intellectual orientation suggested Classicism rather than Expressionism. Shortly afterward, Stravinsky, the doyen of the Neoclassical group, began experimenting with serialism. Avant-garde music since that time has begun to employ the techniques made possible by technological developments in electronics.

Advent of electronic composition. Beginning about 1950, two leading groups began experimenting with electronic music, one in Cologne and the other in Paris. The product of the latter group was referred to as *musique concrète* in acknowledgement of the principle that preexisting, or "concrete," recorded sounds serve as the basis of all sonorities in the finished work. The basic sounds, which may be derived from any source—musical, natural, or mechanical—are modified electronically and arranged in any combination and succession suitable to the composer's purpose. The German group, led by Karlheinz Stockhausen, was concerned with a purer form of

the medium in that its basic sounds are electrically generated instead of being recorded from sources external to the electronic apparatus. The two approaches share one connecting link with music of the past: all sounds have pitch, intensity, duration, and quality. All other concepts of musical organization have been discarded, including the necessity of a performer. Electronic compositions exist on a tape (or disk), and can be made audible by a speaker system. The dehumanizing of music has been carried several steps further by the use of mathematics and even of computers to determine the nature of sound materials—either electronic or that produced by more conventional means—and their organization. At the other extreme is aleatoric music, in which the performer is allowed to choose the manner and order of presentation of materials specified or simply suggested by the composer.

Computer music

Popular music and jazz. Another result of advances in electronics has been the tremendous growth of popular music during the 20th century. New techniques have made possible high-fidelity reproduction of sound and its widespread and rapid dissemination through radio, phonograph, tape recorder, and television. In addition, some of the instruments used in popular music have incorporated electronic amplification as well as sound production. While there has been a popular music as long as mankind has turned to singing and dancing for diversion and recreation, much of it was folk music and existed only as an oral tradition. But popular music in the modern sense originated in the late 18th century, when ballads made popular in ballad opera and dance music received wide circulation. The same types flourished throughout the 19th century and into the 20th, when a new direction was prompted by the emergence of jazz among Negroes in the southern United States. After the original jazz forms came several stylistic modifications such as ragtime, swing, bebop, and rock 'n' roll. Early in the century, the novelty of jazz rhythms and dominance of brass, woodwind, and percussion instruments over strings attracted some serious composers who occasionally incorporated suitable jazz idioms into their works. Since about 1930, the influence has worked in both directions, and popular music has gradually adopted techniques that originated in serious music. Regardless of the interaction of popular and serious music, the popularity of the former is one of the most significant musical developments of the 20th century, especially in view of the widening gulf between the serious composer and the potential audience.

Interaction of serious and popular idioms

It is impossible to arrive at a complete and objective description of a revolutionary movement while it is in progress; only a period of time can provide the necessary perspective. It can be acknowledged, however, that music has never before passed through a more anarchic phase than in the 20th century. The tremendous number and diversity of stylistic distinctions has precluded a characteristic designation for the first half of the century, but one must be forthcoming, for musicians of the future will need the terms modern and contemporary for their own times.

Despite the disproportionate publicity given by communications media to the most radical experiments, the majority of leading composers continue along the moderate path established in the late 1920s and 1930s. And, if one can rely on the lessons of history, the mainstream of music will continue to absorb those new techniques that contribute to expressiveness and communication while discarding the merely novel and sensational, so that music history will remain an evolutionary rather than a revolutionary process.

BIBLIOGRAPHY

Reference works: WILLI APEL (ed.), *Harvard Dictionary of Music,* 2nd rev. ed. (1969); THEODORE BAKER, *Biographical Dictionary of Musicians,* 5th ed., rev. by NICOLAS SLONIMSKY (1958; with 1965 supplement); ERIC BLOM (ed.), *Grove's Dictionary of Music and Musicians,* 5th ed., 9 vol. (1954; with 1961 supplement); FRIEDRICH BLUME (ed.) *Die Musik in Geschichte und Gegenwart,* 14 vol. (1949–68).

Survey histories: DONALD J. GROUT, *A History of Western Music* (1960); RICHARD L. CROCKER, *A History of Musical Style* (1966); PAUL HENRY LANG, *Music in Western Civilization* (1941); *New Oxford History of Music,* 4 vol. of projected 11 (1954–68); HOMER ULRICH and PAUL A. PISK, *A History of Music and Musical Style* (1963).

Period monographs: CURT SACHS, *The Rise of Music in the Ancient World, East and West* (1943); GUSTAVE REESE, *Music in the Middle Ages* (1940), *Music in the Renaissance,* rev. ed. (1959); MANFRED F. BUKOFZER, *Music in the Baroque Era, from Monteverdi to Bach* (1947); ALFRED EINSTEIN, *Music in the Romantic Era* (1947); WILLIAM W. AUSTIN, *Music in the Twentieth Century, from Debussy Through Stravinsky* (1966); ERIC SALZMAN, *Twentieth-Century Music* (1967).

(R.T.D.)

Musical Composition

Both the process of "putting together" a piece of music and the finished product susceptible to repeated rendition are designated by the term musical composition. In Western civilization musical composition refers to the finished work primarily as fixed by a complex system of graphic notation, or, more recently, by sound recordings. This article will examine different cultural approaches to composition as well as changing concepts and specific procedures associated with the stylistic development of Western music.

THE PROCESS OF COMPOSITION

Varying societal conceptions. Whether referring to the process or to the completed work, composition implies the creation of a unique musical event that may or may not be based on original musical materials. At certain cultural levels and in many non-Western societies unique performance characteristics tend to assume greater significance than composition in itself. In oral traditions, related variants of common origin often take the place of unalterable musical entities, so that tune families rather than single autonomous tunes form the collective repertory. Where certain patterns of musical structure have gained broad recognition (as the ragas, or melody types, of India), musicians will as a rule rework such patterns extemporaneously, though in accordance with prevailing attitudes and conventions.

Oral tradition

European music was communicated orally well into the Middle Ages and received important stimuli from a variety of oral traditions even after musical notation had developed to a high degree of precision. Indeed, the lower population strata, especially in rural areas, never abandoned the relative freedom that comes from reliance on the ear alone, and the sophisticated music of the upper strata, throughout its rapid evolution, rarely severed its connection with folk music altogether. Ultimately, the process of composition, as seen by the American musicologist Alan P. Merriam, does "not seem to differ radically between literate and non-literate peoples save in the question of writing." As a conscious act of social communication it always "involves learning, is subject to public acceptance and rejection, and is therefore a part of the broad learning process which contributes, in turn, to the processes of stability and change." Whether explicitly or not, composition is thus subject to rules that represent the stylistic consensus of a specific segment of society at a given stage of cultural development. During the Middle Ages, when man's natural instincts were held in particularly low esteem, musical compositions were often judged primarily in terms of their adherence to the rules. Hence, the supreme authority in matters musical was the *musicus* as theorist; only he was considered sufficiently conversant with musical science, "the discipline which treats of numbers in their relation to those things which are found in sounds" (Cassiodorus, 6th century AD), to vouchsafe its continued existence as the sonorous embodiment of universal truths. And it was because the metaphysical properties of numbers were allegedly embedded in the rules of composition that music, on a par with arithmetic, geometry, and astronomy, attained and retained an honorable place as a constituent member of the quadrivium, the more exalted of the two divisions of the seven liberal arts. Characteristically, music was not classified with grammar, rhetoric, and logic, the "rhetorical arts"

gathered in the trivium. About 1300 musical composition as a mere craft was ranked by Johannes de Grocheo, a shrewd observer of the Parisian musical scene, with shoemaking and tanning.

Musical elements. At its most fundamental level the act of composition involves the ordering of pitched sounds in musical time and space. Pitch relationships are referred to as intervals; their specific occurrence in musical time is determined by rhythm, a concept that embraces all durational aspects of music. Rhythm in turn may or may not be regulated by metre. In metrically organized rhythm, recurring patterns of accented and unaccented "beats" furnish a durational substructure that necessarily affects all the other elements of composition, including the nature of melody, harmony, and texture. Metrical rhythm is nearly always present in dance music because its patterning is largely analogous to that of bodily motions and step figurations. But logogenic, or word-determined, music also often employs metrical patterns, corresponding as a rule to those of the poetic text. The first large corpus of logogenic compositions transmitted through the ages is that of medieval plainchant, consisting of monophonic settings (limited to a single melodic line) of liturgical texts for the entire year, based on a system of eight church modes, diatonic scales abstracted from the melodic motives utilized by medieval singers. Modality—whether referring to a melodic or a rhythmic framework—furnishes compositional frames of reference in a wide variety of essentially monophonic musical styles, especially in Asia. Asian influences upon early European music cannot be ruled out, whether by way of ancient Judaea, Greece, Byzantium, or the medieval Arab invasions. But unlike their Asian counterparts, Europeans at first limited modality to melody, through pitch arrangements. The rhythmic properties of plainchant have largely remained a matter of conjecture, for no systematic discussion of plainchant rhythm survives, and the notation used was noncommittal with respect to rhythm. By the same token, plainchant no doubt owed much of its amazing vitality to the absence of an all-encompassing notation, which made possible the flexibility of performance and regional variation inherent in a partly written, partly oral tradition.

Rhythm and metre

Music like medieval plainchant, in which the lengths of individual tones tend to be rather uniform, is often referred to as nonrhythmic or rhythmless. Such careless terminology denies the very essence of music as a temporal art, which implies by definition the omnipresence of rhythm as "order in musical time." Actually, the relative presence or lack of rhythmic differentiation in the duration of tones can act as a decisive stylistic determinant. Thus the rhythmic equanimity of the monophonic plainchant, at least in the interpretation set forth by the 19th-century Benedictine monks of Solesmes, France, and recognized as authoritative by the Roman Catholic Church, effectively symbolizes an atmosphere of faith and inner peace. By contrast, the strictly metrical organization of rhythm in most 18th-century music reflects the thinking of an age of reason, favouring mathematically definable, hence "natural," structures in its music.

Motive and melody

The smallest melodic-rhythmic unit (minimally two separately perceived sounds) is the motive. Pitched sounds are, however, not of the essence: drum motives are so effective rhythmically precisely because they lack pitch definition. By and large, rhythmic motives are used to endow pitch relationships with identifiable durational characteristics. And consequently rhythmic identity often serves to establish motivic connections between different intervals. A famous case in point is the opening short–short–short–long motif of Beethoven's *Symphony No. 5 in C Minor*, Opus 67, which serves as an effective element of structural cohesion in this large-scale work.

Types of melody owe their aesthetic associations in many instances to their motivic peculiarities. In Western music motivic contrast has been identified with emotional conflict since at least the mid-16th century, when composers of madrigals (Italian polyphonic secular songs) began to set dramatic texts. The opening of Mozart's *Symphony No. 35 in D Major*, K 385 (the *Haffner Symphony*) offers an excellent example. Analogous in its motivic structure to a section of the first act of Mozart's opera *Don Giovanni*, the opening of the symphony engenders emotional contrasts similar to those inherent in the opera's dramatic action when Donna Anna, under the double impact of attempted rape and her father's violent death at the Don's hands, impulsively rejects Don Ottavio's sympathy until, realizing that she has no one else to rely on for help, she reverses herself and induces him to swear revenge. Conversely, melodic lyricism correlates with a high degree of motivic affinity.

HISTORICAL APPROACHES TO COMPOSITION

The Middle Ages. The European written tradition, largely because it evolved under church auspices, de-emphasized rhythmic distinctiveness long after multipart music had superseded the monophonic plainchant. But multipart music might never have gone beyond the most primitive stages of counterpoint (from Latin, *punctus contra punctum*, "note against note") had it not been for the application of organized rhythm to musical structure in the late Middle Ages. This era witnessed the emergence of basic polyphonic concepts identified with European art music ever since. The precise measurement of musical time was simply an indispensable prerequisite for compositions in which separate, yet simultaneously sounded, melodic entities were combined in accordance with the medieval theorists' rules of consonance (specifying the proper intervals to be used between voice parts, especially at points of musical repose). Toward the end of the 1st millennium of the Christian Era church singers had grown accustomed to enhancing their chants through organum—*i.e.*, the simultaneous singing of the same melody at different pitch levels, specifically the "perfect" intervals of the octave, fifth, and fourth (distances of eight, five, or four notes of a diatonic scale). "Parallel" organum of this type was followed, in turn, by "free" organum, which allowed the synchronized voice parts to utilize contrary melodic motion.

The decisive relationship between text and melody in early European music led to stylistic distinctions that have survived the ages. Thus, "syllabic" denotes a setting where one syllable corresponds to one note; "melismatic" refers to a phrase or composition employing several distinct pitches for the vocalization of a single syllable. Late medieval composers made clever use of these distinctions, including an intermediate "neumatic" style (Greek *pneuma*, "breath") to create ever more extensive polyphonic pieces. By the 12th century musicians at Notre-Dame in Paris, led by Léonin, the first polyphonic composer known by name, cultivated a type of melismatic organum that featured a highly florid upper part above a slow moving cantus firmus taken from a suitable plainchant melody. The melismatic sections alternated with strictly measured, or "discant," sections. This very effective procedure possibly was inspired by Middle Eastern practices with which the crusaders must have been well acquainted. In Eastern music, the rhythmically measured portions following the virtuoso singer's florid "outpouring of the soul" are nearly always played or at least supported by instruments. In the 13th century the clausula, a short, textless composition in discant style, tended to be dancelike in its systematic sectionalization, strongly suggesting instrumental derivation, if not necessarily actual performance. The motet, a major genre of the medieval and Renaissance eras, was in its 13th-century form essentially a texted clausula, frequently employing two or three different texts in as many languages. This fact merely reinforces the suspicion that little distinction was made between vocal and instrumental composition in an era that so blithely based dancelike settings of erotic, in a few instances outright obscene, texts on a chant-derived cantus firmus. The point is not without its broader ramifications. For, brought up largely on 19th-century notions about the "purity" of church music, one easily overlooks the fact that even Bach and Mozart had few compunctions about the use of secular—in their cases mostly operatic—styles and specific tunes in church music. Over the centuries, the church has been the most

Rhythmic contrast as a structural device

important employer of composers and has offered far greater outlets for newly created music than any other social institution or category. Thus composers of sacred music have had to satisfy the aesthetic needs and expectations of its highly differentiated "public." The church in turn repeatedly permitted the adaptation of promising secular types of composition, even though instrumental music, because of its more lascivious associations, remained suspect well into the 17th century.

In accordance with medieval tendencies generally, Gothic polyphonic music was conceived in loosely connected separate layers. Thus, two-part motets could be converted into three-part motets, and Léonin's successor Pérotin expanded the organum to three and four parts. Inevitably, as their compositions gained in length and depth, musicians began to search for new integrative procedures. A system of six rhythmic modes (short, repeated rhythmic patterns) evolved rapidly. Pérotin used a single rhythmic mode for the multiple upper parts of his organums so that, separated from their cantus firmus, they resembled the conductus, a syllabic setting of a sacred text for two or three voices sharing the same basic rhythm. Finally, as organum faded into history conductus-type motets were composed outright. Most prominent among the devices used to achieve structural integration in the 13th century were *color*, or melodic repetition without regard to rhythmic organization; *talea*, or rhythmic repetition without regard to pitch organization; and ostinato, or repetition of a relatively brief melodic-rhythmic pattern. Exchanges of melodic phrases between two or more parts in turn led to canon, a form in which all voice parts are derived from one tune—either by strict imitation of the basic melody or by manipulations stipulated in often quite sophisticated verbal instructions (canon = law). For instance, the canon *Ma fin est mon commencement* (*My End Is My Beginning*), by Guillaume de Machaut, the leading French composer of the 14th century, demands the simultaneous performance of a melody and its retrograde version (the notes are sung in reverse order). French musicians of the 14th century were particularly partial to isorhythm, which refers to repetition of the rhythmic organization of all the voices in a given compositional segment. It enjoyed considerable popularity for more than a hundred years.

Melodic and rhythmic manipulation

Meanwhile, though somewhat eclipsed historically by the increasingly abstract nature of polyphony, the primacy of poetry was safeguarded in 13th-century music by the troubadours of southern France and their northern counterparts, the trouvères, as well as the German Minnesingers. These noble poet-composers created a rich tradition of purely monophonic secular song that furnished convenient points of departure for much of the secular polyphonic music in both 14th-century France and 15th-century Germany. By the beginning of the 15th century, European music had also begun to feel the impact of English music. The English emphasis on the rich sonorities of the third and sixth provided welcome relief from the aesthetic consequences of the earlier continental dedication to the "perfect" intervals of the octave, fourth, and fifth. Because the perfect intervals were also those formed by the lowest pitches of the harmonic overtone series (faint tones sounding above a fundamental tone, caused by a sound wave vibrating both as a whole and in parts), their "naturalness" had long been an unassailable theoretical axiom.

Impact of English music

Late 14th-century French secular music literally lost itself in rhythmic complexities without any substantive changes in the basic compositional approach, which continued to favour relatively brief three-part settings of lyrical poetry. But the ensuing 15th century eagerly embraced the simpler melodic and rhythmic ideas associated with the rich harmonies of the English style; often melodies were outright triadic in contour; *i.e.*, they outlined the intervals of the triad, an increasingly important chord composed of two linked thirds (*e.g.*, C–E–G). But the truly amazing stylistic development from the influential English composer John Dunstable (*c.* 1385–1453) to Josquin des Prez (*c.* 1445–1521), the Franco-Flemish composer who stands at the apex of his era, was equally indebted to the flowing cantilenas, or lyric melodies, that characterized the top parts of Italian trecento music. If the French music of the waning Middle Ages was structured essentially from the bottom up, with relatively angular melodic and rhythmic patterns above the two-dimensional substructure of tenor and countertenor, its Italian counterparts were quite often monodically conceived; *i.e.*, a highly singable tune was sparingly yet effectively supported by a single lower voice. Indeed, the passion for melody, if need be to the detriment of other musical elements, has been a constant of Italian music. It sparked the *nuove musiche*, or "new music," of about 1600 and is exemplified in innumerable works of composers as diverse as Claudio Monteverdi (1567–1643) and Luigi Dallapiccola (1904–). But it found its first major artistic expression in the city-states of northern Italy during the lifetimes of such 14th-century literary figures as Giovanni Boccaccio and Petrarch.

The Renaissance. During the latter part of the 15th century, French rhythmic sophistication, Italian cantilena, and English harmony finally found common ground in the style of Renaissance polyphony that, under the aegis of Flemish musicians, dominated Europe for nearly two centuries. Often referred to as modal because it retained the medieval system of melodic modes, Flemish polyphony was characterized by a highly developed sense of structure and textural integration. Although the older cantus firmus technique was never totally abandoned, Renaissance polyphony is identified above all with imitative part writing, inspired no doubt by earlier canonic procedures but devoid of their structural limitations. After a canonic or freely imitational beginning, each of the subunits of such a polyphonic piece proceeds unfettered by canonic restrictions, yet preserves the fundamental equality of the melodic lines in accordance with contrapuntal rules amply discussed by various 15th- and 16th-century theorists and ultimately codified by the Italian theorist Gioseffo Zarlino. Through the works of Giovanni da Palestrina, the model composer of the Catholic Counter-Reformation, Renaissance modal counterpoint has influenced the teaching of musical composition to this day, suggesting the near perfection with which it conveys some fundamental aspects of the historic European ideal of composition as the art of lasting musical structures.

Whereas imitative polyphony affected virtually all 16th-century music, modal counterpoint was paramount in sacred pieces, specifically the motet and mass, probably because of its close kinship with the traditional modality of liturgical plainchant. In contrast, the beginnings of functional harmony (chordal relationships governed by primary and secondary tonal centres) manifested themselves first in the polyphonic French chanson, its Italian counterpart, the madrigal, and related secular types. Under the influence of less sophisticated music, such as that of the Italian frottola, a popular vocal genre, these secular polyphonic genres favoured rather simple bass lines highlighting a limited number of related harmonies. Thus, undisturbed by the theoretical writings from the pens of church-employed musicians, secular musical practice in the later Renaissance laid the foundations for the harmonic notions that were to dominate three centuries of Western art music. The increasing emotionalism of texts taken from the leading Italian poet of the 16th century, Torquato Tasso, and his immediate successors, acted as a further stimulant, as Italian composers, searching for appropriate musical symbols, discovered the expressive possibilities of chordal progressions.

Changing conceptions of harmony

The Baroque era. Inevitably, the strong desire for heightened expression through harmony led at first to new, mostly chromatic, chord progressions (using notes foreign to the diatonic scale). Eventually it precipitated the total abandonment of traditional polyphony about 1600 in the monodic experiments of the Florentine Camerata, a group of aristocratic connoisseurs seeking to emulate the Greek drama of antiquity. The accompaniment for these passionate and heroic solo recitations is based on a simple basso continuo. Only the bass part was

written down; it was played by low, sustaining instruments bowed or blown, while plucked or keyboard instruments supplied the chords suggested by the bass and melody lines. The small figures used to indicate the proper harmonies gave the system the alternative name "figured bass." Monody had its historical antecedents in mid-16th-century solo lute songs and in the plentiful arrangements of polyphonic vocal compositions for single voices accompanied by plucked instruments and for solo keyboard instruments. But it was the attempt to resurrect the spirit of antique drama in the late Renaissance that created the textural revolution that has been equated with the beginnings of modern music: the monodic style with its polarity of bass and melody lines and emphasis on chords superseded the equal-voiced polyphonic texture of Renaissance music. Monteverdi, the undisputed master of the monodic style, recognized the possibility of two basic approaches to composition: the first or polyphonic "practice" and the second or monodic "practice." Thus, with penetrating analytical insight he formulated the basic stylistic dialectic that has since governed the course of Western music.

Aesthetic and technical reorientation

The emergence of an essentially nonpolyphonic style went hand-in-hand with the rise of a variety of specifically instrumental idioms. Not only did accompanied vocal music offer instrumentalists various opportunities for improvisation; the basically chordal style also facilitated the emergence of virtuosity in the modern sense of the term, especially among keyboard artists. But as the singer and composer Giulio Caccini demonstrated in the preface to his influential collection *Le nuove musiche* (*The New Music;* 1602), singers, too, put their newly found freedom to good improvisational and ornamentational use. In short, after two centuries dominated by the highly structured, rationalistic polyphony of the Renaissance, the performing musician reiterated his creative rights. Inevitably, under such forceful pressures, the teaching of composition, previously tied to the laws of modal counterpoint, quickly shifted to the harmonic challenges of the figured bass.

Because the bass-oriented music of the 17th century relied primarily on chord progressions as fixed by the bass notes, it was structurally quite open-ended; *i.e.*, the new technique suited any number of formal patterns. Even so, the incipient rationalism that was to reach its peak in the 18th century soon led to the consolidation of broadly accepted structural types. Indeed, the very concept of musical "form," as generally understood from the late 17th century on, was intimately tied to the growing importance of instrumental music, which, in the absence of a text, had nothing to rely upon save its own organically developed laws. At least for a while, vocal music, which had been so largely responsible for the monodic revolution, continued to adhere to the Monteverdian principle that the words must act as "the mistress of harmony." Both melody and harmony, therefore, reflected often minute affective textual differentiations. And as late as the early 18th century similar musico-rhetorical considerations led to *Affektenlehre*, the theory of musical affects (emotions, feelings), developed primarily in Germany. Following this theory, German musicians dealt with composition systematically in terms of a specific but broadly adopted expressive vocabulary of melodic, rhythmic, and harmonic figures. Meanwhile, the Italians laid the foundations for such lasting categories of instrumental music as the symphony, the sonata, and the concerto. In each instance the structural outline was harmonically determined through juxtapositions of principal key areas acting as focal centres of tonality. As for tempo, the earliest 17th-century solo sonatas had relied on drastic short-range changes in accordance with a general predilection for "instant sensations." Subsequently, as musical composition fell in line with the prevailing rationalistic trend, tempo served above all as a means of differentiation between the various movements, or self-contained sections that constituted the large-scale works of the Italian string school and of French and German instrumental composers as well. Texture, too, was used to provide contrast, particularly within a given movement, as in the concerto grosso with its alternation between small and large groups of players (concertino and tutti).

Emotional expressiveness

Interrelated with the spectacular rise and amazing vitality of instrumental music was its unprecedented variety. By the early 18th century, composers drew freely upon everything from contrapuntal forms like the fugue (an adaptation of the imitative techniques of the Renaissance motet within the context of functional harmony) to stylized popular dances, such as those that make up the suites and partitas of J.S. Bach. The figured bass era took full advantage of the possibilities of variety and contrast through judicious manipulations of all elements of composition. Whereas accompanied solo music pitted bass against treble (the latter often split up into two parts, as in the trio sonata), composers generally liked to juxtapose figured bass and polyphonic textures. Melodically, the far-flung phrases of Italian bel canto, the florid singing style characteristic of opera seria (17th- and 18th-century tragic opera), had little in common with the concise, symmetrically balanced phrases found in music of popular inspiration, whether in opera buffa (Italian comic opera) or the many types of dances. As for the latter, their impact on sophisticated 18th-century music is evident not only in many dance-inspired arias and concerto movements but is also in certain polyphonic compositions. Both the chaconne and passacaglia, related polyphonic types, were based on dancelike ostinato patterns, often with specific harmonic implications. Perhaps the most famous example is Bach's "Chaconne" for solo violin, which concludes the *Partita in D Minor*.

Major structural developments of the Baroque

Even though the Baroque preoccupation with style worked somewhat to the detriment of structural definition, certain closed forms did gradually emerge. The da capo aria distinguished clearly between an initial section (A), a contrasting section (B), and the repeat (da capo) of the initial section, as a rule with improvised vocal embellishment. In instrumental music, the French opera overture began with a slow, stately introduction followed by a fast, often fugal movement, whereas its Italian counterpart had a tripartite fast–slow–fast scheme. Dance-based suite movements were binary in outline: the first of the two sections, each separately repeated, moved to the dominant key (a fifth above the tonic or principal key) or to the relative major (a minor third above the tonic) in the case of a piece in a minor key; the second section, after some modulatory activity (*i.e.*, passing through several key areas), returned to the central key. Even more decisive in its far-reaching historical consequences was the structural organization of a number of the keyboard "sonatas" of the composer Domenico Scarlatti. These works consisted of single, essentially binary movements, the first section of which differentiated not only between two key areas but two contrasting thematic ideas as well.

The Classical period. The Classical era in music is compositionally defined by the balanced eclecticism of the late 18th- and early 19th-century Viennese "school" of Haydn, Mozart, Beethoven, and Schubert, who completely absorbed and individually fused or transformed the vast array of 18th-century textures and formal types. Expansion of the tripartite Italian overture had produced the basic three-movement scheme of the symphony even before the 18th century reached midpoint. Shortly thereafter, the minuet, borrowed from the dance suite, was inserted with increasing frequency as a fourth movement, between the slow movement and the fast finale. The French opera overture in turn lent its slow introduction where needed for structural variety. Texturally, homophony (chordal texture) and polyphony soon assumed rather specific roles, with polyphonic writing usually reserved for the central or development section of the classical first-movement form. The organic fusion of a number of stylistic traits previously associated with strong and immediate contrast is exemplified by the obbligato accompaniment, the texture most typical of Viennese classicism. Here the relative equality of all the melodic parts in a given composition is ensured without denying the melodic supremacy of the treble and the harmonically decisive role of the bass. The evolution of

Integration of homophony and polyphony

this characteristic texture can be traced in the string quartets of Haydn. At first, following earlier 18th-century custom, Haydn wrote strictly treble-dominated compositions with a simplified bass (as compared with the more varied basso continuo); then, with the six *Sun Quartets* (Opus 20), dating from the early 1770s, he defied precedent and concluded each work with a fugue in the "learned style" of Handel. Finally, in his *Russian Quartets,* Opus 33, written, in his own words, "in a new manner," Haydn achieved the fusion of elements of both the learned and the treble-dominated styles. The result was a harmonically oriented, yet polyphonically animated, texture that was to affect both instrumental and vocal ensemble music for generations. It was also at this point, when compositional procedures reached a degree of stability and universality unmatched since Renaissance polyphony, that composition began to be taken seriously as a separate musicianly discipline. Johann Joseph Fux's famous *Gradus ad Parnassum* (*Steps to Parnassus*), published first in Latin in 1725 and subsequently in every important modern language, was still basically a didactic treatise on counterpoint abstracted from 16th-century practice. As such it served its purpose throughout the 18th century, while harmony continued to be taught as the art of accompaniment—*i.e.*, the improvised realization of a figured bass. But eventually the general fascination with comprehensive knowledge, sparked by the French *Encyclopédie,* inspired at first sporadic, then ever more numerous volumes dealing progressively with all aspects of composition. During the ensuing 19th century the rapid institutionalization of musical education in the image of the National Conservatory of Music in Paris, created while the French Revolution was still raging, added further to the academic systematization of all musical studies along lines that have essentially remained in force. Thus the teaching of musical composition reflects to this day the biases of the 19th century, specifically its concern with functional harmony as the principal generative force in music—a doctrine first proclaimed in the 1720s in the name of nature (as being consistent with the harmonic overtone series) by the composer and theorist Jean-Philippe Rameau.

Composition as an academic discipline

The Romantic era. With the onset of the Romantic era in the wake of the French Revolution, composers began to view their own role in society as well as the social function of their work, and hence also its aesthetic prerequisites, in a radically different light. With respect to social function, Beethoven was actually the first musician of stature to achieve emancipation in the sense that his work reflected, with relatively few exceptions, purely personal artistic concerns. He simply took it for granted that patrons would supply funds sufficient for him to pursue his creative career unfettered by financial worries. This attitude represents a total reversal of the basic assumptions of the preceding century, when composers were hired by and large to satisfy the musical needs of specific individuals or institutions.

The view of the composer as artist also changed. If during the Middle Ages the craft of musical composition had been evaluated largely in terms of its strict adherence to established rules, instinctiveness and spontaneity had remained suspect well into the Italian Renaissance. For a 15th-century composer–theorist like Johannes Tinctoris, the value of a musical composition depended on learned judgment as well as spontaneous reaction. Thus his admiration for certain composers of his time stemmed both from the happiness and from the enlightenment that he found in examining their music. But the Swiss theorist Henricus Glareanus, writing 70 years later, explicitly preferred natural talent to the most exquisite craftsmanship. The Renaissance was the first epoch in European intellectual history to recognize that the greatness of a composer rests upon his inherent talent and unique personal style, and that genius supersedes both experience and the observance of theoretical precepts. Likewise, it was the first era in which the process of composition was viewed as linked to powerful internal impulses. The rising tide of academicism notwithstanding, this basic attitude on the whole dominated the European scene more or less consistently from then on. According to E.T.A. Hoffmann, the early 19th-century poet, critic, and composer, "effective composition is nothing but the art of capturing with a higher strength, and fixing in the hieroglyphs of tones, what was received in the mind's unconscious *ecstasis.*" And Romantic composers from Schumann and Chopin to Hugo Wolf and Gustav Mahler did in fact produce much of their very best creative work in precisely such a state of exaltation, in a few tragic instances (*e.g.*, Schumann and Wolf) to the ultimate detriment of their sanity.

The composer as genius

The aesthetic effects of this drastic change in conception of the composer's task and potential were immediate and far reaching. For one, every large-scale composition assumed artistic significance of a type previously accorded only a whole series of works, sometimes a composer's entire output. And concomitantly, many leading composers of the 19th century wrote in considerably smaller quantities than their predecessors. But in exchange they revelled in idiomatic and structural peculiarities even in works that nominally fell into the same formal category. Thus although "characteristic" symphonies alluding to nonmusical ideas occurred occasionally in the late 18th century, virtually every symphonic composition postdating Beethoven's *Symphony No. 3 in E Flat Major,* Opus *55* (*Eroica;* completed 1804), could be so designated. "Characteristic" works like Beethoven's *Symphony No. 6 in F Major,* Opus 68 (*Pastoral;* 1808), or his overture to Goethe's drama *Egmont* are but one step removed from the kind of characteristic scenes that make up the *Symphonie fantastique* of the French composer Hector Berlioz or, for that matter, Felix Mendelssohn's *Hebrides* (also known as *Fingal's Cave*), an overture unrelated to any particular drama, spoken or sung. Franz Liszt, in the free-wheeling forms of his symphonic poems, simply pursued the individualistic line to its ultimate consequences, severing whatever tenuous ties to traditional structures the works of his immediate predecessors had still maintained. The Romantic composer viewed himself basically as a poet who manipulated musical sounds instead of words. But if the composers catered to poetry, writing Lieder (German songs) and attempting to retell stories in instrumental works, the poets looked with awe and envy upon the composers' use of a language so utterly dissociated from material existence. "All art aspires to the condition of music," said Wordsworth. It is thus hardly surprising that opera, whose extramusical connotations had in the past been responsible for some of the most daring stylistic innovations, rapidly incurred the disfavour of progressive composers. Although some, like Berlioz, Mendelssohn, and Schumann, tried their hands at an occasional opera, others, including Chopin, Liszt, and Brahms, felt no inclination whatever to compose for the stage. Instead, each developed personal idioms capable of a depth of expression that words could not match. Mendelssohn spoke indeed for many when he remarked that, as far as he was concerned, music was more precise in meaning than words.

Nonmusical allusions

As in the late Renaissance, harmony once again furnished the primary expressive means. In defining musical structure, too, harmonic and modulatory procedures predominated at the expense of the contrapuntal interplay of motives. Numerous Romantic composers excelled in concise forms of strong melodic-harmonic import, variously entitled Impromptu, Nocturne, Song Without Words, Ballade, Capriccio, Prelude, Etude, etc. The form of these works was nearly always tripartite, with a literal or modified repeat of the first part following a melodically and harmonically contrasting middle section. Works of larger scope often consisted of a series of relatively autonomous subunits tied together either by the same tune presented in different guises (as in variation sets) or by fairly literal recurrences of an initial musical idea (the rondo principle). Compositions of the Classical sonata-allegro type, to which motivic-contrapuntal development was essential, inevitably suffered from the Romantic love for pure, harmonically defined melody. Thus Tchaikovsky frankly admitted in 1878 that, although he could not complain of poverty of imagination or lack of inventive power, his lack of structural skill had frequently caused

Interrelation of aesthetic, technique, and form

his "seams" to show: "there was no organic union between my individual episodes." Composers such as Tchaikovsky were indeed particularly successful with chainlike formations like the serenade or the ballet suite, which comprised a well-calculated number of carefully wrought smaller entities.

In the context of functional harmony, the Classical motivic-contrapuntal approach had no doubt been exploited in the last sonatas and string quartets of Beethoven to the very limits of its potential to define musical structure. The heroic image of Beethoven as one who had overcome every possible personal and artistic difficulty to achieve the highest aims of the art assumed well-nigh traumatic proportions among 19th-century musicians. Not only did composers ill equipped both by training and artistic temperament try to emulate him, but theorists from Adolf Bernhard Marx to Vincent d'Indy based treatises on his works. Thus, unwittingly the Classical Beethovenian inheritance turned into something of an aesthetic liability for Romantic composers swayed by the image of Beethoven and unable or unwilling to face the fact that their particular talents were totally unsuited for any further capitalization of his basic compositional procedures. Confronted with the task of writing in the Beethovenian manner, a great master like Schumann, who had created the near-perfect, totally Romantic suite *Carnaval*, Opus 9 (1835), was clearly out of his element: the development of his *Symphony No. 1 in B Flat Major*, Opus 38 (*Spring;* 1841), offers a prime example of the "rhythmic paralysis" that affected so many large-scale 19th-century works. That this symphony managed nevertheless to maintain itself in the concert repertory, on the other hand, demonstrates the extent to which the best among the German composers compensated for obvious weaknesses in handling motivic development by sustaining above all constant harmonic interest. For their part, the French, always coloristically inclined, turned instrumentation into a principal compositional resource, so that in an unadorned piano transcription Berlioz' *Symphonie fantastique* retains little more than its basic contours. That by the end of the century virtuoso instrumentation had become universal practice is attested by any work of Richard Strauss or Gustav Mahler.

Characteristically, the most unique compositional achievement of the 19th century, that of Richard Wagner, was also the most eclectic. Wagner represents the apotheosis of Romanticism in music precisely because he fused into musico-poetic structures of unprecedented proportions virtually every musical resource that went before him. Seen in this light it may be more than mere coincidence that *Tristan und Isolde*, perhaps Wagner's most perfect music drama, begins with the same four notes that make up the motivic substance of Beethoven's string quartets Opuses 130–133. Unlike most instrumental composers after Beethoven, the dramatist Wagner fully assimilated the motivic-contrapuntal process, even though his texture is principally determined by strong harmonic tensions and by a masterful use of instrumental colour in the vein of Berlioz and French grand opera. Just as he integrated diverse compositional techniques Wagner also achieved a balance of musical and poetic elements so perfect that critics, both favourable and unfavourable, have never ceased to be puzzled by its aesthetic implications. How consciously Wagner proceeded is attested not only by his numerous theoretical writings but also by compositional sketches pointing in some instances to several stages of mutual adjustments involving music and text.

Wagner's fusion of counterpoint with late Romantic harmony

Compositional approaches of the 20th century. Wagner's highly expressive harmonic bequest could not but drive chromaticism eventually beyond the retaining confines of the idea of a central key, for the extensive use of chromatic chords tends to blur the listener's ability to perceive the basic harmonic relationships that define a key. In their nontonal compositional procedures, Arnold Schoenberg and his 20th-century Viennese school abandoned the concept of key, using all notes freely without relating them to the system of functional harmony. They thus represent not so much a reaction to as a logical extension of Wagnerian principles. Wagner's compelling artistic personality certainly spelled near-magic powers over many of his younger contemporaries and successors, overshadowing even Beethoven. But others, too, contributed to "the music of the future." As Schoenberg was to point out in one of his remarkable essays, even Brahms, who looked upon himself as a conservative in the best sense of the term, was, historically speaking, a true "progressive," especially in his propensity for irregular phrasing and complex motivic manipulations.

The growth of political nationalism in the "peripheral" countries of Europe also had significant repercussions in musical composition. In the last half of the 19th century distinctive folk-music elements, previously totally unheeded by Europe's elitist musical cultures, found enthusiastic response in sophisticated circles, exerting an "exotic" attraction similar to that which had accounted earlier in the century for the Romantic infatuation with Eastern civilizations. Thus at a time when the exhaustion of Europe's "civilized" compositional resources appeared imminent, the "untutored" harmonies of the Russian composer Modest Mussorgsky—steeped in the spirit of Russian folk music and based on chord progressions alien to the standard harmonic usage of his day—helped breathe new life into a harmonic language about to succumb to redundant overdoses of functional chromaticism. Mussorgsky thus paved the way for the later whole-tone and pentatonic (five-tone scale) experiments of Claude Debussy (1862–1918) and Béla Bartók (1881–1945).

What was being questioned publicly in many quarters at the dawn of the 20th century was the evolutionistic view of Western art music as man's ultimate achievement in the realm of sound and its logical consequent that 19th-century harmony represented in turn music's most advanced stage of development. This increasing skepticism, given by the nature of late 19th-century music itself, was strongly reinforced by the growing awareness of historical compositional techniques that resulted from the mushrooming discoveries of musicological scholarship. Before long, all manner of pre-19th-century textures and structural principles were seized upon to counteract the type of self-defeating post-Wagnerianism—so tragically exemplified in several of the most ambitious works of Max Reger. The 20th-century search for fresh, flexible techniques extended far beyond the nontonal Viennese school of Schoenberg. In historical perspective, Anton Webern's fascination with 15th-century canonic techniques, Paul Hindemith's predilection for both modal and early-18th-century polyphony, Igor Stravinsky's emulation of Domenico Scarlatti, and, for that matter, Kurt Weill's reinterpretation of John Gay's *Beggar's Opera* (1728), represent but various individual and culturally conditioned manifestations of the same determination: to put the burden of history to positive use in a concerted effort to revitalize an art that seemed moribund by the time World War I changed the socio-economic and political physiognomy of Europe.

Re-application of older techniques

Historically, Schoenberg's formulation of the laws of composition with 12 tones—involving the consistent melodic and harmonic use of a specifically arranged sequence of the 12 notes of the chromatic scale—sprang in the early 1920s from the same obsession with textural and structural clarity that marked the postwar Neoclassic syndrome as a whole. Schoenberg himself may have considered his most fundamental contribution to musical history to be "the emancipation of dissonance"—a relativistic conception of intervals and chords that disregarded the careful regulation of dissonance characteristic of functional harmony. Actually, the 12-tone procedures he developed so consistently served to restore, to an extent far beyond that which Gustav Mahler had been able to achieve within the traditional harmonic framework, the primacy of motivic-contrapuntal development as a musical resource. Thus it was that the profusion of simultaneous melodies that animates Mahler's *Symphony No. 4 in G Major* (completed 1900) found its ultimate potential realized in Schoenberg's most uncompromising polyphonic work, the *Wind Quintet* (1924).

Possibly the most successful attempt to regenerate Beethovenian procedures without the total abandonment of functional tonality is represented by the string quartets and certain other instrumental compositions of Béla Bartók. Drawing upon the rhythmic-melodic properties of Hungarian and Romanian folk music, Bartók produced a unique type of functionally extended harmony determined largely by the contrapuntal interaction of motives. Others, such as Charles Ives in the United States and, under the impact of South American popular music, the Frenchman Darius Milhaud, transcended traditional tonality by writing polytonally (in two or three keys simultaneously).

Whatever their specific approach, progressive 20th-century composers everywhere clearly gave precedence to melodic-rhythmic energies. Even instrumental colour was pressed into the service of melodic definition. Years before World War I Schoenberg had advocated in practice (*Five Orchestral Pieces*) and in theory (*Harmonielehre,* 1911; *Theory of Harmony,* 1947 [the English edition omits the pertinent chapter]) the idea of tone-colour melody, or *Klangfarbenmelodie.* But it was his pupil Webern who in his mature works divided the individual components of melodic phrases over several instruments as an imaginatively coloristic reinforcement of the complex polyphony that characterized his style. After World War II Webern's procedures were adopted enthusiastically by composers on both sides of the Atlantic. Living in increasingly automated societies, the post-Webern composers soon discovered total serialism, a manner of composition in which all musical parameters follow numerical rules laid down in the course of what has been called the precompositional process. Whereas Schoenberg's row technique merely fixed the sequence of the 12 pitches of the chromatic scale in accordance with the motivic context of a given piece, Webern had indeed begun to serialize rhythm and to some extent instrumentation, possibly under the influence of medieval isorhythmic techniques. But total serialization, as practiced by the post-Webernians, left little, if anything, to spontaneous inspiration, and the 1950s thus witnessed the closing of a creative circle initiated during the early Middle Ages when spontaneous inspiration was manifestly suspect. Perhaps inevitably, such a hermetically closed system of composition provoked reactions that moved the aesthetic pendulum violently to the extreme opposite position. Spur-of-the-moment action became the watchword in music as well as in life generally. Aleatoric (chance-music) composition in its more radical manifestations provides only minimal guidelines for performers who are told to improvise freely within certain temporal or spatial limitations, or both.

Serialization of all elements

Thus, as the 20th century enters its last quarter, Western man has returned, insofar as is possible, to the pretypographical stage when musical tradition was essentially oral. More than that, the relatively recent replacement of "the public," once essentially the cultural elite, by a whole host of publics to whom mutually exclusive types of composers cater with great solicitude, suggests that the private musical life of Western man in the immediate past is in the process of being transplanted by experiences serving, among other things, the identification of special group interests and needs. At the same time, novelty rather than originality has become the order of the day. Musical compositions, electronically recorded—more recently also electronically produced—and distributed as salable items by mass-oriented corporations, have attained the status of physical objects that are easily discarded and replaced. Still, there are those who try to transform the burdens of history into significant new forms of composition. A new eclecticism is in the making, destined perhaps to preserve a long tradition that has been among Western man's proudest achievements.

BIBLIOGRAPHY. Because most writings on musical composition were conceived as didactic treatises for would-be composers, their content has virtually no bearing on a better understanding of crucial aesthetic attitudes and mental processes. Instead, depending on the era from which they hail as well as the specific outlook of their authors, such treatises deal for the most part with contrapuntal rules, harmonic laws, and the like. While numerous books of this type appeared through the ages, it was only with the creation of the educational institutions known as conservatories of music, following the lead of the National Conservatory of France established in the last years of the 18th century, that composition became a discipline formally taught and hence requiring comprehensive textbooks including the theory of form and instrumentation. This profitable need was satisfied by the voluminous activities of 19th-century writers from J.J. Momigny and Anton Reicha to Vincent d'Indy in France and from Heinrich Christoph Koch to Hugo Riemann in Germany. Few of these books were translated into English, and they are of primary interest to the specialist. The following is a selection of easily accessible monographs and documentary collections with emphasis on firsthand testimonies of the creative artists themselves.

JACQUES BARZUN (ed.), *Pleasures of Music* (1951), a collection of writings about music, including many from the pens of men of literature; LEONARD BERNSTEIN, *The Joy of Music* (1959), contains Bernstein's seven "Omnibus" television scripts, including his excellent comments on sketch materials relating to Beethoven's *Fifth Symphony* and its ultimate realization; FREDERICK DORIAN, *The Musical Workshop* (1947), a discussion of various aspects of musical composition, including comments on creative procedures; HANNS EISLER, *Composing for the Films* (1947), rare insights into the problems of cinematographic music from a composer who, during his Hollywood years, tried to turn music into an integral aspect of film art; MAX GRAF, *From Beethoven to Shostakovich* (1947, reprinted 1969), a popular study of psychological processes involved in composition; MICHAEL HAMBURGER (ed. and trans.), *Beethoven: Letters, Journals and Conversations* (1960), a concise but well-chosen selection of Beethoven's own words concerning both specific compositions and the problems of the composer in general; LEJAREN A. HILLER, JR., and LEONARD M. ISAACSON, *Experimental Music* (1959), the first and still basic monograph dealing with the philosophy, procedures, and techniques of composition with an electronic computer; PAUL HINDEMITH, *A Composer's World* (1961), a leading 20th-century composer looks at various facets of the composer's world, including questions of musical perception, inspiration, technique, performance, and education; IRVING KOLODIN (ed.), *The Composer as Listener: A Guide to Music* (1958), excerpts from pertinent writings, including letters mostly by important composers of the 19th and early 20th centuries; EDWARD E. LOWINSKY, "Musical Genius: Evolution and Origins of a Concept," *Musical Quarterly,* 50:321–340, 476–495 (1964), excellent documentation of the evolution of a concept that has been associated through the ages more consistently with music than any other form of artistic production; ALAN P. MERRIAM, *The Anthropology of Music* (1964), a comprehensive treatment of music in relation to culture, drawing upon fieldwork on every continent and at all social levels (see especially ch. 9, "The Process of Composition"); SAM MORGENSTERN (ed.), *Composers on Music: An Anthology of Composers' Writings from Palestrina to Copland* (1956), an excellent collection, mostly of letters, dealing with general and specific aspects of the composer's work; ERNEST NEWMAN, *The Unconscious Beethoven,* rev. ed. (1970), a fascinating, if not unproblematic, study of the successive stages of Beethoven's work as revealed in his sketches, by one of Britain's most famous critics; GERTRUDE NORMAN and MIRIAM LUBELL SHRIFTE (eds.), *Letters of Composers: An Anthology, 1603–1945* (1946), one of the most comprehensive of several collections of composers' letters available in paperback; JOSEF RUFER, *Die Komposition mit zwölf Tönen* (1952, 2nd ed. 1966; Eng. trans., *Composition with Twelve Notes Related to One Another,* 1965), a cogent introduction to dodecaphonic aesthetics and technique by a dedicated Schoenberg disciple; WILLIAM RUSSO, *Jazz Composition and Orchestration* (1968), one of the very few serious treatments of a type of composition that has gained respectability only in recent years; ARNOLD SCHOENBERG, *Style and Idea* (1950), a discussion of Schoenberg's artistic motivation and procedures as well as those of the composers he admired most, including Brahms and Mahler; ROGER SESSIONS, *The Musical Experience of Composer, Performer, Listener* (1962), a discussion of fundamental musical problems by America's senior living composer; WILLIAM OLIVER STRUNK (ed.), *Source Readings in Music History from Classical Antiquity Through the Romantic Era* (1950), an indispensable collection of relevant excerpts from the writings of philosophers, musical theorists, and composers from Plato to Richard Wagner; D.F. TOVEY, *The Mainstream of Music, and Other Essays* (1961), a paperback reprint of some of the finest essays by the great British critic who discusses basic compositional issues.

(A.L.R.)

Musical Criticism

Musical criticism, properly considered, is a branch of philosophy—more exactly, of aesthetics. It is chiefly concerned with making value judgments, either about composition or about performance, or both.

Unfortunately, it is difficult to show that a value judgment can stand for anything that is even remotely true about music, as opposed to standing for something that is merely a personal whim on the part of the critic, since there is, as yet, no such thing as an organized body of knowledge called "musical criticism." The entire history of musical criticism can be summed up as a vain struggle to forge itself into a suitable tool for coming to grips with the art of music. Criticism always seems to founder on the same small handful of basic problems. These problems are essentially philosophical. They appear to be insoluble. They are aggravated not only by the esoteric nature of music, but also by the psychological mystery surrounding the very act of criticism.

Attempts at rationalization

There have been three major attempts to place musical criticism on a rational footing. These were by Sir Henry Hadow in *Music and Musical Criticism: A Discourse on Method* (1892), M.D. Calvocoressi in *The Principles and Methods of Musical Criticism* (1923), and Ernest Newman in *A Musical Critic's Holiday* (1925). Newman, in particular, was obsessed by a lifelong search for what he called the principles of "scientific" criticism. He maintained that musical criticism, if it were to succeed, must be raised from the level of a hit-or-miss affair to that of scientific certainty. Hadow and Calvocoressi were no less scientific in outlook, although they did not use the term. All three musicians were concerned that criticism lacked bite; their aim, subsequently, was to give it some teeth. Are there any "standards" in criticism? If so, can they be defined? Are they objective or subjective? If the latter, can they possibly be true? These questions are fundamental. They disclose the full range of the philosophy of criticism. The present article will review these basic issues and propose a direction in which they are likely to be resolved.

THE PRACTICE OF MUSICAL CRITICISM

The function of criticism. Musical criticism has a primary aim: the evaluation of music. How does the critic set about this difficult task? The scientific school of criticism holds that he applies certain standards to the work in question. His evaluation, that is to say, is the result of testing music against his critical yardsticks, and observing how far short it falls. According to this view, a value judgment is like a prize to be won by careful, objective, intelligent effort. This is an attractive notion, particularly among critics. It fosters the view that the critic is in a position of authority, and that he possesses the means to arbitrate over the creative artist. Unfortunately for criticism, there is nothing to suggest that this is anything but an illusion.

The scientific school

If one reflects on the way in which one listens to music, a basic fact is apparent. Music's value is inherent; it resides in the work of art itself. A value judgment is something that comes across as part of musical communication. Paradoxically, a value judgment appears to be necessary before the critical process can start. The consequences of this observation are far-reaching. Rather than critics with standards, there appear to be only works with standards, which critics observe. It is not necessary to prove that Mozart's *Jupiter Symphony* is a masterpiece in order to be certain that it *is* a masterpiece. Its mastery is self-evident. Critics did not bestow value on Mozart; they perceived it in him.

The intuitive nature of value judgments

Music as an autonomous communication. Music is autonomous. It refers to nothing outside itself. This sets it apart from the other arts, which rely upon the outside world for their images. A hat, a man, or an apple tree may all turn up in a painting, a sculpture, or a play. Indeed, they may be part of the very language of visual art, and therefore essential to its understanding. Music has no such aids toward comprehension. It is completely lacking in conceptual crutches. It develops according to its own laws. It is a purely musical truth that is comprehended on a purely musical level. The purity of musical communication is what the German philosopher Schopenhauer was admiring when he said that all art aspires toward the condition of music.

The inexplicability of musical experience

The totally musical nature of musical experience raises a difficulty for the critic. Such experience is virtually impossible for him to describe. Consequently, the critic can hardly be called "the man in the middle," a role frequently assigned to him. Music is not like a foreign language that requires an expert to translate it for a lay audience. It is a universal tongue. It either speaks to each listener directly or it does not speak to him at all. If it speaks, the critic's words are already redundant. If it does not speak, a problem exists that his words cannot solve.

Two important consequences flow from these views. They are among the axioms of criticism. First: since music is autonomous, all knowledge about it must spring from experience of it; practice, in short, precedes theory. Second: because of the purely private and personal nature of musical communication, a work's mastery can only be demonstrated to those listeners who already know about it, who have already experienced it—and they hardly require the demonstration. To those listeners who have not experienced it, a work's mastery is not demonstrable. If it were demonstrable, it goes without saying that critical differences would cease to exist: there would be nothing to prevent those who had experienced it from converting those who had not. Yet critical differences remain.

Objectivity versus subjectivity. A difficulty confronting all critics concerns the subjectivity of their observations. Since music is perceived subjectively, so the argument runs, does this not reduce criticism to mere personal opinion? And if this is so, what makes one critic's opinions any truer than another's? This objection can be disposed of, first, on the broadest philosophical level. Since all things perceivable are perceived subjectively, the charge of subjectivity must either be levelled against every other human endeavour, or it must be withdrawn from criticism. Second, and more to the point, what would be said of a performer who proclaimed to all the world his objectivity, his non-involvement with music? As for the composer, he would be thought strange indeed if he managed to avoid subjective entanglement with his creations. Why is it considered virtuous for performers and composers to enjoy an inner participation with music, and not critics? Quite clearly, there is a contradiction here.

The crucial question facing every critic is how he is to demonstrate the truth of his reaction. Yet all critics cannot be right; many are diametrically opposed to one another. It is no wonder that musical criticism has been described as stuck at the litmus-paper stage: critics take a dip into music and one sees what colour they turn. Plainly, criticism remains indistinguishable from mere speculation until the critic develops the means of confirming the truth of his views. If he wishes to develop such means, it is to the theory of criticism that he must turn.

A definition of criticism

Meanwhile, a definition of musical criticism emerges: Criticism is the rationalization of intuitive musical understanding.

THE THEORY OF MUSICAL CRITICISM

If the practice of criticism, as has been noted, can be reduced to one thing—expressing value judgments—the theory of criticism is essentially one thing, too—explaining them. It is not enough for critics to assert that one work is a masterpiece, another a mediocrity. An attempt must be made to explain why, and this may lead to a central discovery. A masterpiece is not a matter of chance, nor is a mediocrity. Both are symptomatic of deep, far-reaching principles.

Principle of unity. When Rudolph Reti, the Viennese critic, was a young man studying music at the Vienna Conservatory, he once stood up in the middle of a composition class and put the following question to his professor: "Why can't we take the themes of one work and

substitute the themes of another?" Reti did not receive a very convincing reply, and was therefore stimulated to think about the problem for himself. Forty years later, he worked out an answer in his book *The Thematic Process in Music* (1951). Briefly, it was that masterpieces diversify a unity. They grow from an all-embracing idea. Their contrasting themes hang together because each of them represents a different aspect of a single basic thought. This observation was not new. The Austrian composer Arnold Schoenberg had made it years earlier. So, too, had Heinrich Schenker, who used it as the basis for a major theory of aesthetics in his monumental *Das Meisterwerk in der Musik* ("The Masterpiece in Music," 3 vol.; 1926–29). Reti sharpened the concept. He made the critics think again about what, precisely, they mean when they talk about the integrity of a musical structure.

Diversity in unity

Reti's thesis can be vividly demonstrated by taking an existing masterpiece and substituting random themes from another. Even if such themes preserve a semblance of continuity (matching the key, metre, and mood of the themes they displace), they lose the deep sense of unity communicated by the original.

Not all musicians accept this theory. They argue that many composers, notably Bach, have put together works by borrowing materials from other sources. They fail to realize, however, that the act of borrowing is so highly selective that it, too, must be regarded as part of the creative process. The question then becomes: Why was that particular theme or movement borrowed?

Medium and idea. Another question is why a composition expresses itself through its particular medium? Why that medium rather than another?

If a masterpiece is transferred from the instrumental medium for which it was conceived to some other, alien medium, it undergoes a curious distortion. Such distortion offers the clearest proof that a musical law is operating in the original: an identity of the idea with its medium. A master's inner inspiration adapts to his outer terms of reference. Individual instruments lay down individual limitations. If a composer ignores this fact, he can never be certain that his creative aims will ever coincide with the musical results.

Collision of idea and medium

Occasionally, a great composer deliberately engineers a collision between the idea and the medium for a special musical purpose. The fugue in Beethoven's *Hammerclavier Sonata*, for example, is one of the most physically awkward works to play in the entire piano repertory. It has been composed against, rather than for, the instrument. Some bars are strictly unplayable, and Beethoven knew it; they contribute to the sense of struggle that is an essential part of musical communication, which is present even in the greatest performance. The orchestration of this fugue by the Austrian composer and conductor Felix Weingartner does a major service to musical aesthetics by providing an alternative musical experience of the same work; but by rendering the difficult easy, his orchestral version robs the music of its basic characteristics. It is a splendid illustration of the way in which, in Beethoven's original, a creative aim has been identified with a physical limitation.

Thematic chronology. A further question is why the chronology of the themes of a masterpiece cannot be changed. Why does one thematic chronology sound good, another bad? If the movements of a great sonata or a symphony are switched around, the result will be musically inferior. If the themes of one movement are mixed with those from the same work's companion movements, the result may even be an artistic disaster.

The creative principle of contrast distribution

An illuminating exercise in criticism is to "reconstruct" a masterpiece, so that its thematic running order is altered; that is, to transfer the first movement's second subject to the position occupied by the finale's second subject, and vice versa. Any musician can carry out this simple experiment for himself. Nothing is better calculated to reveal to him the presence of a creative principle of contrast distribution in the original. Given the thousands of different directions in which the material of a work could be unfolded, a master chooses the "right" one, the one that maintains structural tension—and, hence, musical interest. The themes of a masterpiece cannot assume one another's functions. They are born to fulfill specific roles. They develop out of each other because they create a musical need for each other.

Principle of economy. Composers through the ages have hinted at a law of economy toward which all great music strives. Brahms, in a characteristic piece of understatement, once said of composing that "the essential thing is that every note should be in its place." Beethoven expressed the same truth another way. Once, after he had heard the "Funeral March" from the opera *Achille* by the Italian composer Ferdinand Paer, he observed: "I must compose that!" To "fix" the idea, to define it, to pin it down—both composers felt that this was of the essence. Notes are redundant that do not stand for precisely those musical thoughts they are supposed to express. To have more notes, or less, than are actually required to communicate musical meaning must render such meaning correspondingly obscure. This law may be divided into three subsidiary principles.

Idea and utterance. First is the principle of identity between the idea and its utterance. There is a concrete musical difference between what a composer intends (the idea) and what he actually says (the utterance). Some musicians contend that the distinction is merely theoretical, that in practice it cannot be made. Fortunately, composers have left ample evidence to the contrary. Consider the act of revision. Revision is an acknowledgment by the composer himself that what he actually wrote was not, on reflection, what he actually meant. Revision is self-criticism. The very word implies that a composer has a *re*-vision of the work, that he returns to the utterance and modifies it in order to be truer to the idea. Revision can sometimes result in criticism on a grand, creative level. Composers occasionally revise the work of other composers with such effective results that the original composition may be eclipsed by the new version, the "criticism," often succeeding where the original work failed. Bach's arrangement for four harpischords (in A minor) of Vivaldi's *Violin Concerto in B Minor*, for example, is more than an adaptation from one medium to another. It is an act of musical criticism *par excellence*. Vivaldi created the idea. But it was left to Bach to give it complete utterance.

Revision as self-criticism

Form and content. Why does music unfold a particular structure? Why that kind of structure rather than another? The textbooks on form remain silent. Yet this is a profound question. It is surely of paramount interest to know why music unrolls in one direction rather than another. Inspired music appears to carry within itself its own blueprint, according to which it propels itself across precise distances and in precise directions. If it is prematurely halted, diverted, or too long continued—all hallmarks of creative immaturity—it loses the sense of punctuality, the feeling of arriving "on time," the knowledge of being in the right place at the right moment, which characterizes each stage of an emerging structure masterfully handled.

The importance of structure

Some musicians have observed that the distinction between form and content is a false one. They rightly argue that no one hears one without the other, that the one is an organic result of the other. Therefore, why not abolish the distinction? They are right in regard to good forms, forms that arise inevitably from the musical material, in which case there is indeed no distinction to abolish. But bad forms, those that are not true to their content, produce a symptomatic division between the inner direction the music was born to follow and the outer direction it was made to follow.

Audibility. The objective of all compositions is to make a total aural impact. There have been some famous miscalculations, intrinsically inaudible passages, which even the most illustrious performance could not render audible. A striking case of inaudibility occurs toward the end of Grieg's *Piano Concerto in A Minor*, in which the "big tune" of the finale returns in full orchestral splendour and obliterates the part of the solo pianist. In the concert hall, it is an extraordinary sight to see the soloist

racing up and down the keyboard, fortissimo, without producing any sound. The observation is beyond all question, and may be checked every time the work is played in public.

In 1937 Schoenberg completed an orchestration of Brahms's *Piano Quartet in G Minor,* Opus 25. As a young man, he had regularly participated in performances of the *Quartet.* Time and again, he was bothered by its intermittent inaudibility: the piano tended to swamp the strings. Schoenberg's orchestration, as he himself claimed, attempted to put matters right. It remains an exercise in musical audibility—one master helping another to communicate. It constitutes an act of criticism on the highest creative level.

Other principles could be formulated to show that a theory of criticism is also a theory of composition. A search for these principles is really a search for the ultimate justification of the feelings of value inspired by great music.

CRITICISM AND PERFORMANCE

Interpretation. Since music does not exist until it is brought to life by the player, two basic requirements are demanded of the critic: a knowledge of the work, and a knowledge of the instrument. Many critics talk loosely about "the work" and "the performance" as though they were separate aspects of musical experience. They are, in fact, different aspects of a total musical experience, and it can be misleading to split them. Much bad criticism results from trying to do so. The Polish-American pianist Leopold Godowsky used to survey his audience before commencing his recitals in an endeavour to discover how many "detectives" there were in the house. Being such a superb pianist he attracted all those critics exclusively interested in keyboard gymnastics. On the other hand, those interested exclusively in the composition may be equally biassed. In this age of authenticity, when the *ur*-text is the thing, many a promising career has been blighted through what, in the profession, is called a "departure from the text." It is not always appreciated that at least a part of the total musical experience is created by the performer, who has a twofold artistic duty: first, to the fundamental character of the work he interprets; and second, to his own artistic conscience, which tells him how the work should unfold. The two are not necessarily mutually exclusive.

The performer's duty

Critics and performers. The criticism of performance is the most public, and publicized, aspect of a critic's function. It is also the least important. Unfortunately, what particular critics think of particular artists accounts for most contemporary music criticism. This is due rather to arbitrary factors than to the critics' sense of priority. Most newspapers insist that a musical event be reported the following day. The critic, consequently, is forced to telephone his review to his newspaper immediately after the concert, limiting himself to a strictly prescribed number of words. Under these conditions, it is not surprising that most criticism consists of predictable accounts of what was played, who played it, and how it was played. Nevertheless, performing artists are obliged to rely on these critical notices if they are to secure further work, even though neither critics nor artists like it. The box-office economics of concert giving are so delicate that bad publicity, or no publicity, can wreck artists and management alike.

THE DEVELOPMENT OF MUSICAL CRITICISM

Beginnings. The criticism of music first gained serious hold in the 17th and 18th centuries. Among the first writer-musicians to make systematic contributions to criticism were Jean-Jacques Rousseau in France, Johann Mattheson in Germany, and Charles Avison and Charles Burney in England. Their work coincided with the emergence of periodicals and newspapers all over Europe. The first journal devoted entirely to music criticism was *Critica Musica,* founded by Johann Mattheson in 1722. Mattheson had a number of successors, notably the Leipzig composer Johann Adolph Scheibe, who brought out his weekly *Der critische Musicus* between the years 1737 and 1740, and whose chief claim to notoriety was his scurrilous attack on Bach. Generally speaking, the criticism of the time was characterized by an obsessive interest in the rules of music and it tended to judge practice in the light of theory—a fatal philosophy. Mattheson, for instance, castigated Bach for ignoring certain rules of word setting in his cantatas.

Mattheson's contribution to musical criticism

At the turn of the century, the age of academicism dissolved into the age of description. Schumann, Liszt, and Berlioz, the leaders of the Romantic era, frequently saw in music the embodiment of some poetic or literary idea. They composed program symphonies, symphonic poems, and lesser pieces bearing such titles as "novellette," "ballade," and "romance." Their literary outlook naturally influenced criticism, the more so as they themselves frequently wrote it. In his pamphlet *On John Field's Nocturnes* (1859) Liszt wrote, in the purple prose of the time, of their "balmy freshness, seeming to exhale copious perfumes; soothing as the slow, measured rocking of a boat or the swinging of a hammock, amid whose smoothly placid oscillations we seem to hear the dying murmur of melting caresses." Most of the Romantics were guilty of this type of descriptive criticism. Its weakness is that unless the music is already known, the criticism is meaningless; and once the music is known, the criticism is redundant, since the music itself says it all far more effectively.

Schumann and Hanslick

The most influential critic of the age was Schumann. In 1834 he founded the periodical *Neue Zeitschrift für Musik* ("New Journal for Music") and remained its editor in chief for ten years. Its pages are full of the most perceptive insights into music and music makers. The first major article Schumann wrote was a laudatory essay on the young Chopin, "Hats off, gentlemen, a genius" (1834), and the last, called "New Paths" (1853), introduced to the world the young Brahms.

During the second half of the 19th century, the critical scene was dominated by the Viennese critic Eduard Hanslick, who is rightly regarded as the father of modern musical criticism. He was a prolific writer, and his book *Vom Musikalisch-Schönen* (1854; *The Beautiful in Music*) is a milestone in the history of criticism. It took an anti-Romantic stand, stressing the autonomy of music and its basic independence of the other arts, and it encouraged a more analytical, less descriptive approach toward criticism. The book was continually reprinted until 1895, appearing in many languages.

Inspired by Hanslick's example, critics in the 20th century rejected the age of description for the age of analysis. Scientific materialism created a climate of rationalism from which music did not remain immune. Critics spoke of "structure," "thematicism," "tonality"—a far cry from Liszt's "dying murmur of melting caresses." A group of musician-thinkers arose who questioned the very basis of musical aesthetics. They included Hugo Riemann, Heinrich Schenker, Hadow, Sir Donald Tovey, Newman, and, above all, Schoenberg, whose theoretical writings show him to be one of the most radical thinkers of the age. Criticism itself was criticized, its basic weakness clearly diagnosed. The search was on to discover the criteria for the evaluation of music. This quest has dominated the work of serious critics ever since. What will become of it is difficult to say, since the language of music itself is now rapidly changing. Under the leadership of such composers as John Cage (1912–) and Karlheinz Stockhausen (1928–), aleatoric techniques have been introduced in which nothing can be predicted, nothing can be repeated, everything is random, and every possible link with the past has been broken. This is a challenge to the theory of criticism that has so far not been taken up.

BIBLIOGRAPHY. The foundation for most of the subsequent developments in musical criticism was laid by SIR HENRY HADOW in his influential essay, "Music and Musical Criticism—A Discourse on Method," *Studies in Modern Music* (1892). This essay cleared the ground for two major inquiries into musical criticism: M.D. CLAVOCORESSI. *The Principles and Methods of Musical Criticism* (1923); and ERNEST NEWMAN, *A Musical Critic's Holiday* (1925). All these books consider the problem of criteria that now dominates the field of musical

æsthetics. Further contributions to the problem have been made by PAUL HINDEMITH in *A Composer's World* (1952); and by ARNOLD SCHOENBERG in *Style and Idea* (1950), containing a key essay, "Criteria for the Evaluation of Music." A later book that challenges the philosophical premises on which criticism is based is ALAN WALKER, *An Anatomy of Musical Criticism* (1966), an attempt to relate æsthetics to psychology. Important supplementary reading exists in the related field of musical analysis. The more significant books include: SIR DONALD TOVEY, *Essays in Musical Analysis,* 6 vol. (1935–39); HEINRICH SCHENKER, *Das Meisterwerk in der Musik,* 3 vol. (1926–29); DERYCK COOKE, *The Language of Music* (1959); and the analytical essay, "Mozart's Chamber Music" by HANS KELLER, in *The Mozart Companion,* ed. by H.C.R. LANDON and DONALD MITCHELL (1956).

(A.W.)

Musical Form

The term used to refer to the structure or shape of a musical composition—that is, the arrangement of individual elements so as to constitute a whole—is musical form. In music the word form is regularly used in two senses: first, to denote a standard type, genre, or species (*e.g.*, symphony, sonata); and, second, to denote the procedures in a particular work. The two meanings are often not kept separate in writing on music. The situation is complicated by the fact that nomenclature for the various musical formal types is determined in different ways—by, for instance, the medium of performance (string quartet, trio sonata), the technique of composition (fugue, variation), or function (prelude, overture, offertory).

PRINCIPLES OF MUSICAL FORM

Music exists in time; as an aesthetician, Susanne K. Langer, put it in *Feeling and Form,* "music is time made audible." The proper perception of a musical work depends in the main on the ability to associate what is happening in the present with what has happened in the past and with what one expects will happen in the future. The frustration or fulfillment, or both, of such expectations and the resulting tensions and releases are basic to the conception of most musical works.

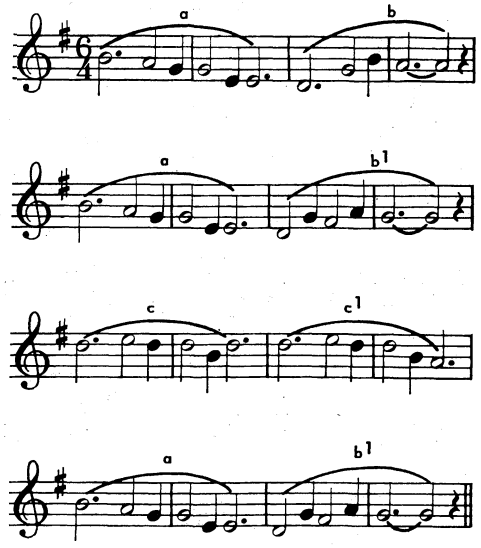
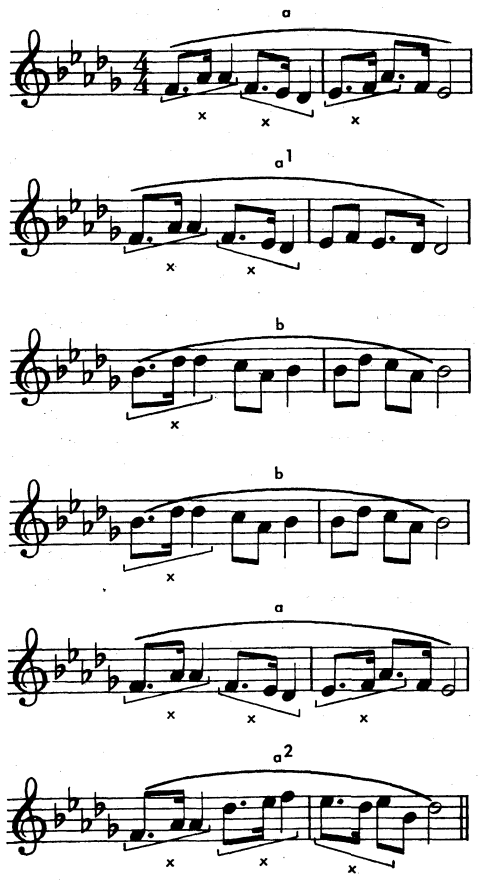
Musical form depends, therefore, on the disposition of certain structural units successively in time. The basic principles can be discerned from a brief consideration of melody, which may be defined as an organized succession of musical tones. This succession of tones consists of component parts, structural units, the principal of which is the phrase—a complete musical utterance, roughly corresponding to what can be sung or played in one breath or played with a single stroke of the bow. A melody, then, ordinarily consists of a succession of phrases, in which there may occur repetition (the same phrase repeated), contrast (a completely different phrase), or variation (the phrase altered, but in such a way that its identity remains perceptible). The relation between these component phrases is important for form. There may, for instance, be a complementary grouping of phrases as antecedent and consequent or "question and answer." The phrases may or may not be equal in length. Some writers, pressing the analogy between music and language, also distinguish larger groupings of phrases: into periods, sentences, paragraphs, and the like. Most musical forms are thus not only additive but also hierarchical: phrases are conjoined to produce a melody, which in turn may be a constituent part of a larger work. A melodic entity that functions as an element in a larger whole is called a theme.

Melody

Coherence may be produced by the use of a motive or figure, *i.e.*, short elements consisting ordinarily of two to four notes. But whereas the motive is usually characterized by a striking interval or rhythmic arrangement, the figure consists of entirely conventional elements (a scale segment, notes of a chord, etc.). Finally, coherence may also be achieved by the consistent use of a rhythmic pattern.

A few examples will serve to illustrate these points. The various phrases have been identified by slurs (phrase marks) and by lowercase letters (the exponential numbers designate variations), whereas larger groupings are designated by capital letters. In the hymn tune "Bethany," by Lowell Mason, shown below, the eight phrases may be grouped in pairs to produce the scheme:

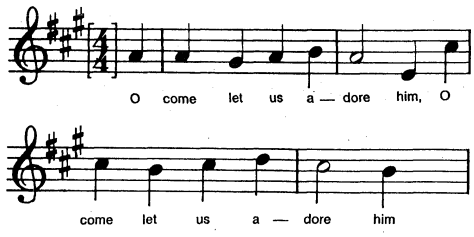
a b a b^1 c c^1 a b^1
A A^1 B A^1

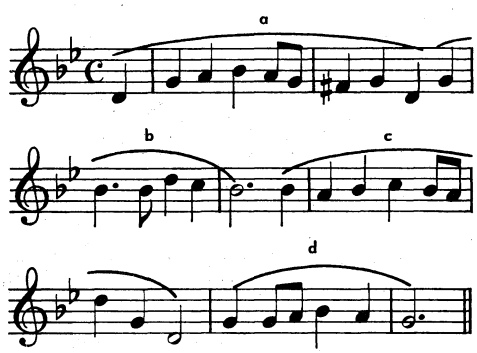
This four-phase arrangement with statement, repetition (here with variation at the cadence), contrast, and restatement is extremely common in the traditional and art music of the West.

Other schemes based on repetition and contrast abound. The famous "Largo" from Antonín Dvořák's *New World Symphony,* Opus 95, goes like this:

Here the scheme is:

a a^1 b b a a^2
A B A^1

This common arrangement is known as three-part form. Coherence is provided by the use of a rhythmic motive, (marked "x" in the second example).

A common device is melodic sequence, in which the phrase is repeated but in transposition (at a different pitch level), as in the refrain of the Christmas carol "O Come All Ye Faithful."



Different from the above procedures is an English traditional song, "How Should I Your True Love Know":

The phrase structure here is a b c d, so that there is no repetitive plan. Such a melody is said to be through-composed. In some through-composed melodies a rhythmic pattern may appear throughout to promote coherence.

Function of pitch

Other elements contribute to formal organization in music. Among those having solely to do with pitch are range or register—whether most of the activity is high, low, or in the middle, or combinations of these, and whether the range of pitches used is large or small; types of melodic motion, whether conjunct (*i.e.*, by step along the scale) or disjunct (by leaps); and the use of different types of scales or modes. Factors included in music's temporal aspect include tempo, whether fast or slow, as well as duration; *i.e.*, whether individual notes are long or short (the gradually increasing use of constantly shorter note values, for instance, is associated with acceleration and accumulation, thus with increasing intensity). Among the harmonic aspects, there is key, or tonality (set of interrelated notes and chords, based on a major or minor scale), whereby the reassertion of a key following the intervention of other keys may produce an effect akin to the restatement of a phrase after a contrasting one has been heard; in this respect, cadences (sections giving the impression of conclusion) are of crucial importance in defining key. Still other factors include the use of dynamics (loud and soft); timbre, or tone colour, especially in the employment of unusual instruments or combinations of instruments; texture, whether monophonic (consisting of a single melodic line) or polyphonic (many-voiced), be it contrapuntal (having simultaneous independent melodic lines) or homophonic (one voice leading melodically, supported by chordal procedures); and, in vocal music, whether the text is set syllabically (one note to a syllable) or melismatically (many notes to a syllable).

FORMAL TYPES

Four basic types of musical forms are distinguished in ethnomusicology: iterative, the same phrase repeated over and over; reverting, with the restatement of a phrase after a contrasting one; strophic, a larger melodic entity repeated over and over to different strophes (stanzas) of a poetic text; and progressive, in which new melodic material is continuously presented (thus synonymous with through-composed). The following discussion deals first with Western and then with non-Western music.

Iterative and reverting types. Iterative types, not common in Western music, may be found in the recitation tones of Gregorian chant, in which, for example, each line of a psalm is sung to the same melodic formula. Far more common, however, are reverting types. In the Middle Ages there existed the fixed forms used in songs, such as the French ballade (a a b), virelai (A b b a A), and rondeau (A B a A a b A B), the Italian ballata (A b b a A) and the German bar form (a a b), where the patterns of repetition and contrast correspond to poetic forms. (In the representations of the reverting types in songs, lowercase letters refer to the same music set to different words, while capital letters indicate that both text and music are the same.) Since the Baroque period (*c.* 1600–*c.* 1750) there has been binary, or two-part form, such as a b. A variety of binary form particularly prominent in the dances of the 18th century is the rounded binary form, the two sections of which are a and b a (*i.e.*, with a final return to original material in the second section), each of which is repeated, part one being heard twice before part two begins: ||: a :|| ||: b a :|| (||: and :|| indicate the enclosed material is to be repeated).

Binary. The rounded binary form took on great importance in the late 18th century, when it was expanded and elaborated into what is known as sonata form (also called sonata-allegro or first movement form), which may be represented thus: ||: exposition :|| ||: development recapitulation :||, whereby the kinship to the ||: a :|| ||:b a:|| structure of rounded binary form is clear. Ordinarily, in the exposition the principal musical themes are stated; in the development they are subjected to a process of working out and variation; and, finally, in the recapitulation they are restated. Sometimes the scheme is enlarged by adding a slow introduction before the exposition or a coda (concluding passage) at the end, or both. This formal principle, usually treated with a certain freedom, has been of basic importance in Western instrumental music since the mid-18th century.

Sonata form

Ternary. Another basic reverting type is ternary (three-part) form, a b a, also known as "song form" because of its frequent use in that genre, as well as in character pieces for piano. The form dominates the aria in late Baroque opera (the da capo aria, in which the final statement of a is not written out, but the performers simply follow the written instruction da capo, meaning "from the beginning," and repeat the first part). The da capo principle also appears in the instrumental minuet and scherzo with trio.

Rondo. On a larger scale are refrain schemes, in which contrasting episodes appear between statements of the refrain. In instrumental music this is found most often in a five-part arrangement, the rondo, frequently a b a c a b a; but many departures from the form occur, most common being the replacement of c by a development passage, usually based on the rondo theme. This important variant, known as the sonata-rondo, is particularly associated with Joseph Haydn. The refrain principle also appears in the rondeau of 18th-century French harpsichord music, in which there is no set limit to the number of episodes. The third movements of concertos, with the reversions of the tutti or ritornello (passage for full orchestra) and the intervening episodes for the solo instrument or instruments, are also of this type, as occasionally are large operatic arias.

Strophic types. The strophic type is seen in hymns and traditional ballads, in which different poetic strophes are set to the same melody. Thus, while the melody of a single stanza may accord with one of the reverting types, the hymn or ballad as a whole is strophic; this also applies to the fixed forms of medieval music and to many other types of song, simple and complex.

Theme and variation

The instrumental equivalent of the strophic type is variation (or theme and variation) form, in which a musical theme, often a complete melody with a harmonic accom-

paniment, is stated and then repeated a number of times, but with variations. A clear example of the relation between variation and strophic form is the chorale-partita of the Baroque era, a keyboard piece based on a hymn, with each varied statement of the hymn tune corresponding to a strophe of the hymn text. But the structure is more common in independent instrumental compositions, often of considerable dimensions (*e.g.*, Beethoven's *Diabelli Variations* for piano). In the Baroque a common type was the ostinato, or variations on a ground, in which the composition was built on a recurring melodic or harmonic pattern, generally in the bass, the accompanying parts being varied with each statement of the pattern, as in Bach's *Passacaglia and Fugue in C Minor* for organ or his "Chaconne" from the *Partita in D Minor* for unaccompanied violin. This procedure is also found in early operatic arias in the strophic variation form, in which each statement of the ostinato corresponds to a strophe of the aria's text. In the 19th century Brahms made impressive use of the ostinato (finales of the *Variations on a Theme by Haydn* and the *Symphony No. 4 in E Minor*).

Progressive types. The progressive type is common in songs and instrumental pieces of the 19th and 20th centuries but is also found in earlier music (*e.g.*, in the melodies used for the Gloria and Credo of the mass in Gregorian chant) and in the prose, or sequence (*c.* 9th–*c.* 12th centuries), the phrases of which are arranged in pairs (a a b b c c d d, etc.), and its instrumental equivalent, the estampie. Polyphonic forms using a cantus firmus or basic melody (often a plainchant excerpt) also belong to the progressive type and include the liturgical organum, the early motet, and the conductus from the medieval era, as well as many chorale-preludes for organ of the Baroque. If, however, the cantus firmus itself is in one of the reverting forms, then the polyphonic setting will frequently follow suit.

The most important forms of Renaissance polyphony also belong to the progressive type, since the characteristic procedure was to give each line of the text its own musical phrase, as in the Renaissance motet and other types of secular polyphonic music. The same applies to the instrumental contrapuntal forms of the late Renaissance and Baroque: the ricercar, canzona, invention, and fugue. Other progressive types include intonations, preludes, toccatas and fantasias for lute and keyboard of the 16th, 17th and early 18th centuries, in which the thematic material consists primarily of figurative elements (scale passages, arpeggiated chords, trills, turns, and the like); in larger works of this kind—by Bach for instance—passages in fugal style are often also present. Finally, there is simple binary form (a b), often found in early dances and in large operatic arias of the Classical period (Mozart and Beethoven).

WESTERN COMPOUND FORMS

Early history. With the larger forms of instrumental music there are extended musical pieces, usually called movements, which in their succession and totality make up a larger whole. An important unifying factor is key: a single key often dominates the work, others being used for contrast. This idea goes back at least to the Baroque, when two formal types were established: the first is the *sonata da camera,* or chamber sonata, consisting of a series of dances in the same key (also known as partita and, later, suite). By J.S. Bach's time (1685–1750) a set arrangement of dances was common: allemande (moderate slow triple time), courante or corrente (fast triple time), saraband (slow triple time), and gigue (fast triple time), usually with some other dance inserted between the saraband and the gigue. The other type is the *sonata da chiesa,* or church sonata, which consists of four movements, again all in the same key, in a slow–fast–slow–fast succession. The distinction between the two types is explicit in Bach's sonatas and partitas for unaccompanied violin: the sonatas are in the "church" form, whereas the partitas are suites. Other large forms of Baroque music are the two types of overture: the French, in two parts, the first slow and stately, the second fast and fugal; and the Italian, in three movements in the succession fast–slow–fast, the middle movement usually in a different key. The instrumental concerto after 1700 usually employed the same scheme as the Italian overture.

The sonata. Around 1750 a combination of these types produced the forms most common in the standard repertory of concert music. The sonata became a piece for either a keyboard instrument or a solo instrument accompanied by keyboard. It ordinarily consisted either of three movements in the arrangement fast–slow–fast or of four, with a minuet inserted between the slow movement and the finale; but there are examples of two-movement sonatas, notably by Beethoven, and even of one-movement sonatas (*e.g.*, by Domenico Scarlatti, the Italian-born composer to the Spanish court). Usually all movements except the slow one (and sometimes the trio, as well) are in the same key. The first movement typically is cast in the sonata form, the slow movement in one of the reverting schemes (often ternary), and the finale either in sonata or rondo form; but variation form may appear in any of these movements.

This large form is also used in chamber music, particularly the string quartet, and in the large form of orchestral music, the symphony, both of which ordinarily have four movements. Notable exceptions to this are the late quartets of Beethoven as well as those of the 20th-century composer Béla Bartók, the latter in two instances using what is called the "arch form," a large reverting arrangement, A B C B A, each element being a separate movement; needless to say, there are innumerable other exceptions. The concerto, on the other hand, adheres more to the older three-movement form. The various kinds of late 18th-century entertainment music (cassation, divertimento, nocturne, serenade, and the like) may employ any of a number of arrangements, ranging from three movements all the way to six or more.

Exceptions to sonata form

Cyclic form. Some authorities believe that from the 18th century on such sonata-form compositions have been organized by the use of a few musical thematic motives, often submitted to considerable variation throughout. Compositions organized in such a way are said to be in cyclic form. While this becomes important in the 19th century, the extent to which it characterizes the Classical period is a matter of some controversy at present.

Fantasias and program music. Simultaneously a much freer form was cultivated, beginning in the late 18th century, the fantasia, primarily for keyboard, notably in the hands of Carl Philipp Emanuel Bach. Consisting of an indefinite number of highly contrasting sections, surprise and expression were of prime significance.

The fantasia, along with the overture to a play or opera, was the precursor of the large forms of orchestral program music of the 19th century, in which an extramusical content (usually a narrative of some kind), called the program, is expressed in the composition. There are two main types: the program symphony, associated with Hector Berlioz, in which the norms of symphonic form are for the most part preserved, and the symphonic poem, associated with Franz Liszt and Richard Strauss, in which the composer allows the extramusical subject matter to determine the structure of the composition. Some 19th-century concert overtures by German composers such as Felix Mendelssohn and Robert Schumann belong to this type of composition. Important here is the association of musical themes with aspects of the program, the themes being used throughout the work, often in varied forms.

The symphonic poem

Another arrangement is called the suite, which no longer consists exclusively of dances but also of instrumental pieces of all kinds. Usually some common element runs throughout: a cyclic theme may be used, as in Schumann's *Carnaval* or the Russian composer Modest Mussorgsky's *Pictures from an Exhibition,* or the music may originally have been intended for use with a play (Mendelssohn's music to *A Midsummer Night's Dream* or the Norwegian Edvard Grieg's for Henrik Ibsen's *Peer Gynt*).

Opera and oratorio. Among the large forms of vocal music, opera and oratorio are the most significant. Both are extended works in which a narrative is set to music.

While an opera is performed in a theatre, an oratorio is a concert piece. Both may be either sacred or secular. A special type of oratorio is the Passion, the setting of New Testament accounts of Christ's crucifixion. The cantata may be regarded as a smaller form of oratorio.

Operas and oratorios ordinarily consist of several musical genres: recitative (imitating the manner of speech), aria, ensemble and choral pieces, often with instrumental interludes and an overture (most overtures of the late 18th and 19th centuries being cast in sonata form). Opera often includes ballets and large sectional finales at the ends of acts. With respect to the oratorio, Handel greatly increased the role of the chorus in his work with this genre (especially *Israel in Egypt*), an example seized upon by his successors. Oratorios also differ from operas in that they frequently make use of a *testo* (narrator), who relates the events of the action, usually set in recitative style. Stravinsky's *Oedipus Rex* (1927) combines the two traditions.

Wagner's new operatic form

Whereas operas are usually composed as a series of enclosed musical forms, the German composer Richard Wagner devised a special kind, known as music drama, in which the music is continuous and in which the distinction between recitative, aria, and ensemble is largely eliminated. Instead, Wagner used a flexible melodic line which he referred to as "tone speech." Wagner also greatly increased the role of the orchestra, stressing the technique of thematic development and transformation borrowed from instrumental music and further associating each theme with an aspect of the operatic plot, such themes being known as leitmotivs ("leading motifs").

The mass. Another large vocal form is the mass, the earliest polyphonic settings of which date from the 14th century. At first the mass was set in cantus firmus style, each movement built on the appropriate Gregorian chant melody, as in the mass of the French composer Guillaume de Machaut. In the 15th century a Burgundian composer, Guillaume Dufay, and his contemporaries developed the cyclic mass, in which a single cantus firmus was employed throughout. This idea was extended in the parody mass, built by elaborating thematic material taken from an existing polyphonic work, usually a motet or chanson; most 16th-century masses are of this kind. In the Baroque mass, each segment of the text is treated as an independent composition (aria, duet, chorus), similar to the procedure in a cantata or oratorio, except that no recitatives are used. J.S. Bach's *Mass in B Minor* is of this type.

In the 16th and 17th centuries, Monteverdi and others grouped madrigals into a kind of cycle around a particular subject; should a dramatic text be involved, the form is known as a madrigal comedy. In the 19th century and after, similarly grouped songs with piano accompaniment are known as a song cycle (*e.g.*, those by Beethoven, Schubert, and Schumann).

Twentieth-century modifications. In the 20th century many composers have continued to use the traditional forms, although in other respects their work makes significant departures from what had been established. Others have virtually discarded them. Radically new elements have been introduced to serve as structural units: instead of the traditional phrases and motives, composers have come to employ what they call "sound events," combinations of textures, types of timbre, aggregates of different simultaneously sounding pitches, and the like. An early example is afforded by the notion of the melody of timbres—the *Klangfarbenmelodie* of Arnold Schoenberg and Anton von Webern. The broadest possible range is included in the conception of a sound event: not merely tones but also noises; in fact, sounds of all kinds. These sound events have been arranged in a complex manner. An important principle of organization has been serialism, originally applied only to pitch, so that the pitch content of the work is decided by reiterations of a series (set or row) of pitches that has been determined in advance (*e.g.*, in the twelve-tone compositions of Schoenberg and Webern). Since 1945 the serial principle has been extended to other aspects of music as well: durations, dynamics, and types of attack and tempi, as in the music of the French composer Olivier Messiaen and the German Karlheinz Stockhausen. Other composers, including Stockhausen and the Greek Yannis Xenakis, have used numerical relationships to determine the structure of compositions mathematically. Still others have conceived of form as a dynamic interaction between contrasting and continually evolving sound events—the French composer Edgard Varèse and the American Elliott Carter. Simultaneously, there are works, such as those of the American John Cage, in which the form is not predetermined by the composer but left to chance, such pieces being called indeterminate or aleatoric. These statements apply to electronic music as well.

"Sound events"

NON-WESTERN FORMS

Elements of structure. In many non-Western civilized cultures, such as those of India and Near Eastern Islāmic countries, music for the most part is not written down in advance of the performance but improvised upon framework-like patterns. In effect a composition exists only in its performance. Problems are presented by the different scales and intervals, rhythmic patterns, timbres, inflections, and the like. Or, again, a type of polyphony known as heterophony may result from the discrepancies that occur when several different singers or players simultaneously vary the same melodic line. In Japanese *gagaku* music, in which improvisation is not a factor, the beats often vary in duration, producing what is known as "elastic rhythms," akin to the Western tempo rubato. Much non-Western music is functional: each piece has a specific purpose and is associated with certain events, actions, or situations, which often determine the musical structure.

Important elements of form in non-Western music are melody types and rhythmic patterns. By melody type is meant a melodic formula using a recognized scale, stressing particular tones and using specific configurations of intervals, which can provide the basis for a larger piece. In a given society or culture there ordinarily exists a large repertory of these melody types from which the musician, following convention, selects those for use in his rendition. The use of rhythmic patterns for form is seen, for instance, in colotomic structure, in which the composition is marked off in temporal entities by the differing but regularly recurring entrances of particular musical instruments in a set order.

Rhythmic patterns used for form

In the music of nonliterate cultures, simple iterative and strophic structures predominate. These are characterized by short phrases and the frequent use of alternation between a soloist and the group (the "call-and-response" pattern). Such pieces, often associated with dancing, usually belong to specific rituals of one kind or another.

In literate cultures comprehensive theoretical music systems often exist, with elaborate musical forms. Nonetheless, the conception of propriety, that each musical piece has a certain function, is dominant. While strophic forms continue to be important, as in the Indian *gat,* the *umui* religious chants of the Ryukyu Islands (both of which also use variation), and in the popular music of most cultures, there are also such large musical forms as those related to the Indian *rāgas*, the Arabic *maqāmāt* (melody types), and the music of the *gamelan* (gong and percussion ensembles) of Java and Bali.

Specific formal patterns. The term *rāga,* meaning colour or passion, refers not only to a scale but also to the melody type. It has given rise to several musical forms, among them the ancient northern *dhurpad* and the shorter type known as *khyāl* in the North and *kīrtana* in the South. These may be vocal, instrumental, or both, the main performer or performers being accompanied by a drone played by a bass lute and by the reiteration of an extended rhythmic pattern (the *tāla*) by a drum. While much variety is possible, a typical formal arrangement includes a prelude which states the *rāga* (scale and melody type) in its simplest form and continues with a number of contrasting sections that differ in the degree of elaboration of the *rāga*, the register exploited, and the rapidity of the notes and figuration, usually reaching a climax near the end, in which daring variations on the *rāga* in the uppermost register are played or sung in very

rapid tempo; the performance may conclude with a return to the character of the prelude. Interludes or linking passages, related to the prelude, are often present, so that an irregular reverting scheme results.

The Arabic *maqām*, roughly equivalent to the *rāga*, provides the basis, among others, for the *bashraf*, a reverting type not unlike the Western rondo, and the *nawba*, an extended multisectional performance with some resemblance to the suite. In Java and Bali particularly noteworthy is the *gending* (musical composition) for the *gamelan*, which may take one of several progressive sectional forms in which the colotomic principle is important: the large *gending agen*, usually bi-partite, a slow movement followed by a fast one, and the shorter *gending gangsaran*, used as preludes and interludes.

BIBLIOGRAPHY. Comprehensive treatments of form in Western music are HUGO LEICHTENTRITT, *Musikalische Formenlehre*, 3rd ed. (1927; Eng. trans., *Musical Form*, 1951); and R.E. TYNDALL, *Musical Form* (1964); a shorter treatment in GROSVENOR COOPER, *Learning to Listen* (1957). For form in traditional and non-Western music, see BRUNO NETTL, *Folk and Traditional Music of the Western Continents* (1965); and *Music in Primitive Culture* (1956); and WILLIAM P. MALM, *Music Cultures of the Pacific, the Near East and Asia* (1967). See also, generally, the relevant articles in WILLI APEL, *Harvard Dictionary of Music*, 2nd ed. rev. (1969).

(F.E.Ki.)

Musical Instruments

A device that produces musical sounds is a musical instrument, whether it is used for ritual or ceremonial purposes, for entertainment, or for private enjoyment. The fact that a number of instruments produce sounds of no definite pitch does not exclude them from this definition, since rhythm is an essential element in all music. The beat of a drum can be as important in a composition as a melody played by violins. Many instruments seem to have developed from the imaginative use of objects originally designed for other purposes. The drum has its origin in the pot and the pitcher; the rattle was used to mark the steps of a dance and also as a charm against evil spirits. Composers of recent times have used other objects not normally associated with music. Stage properties such as the wind machine and the thunder machine have been called for in works of a dramatic character. Iron chains, typewriters, and milk bottles are not unknown in 20th-century music. None of these, however, has established itself in normal use; and to extend the definition of musical instruments to include such objects would involve admitting any solid body that can make a sound when struck.

The material of which an instrument is made does not necessarily affect the sound that it produces. In the case of instruments that are struck with some kind of beater, the difference may be very marked. Wood and metal when struck produce quite different sounds. On the other hand, a flute made of metal does not produce a substantially different sound from one made of wood, as in this case the vibrations are in the column of air in the instrument. Oboes have been made of ivory, clarinets of ebonite. The characteristic tone of wind instruments depends on other factors, notably the length and shape of the tube and the means by which the air column is set in motion. The length of the tube not only determines the pitch but also affects the tone: the piccolo, being half the size of the flute, has a shriller tone; the bass clarinet, being larger than the clarinet, produces a darker sound. The shape of the tube determines the presence or absence of the "upper partials" (harmonic or nonharmonic overtones), which give colour to the single note. In the oboe the tube, or bore, is conical, expanding gradually from the reed mouthpiece to the bell; this means that a strong and uninterrupted series of upper partials is present, and this and the fact that the air column is set in motion by the vibrations of the reed give the instrument its characteristic tone. In the clarinet, on the other hand, the bore is cylindrical, and the tube is stopped at one end by the mouthpiece, with the result that up to a point only the odd-numbered upper partials are present. Hence, the clarinet has a mellower, less reedy tone than the oboe.

The volume of sound produced by an instrument depends on resonance, whether created by a vibrating air column or by a solid body vibrating in sympathy. In bowed string instruments the hollow box over which the strings are stretched vibrates in sympathy when the bow is drawn across the strings. Similarly, the volume of tone produced on the piano depends on the vibrations of the soundboard, without which the mere impact of the hammers on the strings would produce very little effect.

Cultural aspects. The use of instruments purely for entertainment is a development of civilized societies. In primitive societies instruments were used as an adjunct to dancing and labour, and they also had magical properties that could serve in religion and medicine. These functions naturally overlapped. Bells or flutes could be used to attract the opposite sex and also to ward off evil spirits. It is also significant that most instruments were assigned to men. As an element in magic, instruments could be used equally to repel evil or to invoke a deity. The use of instruments for religious ceremonies has continued down to the present day, though at various times they have been suspect because of their secular associations. The many references to instruments in the Old Testament are evidence of the fact that they played an important part in Jewish worship until for doctrinal reasons they were excluded. It is also clear that the early Christians in the eastern Mediterranean used instruments in their services, since the practice was severely condemned by ecclesiastics, who insisted that the references to instruments in the Psalms were to be interpreted symbolically. As a result of this official attitude the music of the Christian Church became for several centuries restricted to voices. The organ, which seems to have been developed in Alexandria in the 3rd century BC, was a purely secular instrument in the Greco-Roman world. It remained so at Byzantium after the official acceptance of Christianity: it was used for ceremonial occasions such as banquets and chariot races but not for church services. The use of the organ in church in western Europe appears to date from the 8th century AD, when the Frankish monarch Pepin III installed at Compiègne an instrument he had received as a present from Byzantium.

The invention of the organ

Belief in the magical properties of instruments is found in many societies. The Jewish shofar (a ram's horn), which is still blown on Rosh Hashana (New Year) and Yom Kippur (Day of Atonement), must not be seen by the congregation. Similar secrecy is found among primitive peoples. The power of the shofar is illustrated by the story of Joshua at the siege of Jericho: when the priests blew their shofars seven times the walls of the city fell flat. In India, according to legend, when the deity Kṛṣṇa played the flute, the rivers stopped flowing and the birds came down to listen. The birds are said to have done the same in 14th-century Italy when the composer Francesco Landini played his organetto, or portative organ. In China, instruments were identified with the points of the compass, with the seasons, and with natural phenomena. The Melanesian bamboo flute was a charm for rebirth. Many primitive instruments were phallic symbols.

Most of the instruments used in the West in medieval times came from the Near East. Trumpets, long associated with military operations, had a ceremonial function in the establishment of kings and nobles and were, in fact, regarded as a sign of nobility. In the later Middle Ages and for long afterward, they were associated with kettledrums (known originally as nakers, after their Arab name, *naqqārah*), which were often played on horseback, as they still are in some mounted regiments. Trumpet fanfares, heard on ceremonial occasions in the modern world, are a survival of medieval practice. In the 17th century, trumpets and kettledrums were introduced into opera and orchestral music of a festal character. In the same way, horns, originally hunting instruments, became normal members of the orchestra in the 18th century.

Technological developments. The earliest instruments appear to have been natural objects capable of producing sound; for example, the conch (marine shell) and the bone pipe. The conch and the primitive trumpet at first served only as a means of amplifying the human voice.

The discovery that tubes possessed acoustic properties that enabled them, by vibration of the lips, to sound more than one note was a later development. From using natural objects, it was an obvious step to construct instruments that would serve a similar purpose, using wood or terra-cotta or any other available material.

Development of string instruments

Another development of natural resources was the discovery that a stretched string could produce a musical note. This resulted in the construction of the musical bow, the ancestor of all plucked string instruments. Its shape is similar to that of the hunter's bow, but it seems to have had an independent origin. Since the sound of the simple string is feeble, it is reinforced by means of a resonator—either a pot placed on the ground beneath the instrument or a gourd attached to it or the player's mouth. The string can either be struck with a stick or be plucked with the fingers. In the former case it could be regarded as the forerunner of the Persian dulcimer, in which numerous strings are struck with a stick, and ultimately of the piano.

The construction of all but the simplest instruments requires skill and craftsmanship: the piercing of a tube to a uniform or expanding width, the flaring of the bell of a wind instrument to increase sonority, the measurement of the bars of a xylophone or glockenspiel, the curvature of the back of a lute, the internal and external structure of the body of a violin. All of these involve accurate workmanship from experts in wood and metal and, from present-day designers, a knowledge of the mathematics of sound. The latter is particularly true in the case of woodwind instruments, where the holes must be accurately pierced in order to produce a scale that is in tune. Of all instruments, the organ demands the fullest cooperation from workers in different skills: in the construction of pipes in metal and wood, in fashioning them to the right length, in securing uniformity of tone throughout a single row, and in adjusting the great variety of sounds so that they blend happily together; in the provision of leather pallets to allow the passage of air; in providing direct and immediate contact between the keyboard and the pipes; and in constructing a blowing apparatus that will be adequate for the pressure required.

Woods used for making instruments

The first step in the building of any instrument is the selection of material. Wood, whether cocus, rosewood, or maple (to mention three that are favoured for woodwind instruments) or pine, used for string instruments, needs to be seasoned. This applies equally to the reeds of oboes, clarinets, and bassoons, which are cut from the plant *Arundo donax,* now grown for this purpose, particularly in the south of France. Climate is also an important consideration. The moist, damp climate of tropical countries has often proved disastrous to instruments built for use in Europe.

Our knowledge of medieval instruments is derived almost entirely from pictorial illustrations, and the men who made them remain anonymous. Their skill, like that of the cabinetmaker and the silversmith, was developed by long practice, and the principles that determine both tone and intonation were discovered by trial and error. The growth of instrumental playing in the 16th century stimulated the production of instruments to be used not merely for ceremonial and official entertainment but also for social occasions and private pleasure. From this time we begin to know the names of makers, many of whom established family businesses that lasted for several generations. These include, for example, Andrea Amati (16th century), violin maker in Cremona; Hans Ruckers (late 16th century), harpsichord maker in Antwerp; and Johann Haas (1649–1723), trumpet maker in Nürnberg.

Modern technology has in many cases simplified or improved the construction of instruments. In the past, for example, the tubes of horns and trumpets were made from a sheet of brass cut to the right width, which was rolled into shape, leaving the edges to be joined by brazing. In modern manufacture the tube is drawn in one piece, and there is no seam. Evolution of design has been particularly notable in the construction of wind instruments: in the case of woodwind instruments, not only in the fixing of the metal keys and the mechanism that controls them but also in the piercing of holes in such a way that they are both acoustically correct and under the direct control of the fingers. This achievement, which dates from the first half of the 19th century, was due mainly to the pioneering work of Theobald Böhm (1794–1881). His system, designed for the flute, was also applied to the clarinet, and some of its mechanical principles were adapted for the oboe and bassoon. The early 19th century also saw a revolution in the manufacture of brass instruments: the addition of pieces of tubing of different lengths through which the air could be directed by the depression of valves (or pistons), enabling an instrument to produce all 12 notes of the chromatic scale, in place of its earlier limitation to the notes of the natural harmonic (overtone) series.

Modern changes in design

The discovery of plastics in the 20th century has also influenced the manufacturers of instruments; for example, plastic has been used instead of quill or leather for the plectra that pluck the strings of the harpsichord, and plastic recorders have been built. Mechanization has made possible the mass production of instruments of all kinds. Insofar as this makes it possible for people to acquire an instrument at a moderate price, mass production is a good thing, and in education it has been beneficial to schools working on a small budget. A natural development has been the provision of kits consisting of the separate parts of an instrument, which can then be assembled by the purchaser. It remains true, however, that the production of an instrument of the finest quality still demands the highest degree of individual skill. A violin glued together from mass-produced parts cannot be the equal of one that has been meticulously constructed from the first by an individual craftsman who will not be satisfied with his work until every detail of it has been tested.

Importance of craftsmen

Technology has been of service to music by providing composers with instruments they have asked for, by eliminating the defects of instruments and facilitating the work of performers, and by making instruments more widely available to the community in general. Richard Wagner suggested the heckelphone, a double-reed instrument akin to the oboe, and ordered a special type of tuba for use in his music drama tetralogy *Der Ring des Nibelungen.* Not all the improvements in construction have been wholly advantageous. It is easier to play in tune on a modern woodwind instrument, but the older examples, being less cluttered with metal fittings, had a purer tone. Similarly, the modern horn, which is virtually two instruments in one, is inferior in tone quality to its 18th-century predecessor. Only the trombone and the string instruments have remained, apart from minor modifications, unchanged in structure over the centuries, though the substitution of wire strings for gut has materially altered the tone of the violin.

The older makers of instruments were craftsmen who took delight in the appearance of their work. Pictures were painted on the inside of harpsichord lids; delicate inlay and carved heads adorned string instruments. Rings of ivory concealed the joints in woodwind instruments; the bells of trumpets were richly chased. Sometimes there was extravagance—for instance, making the bell of a wind instrument in the shape of a dragon's head. Modern taste in most cases rejects ornamentation. The coats of arms painted on military bass drums are there principally to indicate their regimental function. But overall design and finish are as important as they have always been. If instruments were not meant for use, those that are produced today would deserve, on aesthetic grounds, to find a place in museums, side by side with the musical treasures of the past that lead a beautiful but generally silent existence there.

Classification of instruments. Instruments have been classified in various ways, some of which overlap. The Chinese divide them according to the material of which they are made—as stone, wood, silk, and metal. Although applicable to Chinese instruments, this system will not do as a classification of instruments in general. To take a simple example, a saxophone is made of metal, but its method of tone production associates it with the clarinet,

which is, at least in origin, a wooden instrument. Writers in the Greco-Roman world distinguished three main types of instruments: wind, string, and percussion. This classification was retained in the Middle Ages and persisted for several centuries: it is the one preferred by some writers, with the addition of electronic instruments, at the present day. Some 16th-century writers excluded certain instruments from this classification; the musical theorist Ludovico Zacconi (1555–1627) went so far as to exclude all percussion instruments and established a fourfold division of his own—wind, keyed, bowed, and plucked. A different fourfold classification was accepted by the Hindus at least as early as the 1st century BC: they recognized string instruments, wind instruments, percussion instruments of wood or metal, and percussion instruments with skin heads (that is to say, drums). This ancient system—based on the material producing sound—was adopted by the Belgian instrument maker and acoustician Victor-Charles Mahillon (1841–1924), who named his four main classes autophones, or instruments made of a sonorous material that vibrates to produce sound (*e.g.* bells, rattles); membranophones, in which a stretched skin is caused to vibrate (*e.g.* drums); aerophones, in which the sound is produced by a vibrating column of air (wind instruments); and chordophones, or string instruments. In their highly influential studies of musical instruments, the Austrian musicologist Erich von Hornbostel (1877–1935) and his German colleague Curt Sachs (1881–1959) accepted and expanded Mahillon's basic division, creating the classification now used in most systematic studies of instruments. The name idiophones was substituted for autophones, and each class was subdivided according to a method similar to that used by botanists. A fifth class, electrophones, in which vibration is produced by oscillating electric circuits, was added later. The Table gives examples of instruments that come within the various categories.

Systems based on the material sounded

Classification and Examples of Some Musical Instruments

Idiophones	
Struck	against each other: cymbals, castanets
	with a beater: triangle, glockenspiel, xylophone, slit drums
Shaken	rattle, jingles
Scraped	scraper
Plucked	jew's harp, music box
Rubbed	musical glasses
Membranophones	
Struck	tambourine, side drum, bass drum, timpani
Rubbed	friction drum
Blown	mirliton (or kazoo)
Chordophones	
Plucked	without keyboard: harp, lute, violin, zither, lyre
	with keyboard: harpsichord
Struck	without keyboard: dulcimer
	with keyboard: clavichord, piano
Rubbed (with a bow)	viol and violin families
Aerophones	
Flues	without keyboard: recorder, flute
	with keyboard: organ
Reeds	without keyboard: single reed—clarinet, saxophone
	double reed—shawm, crumhorn, oboe, bassoon, sarrusophone
	with keyboard: harmonium, American organ (melodion)
Lipped	horn, cornet, trumpet, trombone, tuba, serpent
Electrophones	
Monophonic	producing a single line of melody: theremin, trautonium, Ondes Martenot
Polyphonic	producing harmony and simultaneous melodic lines: Hammond organ

The classification shown in the Table is convenient, but it is not completely logical. Some instruments may belong to more than one class or more than one subdivision. The tambourine is a membranophone in so far as it has a skin head which is struck; but, if it is only shaken so that its jingles sound, it should be classed as an idiophone, for in this operation the skin head is irrelevant. Flue pipes are the foundation of organ tone, but the instrument also has a number of reed pipes, so that the organ belongs equally to the first and second order of aerophones. The violin is undoubtedly a chordophone, but if the strings are struck with the wooden part of the bow it belongs to the second order of this class, not the third. Modern composers not infrequently require players to slap their instruments instead of playing them in the normal way. A slapped guitar or a slapped trombone is technically an idiophone. The classification according to the source of vibrations also ignores common elements in means of production. It is normal practice to speak of keyboard instruments as a group, including piano, harpsichord, organ, and so on; but in the Hornbostel–Sachs classification these are either chordophones or aerophones, and, if the keyed glockenspiel is included, there is an idiophone as well.

History and evolution. Men have speculated about the origin of instruments since antiquity. Older writers were generally content to rely on Greek mythology or Old Testament legends. In the 19th century there was an attempt at a more scientific approach, partly as a result of theories of evolution put forward by Darwin and Herbert Spencer. The British writer John Frederick Rowbotham argued that there was originally a drum stage, followed by a pipe stage, and finally a lyre stage. The Austrian writer Richard Wallaschek, on the other hand, maintained that, although rhythm was the primal element, the pipe came first, followed by song, and the drum last. Sachs, instead of relying on anthropological evidence, based his chronology on archaeological excavation and the geographical distribution of the instruments found in them. Following this method, he established three main strata. The first, of wide distribution, consists mainly of noisemakers, such as the rattle and the bull-roarer. The second, found in several continents but not universally, includes drums, trumpets, and plucked strings. The third, occurring only in certain areas, includes more advanced instruments, such as the xylophone and the flute. That percussion instruments should appear at an early stage is not surprising. From stamping out a rhythm with his feet, primitive man learned to do the same with an implement: these implements became more elaborate as time went on and were followed by further experiments in extracting sound from sonorous objects.

The development of musical instruments among ancient high civilizations appears to have been most substantial in Asia and North Africa. By contrast, instrumental performance remained confined to winds and percussion in Central and South America. It is not always easy to say whether instruments are indigenous to a particular area, since their cultivation may well have spread from one country to another through trade or migration. The harp was used from early times in Mesopotamia, Egypt, and India but was not imported into China before the end of the 4th century AD. In Greece it was regarded as a foreign instrument: the standard plucked instrument was the lyre, known in its fully developed form as the kithara (or cithara). Apart from the trumpet, the only wind instrument in normal use in Greece was the aulos, a double-reed instrument akin to the modern oboe. The Egyptians used wind instruments not only with double reeds but also with single reeds and thus may be said to have anticipated the clarinet. Peculiar to China was the *sheng*, or mouth organ; the Chinese also used as an artistic instrument the panpipes (*hsiao*), which in Greece had merely a rustic function.

Early use of the harp

The most notable development in western Europe was the practice, originating apparently in the 15th century, of building instruments in families, from the smallest to the largest size. A typical family was that of the shawms, which were powerful double-reed instruments. A distinction was made between *haut* (loud) and *bas* (soft) instruments, the former being suitable for performance out of doors and the latter for more intimate occasions. Hence, the shawm came to be known as the *hautbois* (loud wood), and this name was transferred to its more deli-

cately toned descendant, the 17th-century oboe: the Italian name was a phonetic transliteration of the original French. By the beginning of the 17th century the German musical writer and composer Michael Praetorius, in his *Syntagma Musicum* ("Musical Treatise"), was able to give a detailed account of families of instruments of all kinds—recorders, flutes, shawms, trombones, viols, and violins.

Origin of viols and violins

Stringed instruments. The idea of playing a string instrument with a bow appears, from the available evidence, to have originated in Persia in the 9th century AD. The medieval fiddle, like its Eastern prototype, was originally a folk instrument and existed in various forms: by the 16th century these had settled down into two distinct types—the viol, known in Italy as viola da gamba (leg fiddle), and the violin, or viola da braccio (arm fiddle). The two types differ in several respects. The viol has a flat back and sloping shoulders, the violin a rounded back and rounded shoulders. The viol has six or seven strings, the violin four. The viol, unlike the violin, has frets—pieces of gut wound at intervals around the fingerboard—which make every stopped note (*i.e.*, the string being pressed by the finger to produce a higher pitch) sound like an open (unstopped) string. The violin, being the smallest member of the family, came to be known by the diminutive violino: the tenor of the family was called simply viola (in France *alto*; in Germany *Bratsche,* a corruption of Italian *braccio,* "arm"), while the bass acquired the name violoncello, a diminutive of violone ("big fiddle").

The clavichord, harpsichord, and piano

Keyboard instruments. Of all instruments, the organ showed the most remarkable development from the early Middle Ages to the 17th century. Originally, sound was admitted to the pipes by withdrawing sliders or depressing levers. Both of these methods were clumsy: they gave way to a reduction in the size of the levers, which eventually could be depressed by the fingers, while the larger pipes were controlled by pedals. A further development was to separate the various rows of pipes, so that each row could be brought into action or suppressed by means of a draw stop. Once a manageable keyboard had been produced, it could be applied to the portable organ, carried by the player, which was already in use by the 12th century. Scientific experiments with the monochord, a stretched string that could be divided into various lengths by means of a metal tangent, were followed by the construction of an instrument with a whole range of strings and a keyboard similar to that of the organ—the clavichord. A similar adaptation of the plucking of string instruments led to the harpsichord, the ingenious mechanism of which had been perfected by the 16th century. It is curious that a similar method was not applied to the dulcimer, which was struck with hammers, until the early 18th century, when the Italian maker Bartolomeo Cristofori constructed the first pianoforte, so-called because, unlike the harpsichord, it could vary the tone from soft (*piano*) to loud (*forte*).

The effect of the orchestra. The establishment of orchestras, as opposed to chamber groups, in the early 17th century favoured the violin family at the expense of the viols. The latter gradually dropped out of use, with the exception of the bass viol, which was still used in the first half of the 18th century as a solo instrument and as the bass of an accompaniment. It was not until the end of the 19th century that the viols were revived, in order to play the music written for them in the past. Other instruments besides the viols fell into temporary oblivion. The medieval crumhorn, a double-reed instrument with the reed inside a pierced cap, barely survived the 16th century. Another old instrument, the cornett, consisting of a wooden tube with finger holes and played with a cup-shaped mouthpiece, dropped out of use in the 17th century, except in Germany, where it was still played by *Stadtpfeifer,* or "town musicians," in the 18th century. Both Bach and Handel wrote for this instrument, but its tone was too weak for later orchestral music, and it had to wait until the 20th century for a full-scale revival.

Wind instruments. The practice of constructing instruments in families continued from the 17th century onward. English composers wrote for the tenor hautbois, which in various forms was known as oboe da caccia ("hunting oboe," possibly from its curved shape) or cor anglais ("English horn"). There was also an intermediate instrument called oboe d'amore. The bass, or baritone, oboe originated in the 18th century but was not perfected until the 19th. The clarinet (the name means "little trumpet") emerged at the end of the 17th century and, like the oboe, developed into a family extending to a contrabass clarinet in the 19th century and later a subcontrabass. It established itself only gradually in the orchestra in the course of the 18th century.

The work of Sax

The improvement of brass instruments by the addition of valves in the early 19th century led to the creation of others. A pioneer in this field was the Belgian instrument maker Antoine Joseph Sax, who in 1845 built a family of valve instruments called saxhorns, using the bugle as the basis for his invention. Similar instruments were soon being manufactured all over Europe. They were widely adopted in military and brass bands, but only the bass, under the name bass tuba, became a normal member of the orchestra, replacing the older ophicleide and serpent. Sax also invented the saxophone, a single-reed instrument like the clarinet but with a conical tube. This, too, was made in various sizes, which came to be used both in military bands and in jazz ensembles. The saxophone never became a normal member of the orchestra, but the alto and the tenor have been used by composers, largely as solo instruments, and occasionally a complete quartet of four different sizes has appeared in an orchestral work.

Electrically powered instruments. The development of electricity led not only to its use for mechanical purposes—for example, to control the key action and wind flow in the organ—but also as a means of amplification (*e.g.,* in the vibraphone) and as a source of vibrations that can be converted into sound. In the Hammond organ (1935), the player, by selecting the appropriate harmonics that comprise their timbres (tone colours), can produce a tolerable imitation of a variety of orchestral instruments and of certain organ stops, though the tone is not likely to be confused with the sound produced by pipes. A number of keyboard instruments working on similar or related principles have been produced in the course of the 20th century. The use of electronic equipment such as sound synthesizers and tapes to produce and combine sound unrelated to the musical scale lies outside the scope of this article.

BIBLIOGRAPHY. ANTHONY BAINES, *European and American Musical Instruments* (1966), a comprehensive summary, with 824 illustrations; and (ed.), *Musical Instruments Through the Ages,* rev. ed. (1966), chapters on individual instruments written by experts and numerous illustrations and figures in the text; NICHOLAS BESSARABOFF, *Ancient European Musical Instruments* (1941), a detailed account of the instruments in the Boston Museum of Fine Arts (some illustrations); ALEXANDER BUCHNER, *Les Instruments de musique à travers les âges* (1957; Eng. trans., *Musical Instruments Through the Ages,* 1961), a brief history followed by 323 illustrations, some in colour; ROBERT DONINGTON, *The Instruments of Music,* 3rd ed. (1970), a short account of the acoustics and history of European instruments (some illustrations); FRANCIS W. GALPIN, *A Textbook of European Musical Instruments* (1937), a summary history by categories, covering a wide field in a moderate space; KARL GEIRINGER, *Musical Instruments: Their History in Western Culture from the Stone Age to the Present Day,* 2nd ed. (1959), similar in scope to the preceding; FRANK L. HARRISON and JOAN RIMMER, *European Musical Instruments* (1964), a brief account, with 248 illustrations; CURT SACHS, *The History of Musical Instruments* (1940), the most comprehensive work of its kind (some illustrations); EMANUEL WINTERNITZ, *Musical Instruments of the Western World* (1967), a lavishly illustrated anthology of instruments, with an introduction and critical descriptions (251 plates, many in colour).

(J.A.W.)

Musical Notation

Broadly defined, musical notation is a visual record of heard or imagined musical sound, or a set of visual instructions for performance of music. It usually takes written or printed form and is a conscious, comparatively

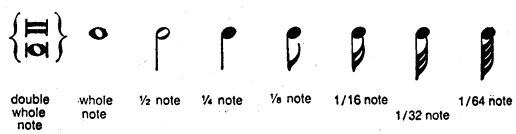
laborious process. Its use is occasioned by one of two motives: as an aid to memory or as communication. By extension of the former, it helps the shaping of a composition to a level of sophistication that is impossible in a purely oral tradition. By extension of the latter, it serves as a means of preserving music (although incompletely and imperfectly) over long periods of time, facilitates performance by others; and presents music in a form suitable for study and analysis.

The primary elements of musical sound are pitch, or the location of musical sound on the scale (hence interval, or distance, between notes); duration (hence rhythm, metre, tempo); timbre or tone colour; and volume (hence stress, attack). In practice, no notation can handle all of these elements with precision. Most cope with a selection of them in varying degrees of refinement. Some handle only a single pattern—*e.g.*, a melody, a rhythm; others handle several simultaneous patterns.

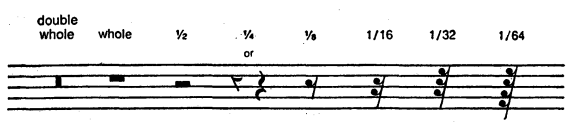
WESTERN STAFF NOTATION

The position of staff notation as the first notational system to be described in this article acknowledges its international acceptance in the 20th century. As an indirect result of colonization, of missionary activity, and of ethnomusicological research—not because of any innate superiority—it has become a common language among many musical cultures.

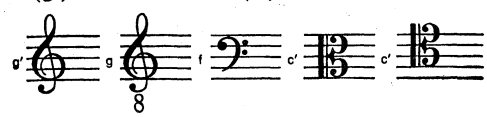
General principles. *Pitch and duration.* Staff notation, as it has developed, is essentially a graph. Its vertical axis is pitch, its horizontal axis, time; and note heads are dots plotting the graph's curve. The five horizontal lines of a musical staff function like horizontal rulings of graph paper, bar lines like vertical rulings. In practice, the system is far more complex and sophisticated than this. The vertical axis of pitch operates to represent melodic contour in music for a single instrument or voice; but when several staves are combined to form a score, the principle breaks down, each staff being a self-contained vertical system. Representation of time (duration) by horizontal spacing is used only in a very limited way. It is in reality made almost redundant because the symbol for a note gives the necessary information itself: not its absolute duration but its duration in relation to the notes around it. These symbols are as follows; each has half the duration of its neighbour to the left:



A system of "rests" measures silence in the same way:



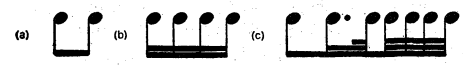
A dot placed to the right of a note head increases by half the duration of that note. Such symbols when placed on a staff may indicate relative pitch and relative duration. In the grid, lines represent alternate notes of the scale and the spaces the intervening notes. Pitch and duration may be fixed by supplying two further indications: a clef and a tempo mark. The clef assigns a definite pitch to a given line of the staff; the first clef fixes the second line up as the G(g′) above middle C(c′):



Tempo and duration. The tempo mark is a sign that lies outside the staff. It appears above and may be a precise fixing of one duration ("♩ = 120 MM" means that the quarter note lasts $^1/_{120}$ of a minute, or one-half second), or it may be an approximate verbal indication setting tempo by reference to accepted conventions (allegro, or quickly; moderato, or moderate speed; etc.).

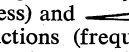
Staff notation is well adapted to two fundamental aspects of Western music: harmony and rhythm. For harmony, note symbols can easily be placed vertically together on a single stem, and these notes need not be all of the same duration; or more than one stem may be used to indicate multiple melodic lines in the musical texture. For rhythm, the existence of an underlying regular pulse, or stress, must be indicated. This is achieved by two devices: the bar line and the time signature. The bar line primarily indicates a point of main stress. Bar lines are usually equally spaced as to duration, though there are numerous exceptions. A time signature indicates, first, the duration of the space between two bar lines (a measure, or bar); and, second, the subsidiary stress patterns within that space. A supplementary system for indicating stress is the device of linking successive notes together by "beaming." Two eighth notes may be linked together as shown in (a); four sixteenth notes (b); or a mixed group of values (c):

The function of a time signature



The implication of such grouping is generally that the first note carries a stress of some sort. Beaming thus may be used either to reinforce the stress patterns of the time signature (the metre) or to contradict it and set up a cross rhythm.

Accidentals. Staff notation rests firmly on the Western system of scales, within which all notes are assumed to be natural unless either accidentals precede them or a key signature is in use. An accidental (♭, or flat; ♯, or sharp) is a temporary raising or lowering of pitch by a semitone; a key signature is the use of the same signs on a more permanent basis, valid to the end of a piece or until countermanded by a new signature. Another accidental, the natural (♮), cancels out a previously indicated flat or sharp and may be used to alter one note or in a key signature to emphasize a key change. Any combination of sharps or flats is theoretically possible in a key signature, but the actual combinations are usually governed by the Western system of keys, or groups of interrelated notes and chords.

Auxiliary signs. Timbre and volume are specified through a variety of additional signs: symbols such as > (stress) and  (increase in volume), verbal instructions (frequently in Italian) such as forte (loud) and col legno (with the wood of the bow) placed above or below the staff wherever space permits. Additional symbols may also provide information about pitch and duration: the dot for staccato, the hold sign (𝄐), the phrase mark, indications of amount of vibrato, and so forth. Other verbal instructions indicate the general manner of performance (pesante, "heavy"; cantabile, "songlike"; etc.) or expression (con dolore, "with suffering"; giocoso, "playfully"; etc.). Further, there are for each type of instrument certain technical signs, as for bowing, breathing, tonguing, or use of mutes.

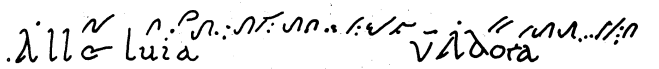
Other auxiliary signs are in the nature of shorthand. Most important are certain symbols indicating performance of notes not shown on the staff. An ornament sign placed over a note may call for several additional notes to be played within the value of that note. It may even delay the sounding of the main note. The precise meaning of such an ornament will vary from one style of music to another and must be interpreted according to the set of conventions governing a particular style.

Types of musical "shorthand"

Comparable to the use of "shorthand" signs for ornaments is the system of placing arabic numerals beneath a bass line in keyboard music of the 17th and 18th centuries. A numeral, or "figure," signifies a harmonic (*i.e.*, a vertical) interval; thus a "6" indicates a note six degrees of the scale above a given bass note (A above C, for example). It is in itself an imprecise measurement, specifying neither whether the interval is major or minor nor in which octave register that upper note should be played. But the figures are governed by the same prevailing key signature as notes on the staff and can, like notes, carry their own accidentals. They are thus not an independent type of notation but a hybrid representation of interval/

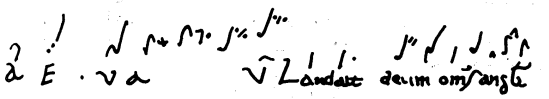
pitch that works in conjunction with staff notation. Its purpose is to indicate the harmonies implied by a bass line (even absence of figuring has a meaning) while at the same time leaving the player free to choose the precise notes that he plays. The systems of letters and figures used by jazz musicians have this same imprecision; they are less dependent upon conjunction with staff notation but lack clear rhythmic significance unless allied to staff notation in at least a simplified form. They operate by defining a harmony in relation to the tonic chord (the chord built on the key note, or tonic), rather than by interval or pitch.

Evoluton of staff notation. *Neumes.* Staff notation has its roots in the neumatic notations of plainchant and secular song of the 9th–12th centuries. Neumes were graphic signs indicating essentially the rise and fall of the voice. Their origin lies probably 1,000 years earlier in signs devised by Greek and Roman grammarians to guide declamation, such as / *acutus* (high voice), \ *gravis* (low), and ∧ *circumflexus* (falling). The musical adaptations of these signs took many different regional forms. Unlike note symbols in staff notation, neumes, with two exceptions, comprised two, three, four, or more notes each and indicated their approximate relative pitches. Each comprised the notes belonging to a single syllable of text, though in florid chant the notes of a single syllable might be split up into several neumes:



Neumes were by intention only a memory aid to singers who knew words and melody by heart. Between the 10th and 12th centuries, however, there occurred significant developments toward a notation that could be sight-read. "Heighted" or "diastematic" neumes were ones that had been carefully spaced on the page in relation to each other, so that an entire line of them formed a continuous graph of pitch over the words of text:

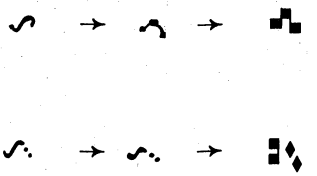
Sight-reading developments

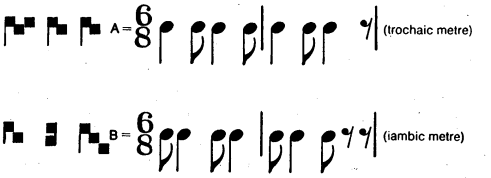


Eventually, precision of pitch was further achieved by using horizontal scratched lines as a grid on which to space the neumes, so that degrees of the scale fell alternately on a line or in a space, and by colouring one line red to signify the pitch "F" and possibly another yellow to signify "C"—or simply by placing a letter "F" or "C" at the beginning of the appropriate scratched line. Together, the two devices fixed the relative pitches of all notes by indicating where the semitones of the scale occurred (that is, immediately below the marked lines: E–F or B ♮–C). In the 11th century two signs from a quite different system of notation (alphabetical notation; see below) were incorporated as accidentals before the pitch "B": b, the ancestor of ♭; ♮, the ancester of ♮ and ♯, and also of the German "h," which refers to b ♮. These two signs were progressively applied to other pitches in the following centuries. By the 13th century a four-line staff ruled entirely in black or red had become established, using stylized forms of the letters *f*, *c*, and *g* (ancestors of the modern 𝄢, 𝄡, and 𝄞) as clefs. For polyphonic music (having several simultaneous voice parts) a five-line staff became standard by the 14th century, but keyboard music in some countries used six- or seven-line staves as late as the mid-17th century.

Adoption of the five-line staff

During the 12th century, in northern and northeastern France the thin, curved lines of neumes were drawn more thickly at the points corresponding to the separate notes within them. In time, a firmly rectilinear notation of heavy horizonal pen strokes, diamond-shaped dots, and hairline vertical strokes emerged, whose groups of notes are called "ligatures":



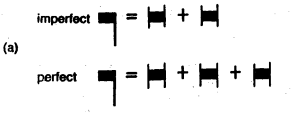
This was the notation of the troubadours' and trouvères' songs; also of plainsong from the 13th century to the present day. It was also used in 12th-century polyphony for the upper voices, which were without text. Freed from syllabic considerations, the grouping of notes into ligatures took on rhythmic significance, specific groupings representing short, repeated patterns called rhythmic modes:

A ligature did not yet have a single, unvarying meaning. Its rhythmic pattern depended upon context.

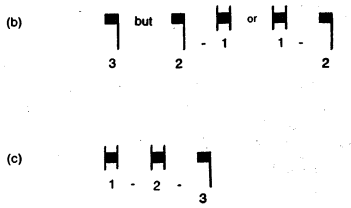
Mensural notation. The freeing of ligatures from considerations of context occurred during the early 13th century. Time values for ligatures, single notes, and rests were codified around 1260 by the influential theorist Franco of Cologne. The notes then in use included the duple long, later called maxima (▬); long (▮); breve (■); and semibreve (◆). In French music a shorter note value was created: the minim (▮).

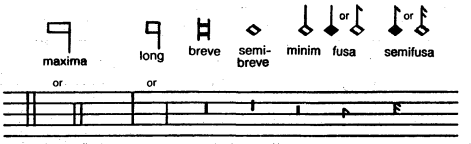
These note symbols provided the basis for notation from the late 13th to the late 15th centuries. This system, called mensural notation, was based on several fundamental principles that determined the value of a note relative to that of its neighbours. In the terminology of mensural notation a given note might be either perfect—*i.e.*, divided into three notes of the next lesser time value; or imperfect—*i.e.*, divided into two notes of the next lesser value. Thus, as in part (a) of the example below, a long might be perfect, containing three breves; or imperfect, containing two breves. To determine which note symbols were perfect and imperfect for a given piece, special symbols were devised: ⊙, ◯, ⊙, ◯, ◯, ¢, 3/2, 3/1, 4/3, etc. (Of these 𝐂 and ¢ survive, together with fractions such as 3/2, as modern time signatures.)

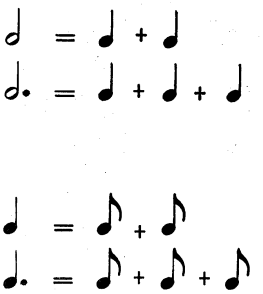
Specific rules provided for lengthening or shortening the value of notes in certain instances. To "imperfect" a long meant to shorten it by one-third of its value; this occurred when the long was preceded or followed by a breve or by notes equalling a breve in value, as shown in part (b) of the example below (for purposes of illustration, numbers have been placed under the notes showing their time value relative to the shortest note). Part (c) of the example shows another common alteration of time value: in a composition in which the long is perfect, a breve (brevis) preceding it may be doubled in value under certain conditions. The time-value relationships between breve and long similarly applied to the other pairs of notes: maxima and long (longa); breve and semibreve (semibrevis); and semibreve and minim (minima).

As the system of mensural notation evolved, another device, coloration, came into use. If a composer wished to





render a potentially perfect note imperfect, he could write it in red or as a hollow note (as 𝅗𝅥, 𝅗𝅥, or ◇); these two devices had, however, various other, less common meanings. About 1400, hollow note shapes were adopted where full black notes had hitherto been used, and full black served as coloration. The notes then current and their corresponding rests were as follows:

Transition to modern staff notation. In 16th-century manuscripts and, rather later, in printed music, the diamond-headed notes became rounded. Ligatures were used less and less often in the later 15th century. The principles of perfection and imperfection gave way to the modern relationship of 2 to 1 between adjacent note values, with the dot adding an extra half value to give a 3 to 1 relationship.



Shorter note values were also introduced, and the old, longer ones became obsolete. Yet, because of a paradoxical survival from 15th-century practice, slow music has tended to be written in short values (for example, Beethoven's slow movements) and fast music in long values.

Introduction of the bar line

The bar line as a measure of metre arose first in 15th-century tablatures (notation showing playing position rather than pitch, as for lute). Barring entered staff notation in the 17th century, but regularly spaced barring became a practice only in the 18th century. Separate tempo indications, arising first in the 17th century, were verbally expressed; for example, adagio, largo, presto. The range of these terms greatly increased during the 18th and 19th centuries, and the metronome mark, an absolute indication of tempo, has never superseded them since its arrival in Beethoven's day. The bulk of the shorthand devices emerged during the 17th century, figured bass early in the century, the majority of ornamental signs later. Indications for loud and soft arose early in the century, expressed as words (forte, mezzoforte, piano) and later as abbreviations (*f, mf, p*). Graphic signs for dynamic and attack (staccato dot, crescendo mark, for example, and also phrase marking) appeared in the 18th century. A great proliferation of dynamic instructions and signs occurred in the late 19th and early 20th centuries.

Twentieth-century notation. Two developments in musical style in the 20th century have placed great strain upon staff notation: integral serialism—in which the music is controlled by a mathematical system—and indeterminacy, or chance music. In the former, every note in a texture may have its individual dynamic marking and type of attack (for example, in Olivier Messiaen's *Mode de valeurs et d'intensités* and in parts of *Structures I* and *II* of Pierre Boulez). There may also be extensive use of unconventional playing techniques. Since staff notation indicates volume, attack, and technical effects in a comparatively clumsy manner, the written page becomes cluttered and unclear. In indeterminate and aleatoric music (the latter type allowing the performer a limited degree of freedom), the notation must offer choices to the performer or be deliberately imprecise. Staff notation is for these purposes often too specific. In addition, electronic music, composed with such devices as graphs, mathematical symbols, and diagrams, is not easily translated into a readable "score" for publication.

Notation that offers the performer a choice

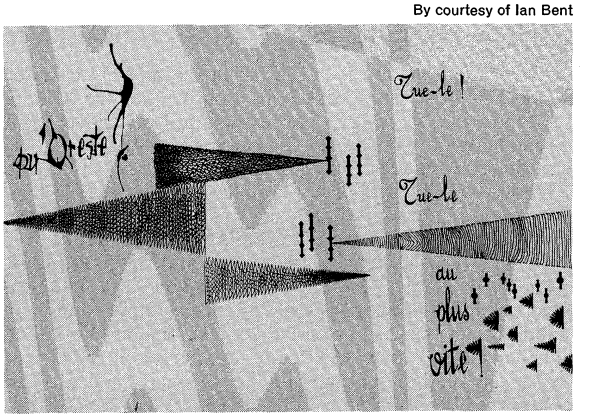
New systems. Notation, in face of this, has moved in two directions: toward adaptation of staff notation and toward the devising of new notational systems. Music using microtonal intervals (less than a semitone) has tended to adapt by modifying the standard accidental signs:

—meaning one-third sharp, two-thirds sharp, and so on (*e.g.*, in Krzysztof Penderecki's *Anaklasis*). So-called space time notation is a further adaptation that reasserts the graphic nature of staff notation. It abandons symbolic indication of note values and replaces it by the spacing out of note heads horizontally on the staff, accompanied by an instruction such as "1 inch = 1 second." The principle may then be amplified by using different note heads (𝅝, □, ◇, etc.) to signify short, medium, and long sustaining of a note, thus obviating the use of rests, and by beaming together notes to be performed in one breath or bow (*e.g.*, Witold Lutosławski's *Trois poèmes d'Henri Michaux* and John Cage's *Music of Changes*).

Indeterminacy. Indeterminate music requires constant experimentation with notation. A composer may offer directions for one element of the music—as rhythm or pitch contour—and leave the performer to improvise the remaining elements. Or he may simply describe the general character of a passage by resorting to a specially designed symbol, a verbal description, or even an impressionistic drawing (as in Earle Brown's *Hodograph*). At the extreme, John Cage supplies "materials" (lines, dots, shapes) and leaves the performer to attach musical significance to them. For electronic music, published scores have so far adopted either strict graph form (Karlheinz Stockhausen, *Electronic Study II*) or pictorial form, using patterned drawings to represent different timbres (as in Henri Pousseur's *Electre* and Stockhausen's *Kontakte*).

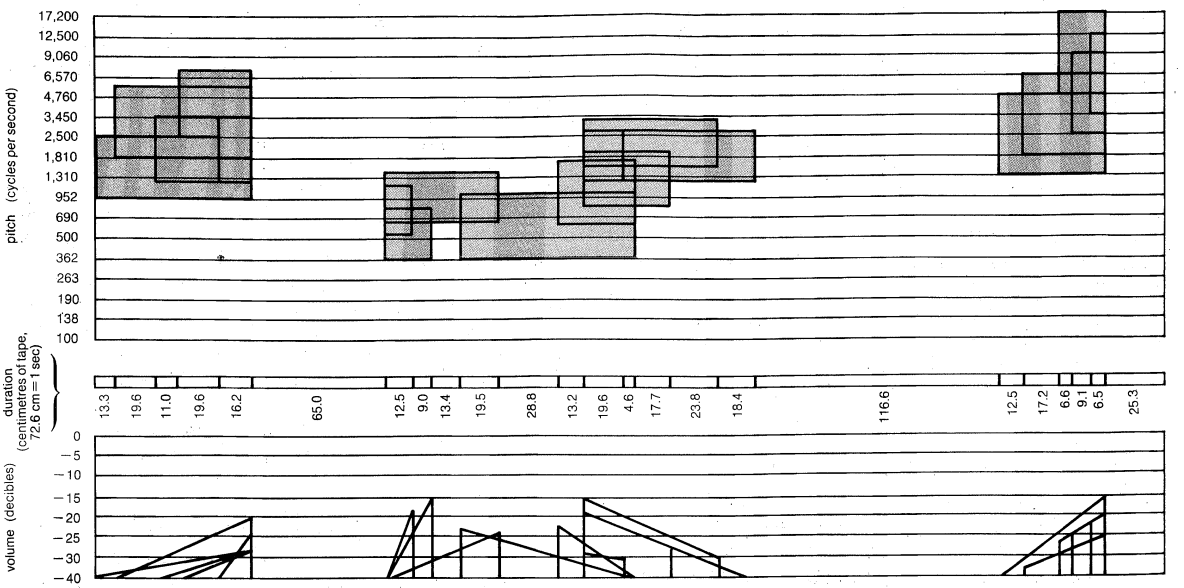
Cage's "materials"

By courtesy of Ian Bent



Page from the score of *Electre* by Henri Pousseur.

Adaptation to non-European music. Notations evolve with the musical styles they serve, and they reflect the underlying aesthetics of their own cultures. Thus staff notation is ill-equipped to cope with non-Western scales and tunings, with music to which the idea of the "note" (a stable, sustained pitch) is foreign, or with music whose subtlety lies as much in delicate gradations of volume or timbre as in pitch and rhythm. Ethnomusicologists have developed a range of supplementary symbols—*e.g.*, for notes of uncertain pitch, glissandi (slides), slight length-



Page from the score of Stockhausen's *Electronic Study No. 2.*

© 1956 Universal Edition Ltd., London. Used by permission of the publisher, Theodore Presser Company sole representative in U.S.A., Canada, and Mexico

ening of a value, half-voiced notes, and other sounds. They have also experimented with staves of fewer or more lines. The Western system of proportional note values (for example, quarter note = half of a half note) does not easily cope with fine fluctuations of value; instead, constant changes of metronome tempo mark may be necessary. Among the most complex uses of staff notation in ethnomusicology are the transcriptions of Serbo-Croatian and Romanian folk song by the Hungarian composer Béla Bartók. Other transcribers have used graph paper to draw a curve of pitch against time. Many significant mechanical methods of transcription have been devised. The two most notable are the melograph, invented by the ethnomusicologist Charles Seeger, which traces a pitch/time graph immediately above a volume/time graph, and a device developed by Dahlbeck, which produces two similar graphs by means of a cathode-ray tube. These methods can reveal a level of interpretation by the performer that aural transcription into staff notation fails to bring out.

OTHER SYSTEMS OF NOTATION

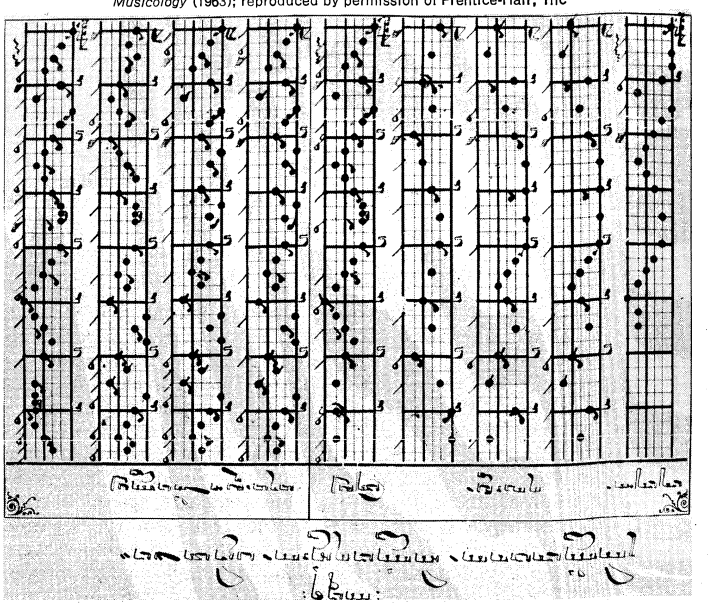
Written notations are to be found in the musical cultures of the Far East, Southeast Asia, South Asia, the Middle East, and the West. The earliest examples survive from Ancient Egypt and Greece. The written symbols of notation may be classified into two broad categories: phonetic symbols—words, syllables, abbreviations of these, letters, and numbers; and graphic signs—accentual signs for the rise and fall of the voice (developing into neumelike "ecphonetic" signs), curves, lines, dots, and other symbols, perhaps originally depicting hand signs, and neumes. Symbols in both categories may denote simple sounds or stand for groups of successive sounds. In the West they are read in lines from left to right, whereas in the Orient many are read from right to left or read vertically, in columns.

Major categories of notational systems

A second fundamental distinction is that between representational notations, which depict the sound of the music—leaving the player to produce that sound as he wishes—and tablatures, which instruct a player as to the technical means of producing a sound. Phonetic symbols play an important role in both types of notation, while graphic signs contribute mainly to representational notations. A prime example of non-Western representational notation is the *kraton* notation used in music for the Javanese gamelan orchestra, its grid using the "graph" principle found also in Western staff notation but oriented at a 90° angle relative to the latter.

Verbal and syllabic notations. In oral traditions of music, solmization (the naming of each degree of a basic scale with a word or syllable) plays an important role. The modern European sol-fa method ("do," "re," "mi," etc., representing the notes of the rising major scale at any pitch transposition) is such a system. The Indian

From Mantle Hood, "Music the Unknown" in Harrison, Hood, and Palisca *Musicology* (1963); reproduced by permission of Prentice-Hall, Inc



Javanese *kraton* notation.

syllables *ṣa, ṛi, ga, ma, pa, dha, ni* are similar, as are the Balinese *ding, dong, deng, dung, dang,* the ancient five-note Chinese scale *kung, shang, chiao, chih, yü,* and the Korean *tŭng, tung, tang, tong, ting.* Slightly different are the 12 chromatic Chinese *lü* syllables: each pitch bears the name of a bell—as *huang chung* ("yellow bell"), *ling chung* ("forest bell")—and its name is reduced to a syllable—*huang, ling, t'ai,* etc. Though primarily for reciting or singing as a melody is being learned, these syllables can be used to write down the notes of a melodic line (each appearing as a single ideogram in the Chinese examples) and thus form a simple syllabic pitch notation. Of the five Balinese syllables, only their vowels are used in written form—*i, o, e, u, a*—so that a letter notation results; this is still essentially an abbreviated syllabic notation, not an alphabetical notation.

Words and syllables are not only applied to pitch. In Western plainchant, abbreviated words were used to indicate duration (for example, *c* = *cito* or *celeriter* =

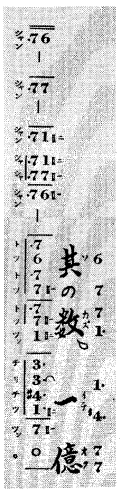
Notation for duration

"short" value) and direction of melodic movement (*l* = *levare,* and *s* = *sursum* = "upward"). In the notation of early Ethiopian church music a single letter or a pair of letters (short for a passage of text) signified a group of notes, even a complete melodic phrase. The drum syllables of North Indian music are a solmization of timbre (as *na, ta, dhin*) and often also of rhythmic patterns (as *tirikita, dhagina*) and can be written down to make a notation.

Alphabetical notations. Alphabets are historically a phenomenon of the Middle East, Europe, and the Indian subcontinent. Their ordering of letters provides a convenient reference system for the notes of musical scales in ascending or descending order. Alphabetical notations are among the most ancient musical scripts. Two Greek notations were of this type, the earlier using an archaic alphabet and the later using the Classical Greek alphabet. Many comparable notations arose in the Middle Ages, and the modern note names, A to G, are an outgrowth of these. The clefs of staff notations are a formalized survival. The system of pitch notation devised by the 19th-century German philosopher and scientist Hermann Helmholtz was derived from the Greek system, using dashes for octave register but employing Roman letters: $A_{,,}$, $B_{,,}$, $C_{,}$–$B_{,}$, C–B, c–b, c′ (middle C)–b′, c″–b″, c‴–b‴, etc.

Numerical notations. The notation of pitches by numbers is rare. A modern Javanese system allots numbers 1–7 to the pitches of the seven-note *pelog* scale, and a similar five-note system exists in Bali. Modern Japanese notation for the samisen (a type of lute) uses 1–7 for the diatonic scale, 7 being the lowest note; and modern China has a similar system for publishing popular songs. An Arabic notation of the 16th century used the first seven Persian numbers to signify pitches of a seven-note scale. Numbers lend themselves more directly to expression of duration, of metre, or of the division of a basic time unit into a number of equal parts (for example, the Western triplet $\overset{3}{\frown}$ and its extensions in the works of Chopin $\overset{11}{\frown}$ $\overset{21}{\frown}$; and the Western time signatures). Arabic alphabetical notation of the 13th and 14th centuries used arabic figures, placed beneath the pitch letters, to indicate durations of individual notes.

From William P. Malm, *Japanese Music;* reproduced by permission of Charles E. Tuttle Co., Inc., Tokyo



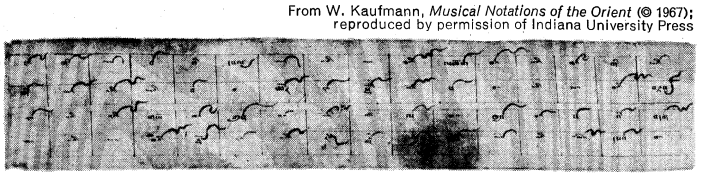
Notation for Japanese samisen from the piece *Tsuru kame.*

Graphic notations. The character of neumes and of accentual signs has already been described under Western staff notation. The Vedic chant of southern India uses a form of accentual notation: a dot beneath or above a syllable of text indicates a lower or upper reciting pitch. Analogous systems, involving dots and dashes, formed a notation for ancient Jewish cantillation and for early Syrian Christian chanting. Such signs are really only an extension of punctuation. A more developed form ("ecphonetic" notation) was used for recitation of Byzantine liturgical chants. In addition to simple signs for vocal inflection, it had also more elaborate, compound signs, such as

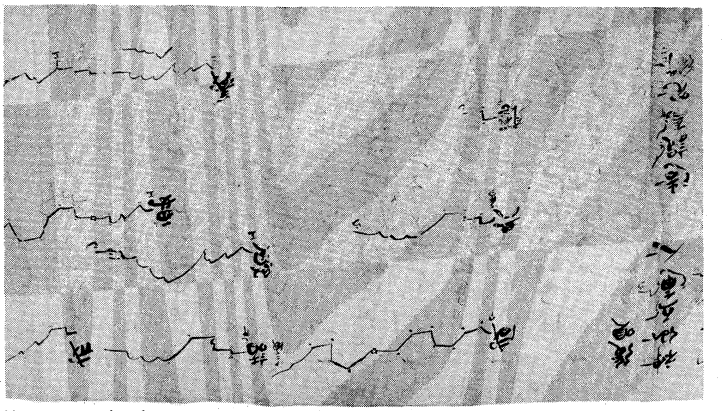
capable of signifying degrees of loudness and changes of voice production. The Western phrase mark and crescendo and decrescendo symbols are graphic signs of this type.

The dividing line between compound ecphonetic signs and neumes is slender. Neumes are concerned not with inflection of the voice between high, medium, and low but with groups of sung pitches rising and falling over a quite narrow range: a neume may represent a given pattern of intervals whether it lies high or low in the voice's compass. Neumes made up of curves, hooks, lines, and dots are found in Tibetan Buddhist chant books; in vocal notations for Japanese Buddhist chant, Nō plays, and *gagaku*

From W. Kaufmann, *Musical Notations of the Orient* (© 1967); reproduced by permission of Indiana University Press



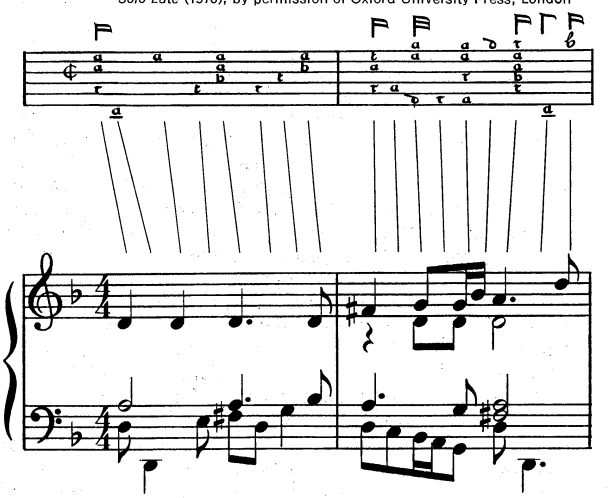
Neumes for a Tibetan Buddhist chant.



Neume notation from the third part of the psalm of Upāli.
By courtesy of Todaiji, Nara, Japan

(court music); and, in simplified form, they are used in Chinese notation in conjunction with pitch syllables.

Tablatures. A tablature notates music as a series of playing positions. Hence a tablature for a plucked stringed instrument guides the player's fingers to the string he must pluck and to the fret at which the string should be stopped. Similarly, a tablature for a wind instrument guides a player's fingers to cover certain holes, and a tablature for percussion directs a player as to which drum to strike, which hand and stick to use, and what type of stroke to execute. Each single instruction in a tablature corresponds to one action by the performer. The order of his actions is automatically prescribed, and more precise rhythmic indication can be given quite easily as the length of time between successive actions—rests are unnecessary. Thus, a tablature for a plucked instrument (for example, a lute, guitar, or zither), requires signs for: each string, each fret, and possibly also each right-hand plucking finger, direction of stroke, and ancillary techniques such as harmonics, vibrato, and left-hand plucking. To indicate these, the tablature may use letters, numbers, and graphic signs. Thus 16th-century French lute tablatures indicate the strings (more properly the "courses") graphically as parallel horizontal lines, frets as letters, and right-hand fingering by dots:

From Francis Pilkington, Brian Jeffrey (ed.), *Complete Works for Solo Lute* (1970); by permission of Oxford University Press, London

Italian lute tablatures use numbers in place of letters.

The tablature for the Japanese koto zither is simpler in that its 13 strings are not stopped. The pitch of each string is indicated at the beginning of a tablature, and thereafter the strings are represented by numbers combined with graphic signs for special technical effects. The tablature for the Chinese ch'in, a zither whose strings are stopped, uses a combination of numbers (for strings and for stopping points) and ideograms (for other technical details). These are grouped close together in composite symbols. One composite symbol may contain an ideo-

Tablature for the Chinese ch'in.

From Mantle Hood, "Music the Unknown" in Harrison, Hood, and Palisca, *Musicology* (1963); reproduced by permission of Prentice-Hall, Inc.

gram for the left-hand finger; a number for the stopping-point; another number for the string itself; an ideogram for the right-hand finger; and possibly an ideogram indicating loudness, legato, glissando, etc. German 16th-century lute tablature achieved economy of space by indicating string and fret as a single letter (as, k = first string stopped at third fret). No grid of lines was necessary.

BIBLIOGRAPHY. WILLI APEL, *The Notation of Polyphonic Music, 900–1600*, 5th ed. (1961), a standard textbook, including staff notation and tablatures; "Solmization," "Notation," and "Tablatures," in *Grove's Dictionary of Music and Musicians*, rev. ed. (1955); ERHARD KARKOSCHKA, *Das Schriftbild der neuen Musik* (1966), an excellently documented study of contemporary notation; WALTER KAUFMANN, *Musical Notations of the Orient: Notational Systems of Continental East, South and Central Asia* (1967); CARL PARRISH, *The Notation of Medieval Music* (1957), excellent facsimiles; EMANUEL WINTERNITZ, *Musical Autographs from Monteverdi to Hindemith* (1955), a study of musical handwriting, with many facsimiles.

(I.Be.)

Musical Performance

Musical performance is that step in the musical process during which musical ideas are realized and transmitted to a listener. In Western music, performance is most commonly viewed as an interpretive art, though it is not always merely that. A performer to some degree determines aspects of any music he plays. Issues of tempo, phrasing, dynamics and, in some types of music, pitches and instrumentation are subject to a performer's discretion. In this article, the various aspects and traditions of musical performance in Western and non-Western civilization are treated and their development traced.

Because the pleasure men derive from sounds has always been closely related to the pleasure they derive from making the sounds themselves, it is difficult to conceive of the origin of music as separate from an act of musical performance. Models for the establishment of rhythm may be found in heartbeat and breathing, and in the inflections of speech and cries of grief, pleasure, and desire are found the source of what became song.

The earliest visual manifestations of musical performance are found in rock paintings and excavated objects. While the interrelationship of music and ritual is clear, there is evidence that music was performed for dancing, in various work activities, and primitive games as well. Flutelike instruments of many sizes, made from bones and wood, and elaborate percussion instruments figure prominently in all primitive cultures, in which these instruments often were assigned symbolic significance associated with forces of the supernatural.

Singing is most probably the oldest musical activity. Even in the most primitive cultures the singer has had a special, defined position. In primitive singing there are three classes of sound: the first is called logogenic, in which words form the basis for the wavering musical incantation; the second, called pathogenic, consists of harsh, forceful, percussive, nonverbal sounds emitted to express strong feeling; in the third category, called melogenic, the sounds of the two previous categories combine to form a contour of pitches which pursue a course seemingly dictated by the weight of tensions inherent in the sequence of pitches and hence melodic in effect.

Primitive societies evolved several means to relieve the monotony of one person's singing. A principle device is called antiphony, which involved two groups that sang in alternation or a leader who sang and was answered by a group of singers. In the latter may be seen the origin of responsorial singing, which continues today and which may be the point of origin for several types of musical-phrase structures. Polyphony was also anticipated in primitive musical performance. It appeared through haphazard rather than intentional manifestations, such as the singing of the same melody with the parts starting on different pitches or at different times.

Primitive antiphony

WESTERN MUSICAL PERFORMANCE

The performer as interpreter. Music as an interpretive art is a relatively recent phenomenon. In primitive societies, music plays a ritual role based on an oral tradition, and each performer in a sense interprets the tradition, but, more importantly, he renews it and transforms it through his own performance.

The development of the performer's role as interpreter coincided with the development of musical notation. Because composers for so many centuries were in a position to supervise the performances of their music, certain aspects of performance were not notated. Notation has grown increasingly complex as the dissemination of printed music has become more widespread. Ultimately, the degree of judgment a performer may exercise is determined by the period in which the music was composed. For music of certain periods, even though their notational systems are incomplete and give few indications of how the music should sound beyond pitch and rhythm, musical scholarship has amassed much information concerning proper instrumentation, ornamentation, improvisation, and other traditional performing practices that determine to a large degree the sound and stylistic character of the music. A performer as interpreter operates within a range of limitations imposed upon him by his understanding of the printed page, whatever knowledge may be available concerning the tradition that surrounds the music he is playing, and the extent to which his personal tastes coincide with this information. Certain aspects of the musical taste of the past sometimes cease to be expressive and gradually disappear from usage. Just as often, with the passage of time, performers tend to reassess the literature of previous ages and find renewed interest in practices that an earlier generation may have set aside. In any case, the performer as interpreter speaks to and with the tastes of his own time. And his task, no different from that of the performer in primitive music, is to renew, to refine, and to enrich the materials and traditions he inherits.

Mediums of performance. The mediums for musical performance are extraordinarily various. Western technology has had a tremendous impact on the development of musical instruments and has thereby greatly expanded the means whereby music is made. Performance may be vocal, instrumental, or electronic. Vocal performance is the oldest and the primary influence for the development of all subsequent musical gestures and materials. Instrumental music began with the development of percussion instruments and crude horns; string instruments came later. Electronic music is a 20th-century development involving the reproduction of traditional performance mediums through electronic means, at the same time that it is also evolving composition and performance of its own. At first it reproduced natural sounds by electronic means;

Vocal, instrumental, and electronic performance

later, composers and technicians began to invent electronic sounds and to discover new sound relationships.

Solos. In all musical mediums the solo performance is the most spectacular. The power of music to compel attention and to stir emotions lends to the solo performer an especially fascinating aura. This is the domain of the virtuoso, that musical performing phenomenon of prodigious technical mastery, invention, and charisma. Most solo literature includes another instrument or group of instruments, and the literature varies from one medium to another according to the expressive range and technical capabilities of the solo instrument.

The largest solo literature for a single instrument is for keyboard instruments. Vocal solo literature is very important and extensive, and the string instruments also have a distinguished solo repertory. The wind, brass, and percussion solo literature is more restricted.

Ensembles. In vocal and instrumental chamber ensemble performance, the performing groups are divided into duets, trios, quartets, quintets, sextets, septets, and octets, which exist for every medium and combination. Of particular importance is a string quartet consisting of two violins, viola, and violoncello. Dating from the 18th century, this instrumental ensemble is analogous to the vocal ensemble consisting of soprano, alto, tenor, and bass.

Beginning of symphonic music

Symphonic music dates from the 17th century. With the rise of the middle class and its aspirations for culture, music as an art required performing situations which would accommodate more people. Larger halls required ensembles acoustically suited to the expanded performing area. The primary result of this development was the symphony orchestra with its multiple string, wind, brass, and percussion instruments.

Ensemble performance places a special responsibility on the concentration of the performer, who must attend not only to his own playing but also to that of all the others in the ensemble. All aspects of the performance depend on this mutual awareness. The leader of most small ensembles is one of the performers, the first violinist, a keyboard player, or one of the singers who indicates tempi, entrances, and musical character and supervises rehearsals. As ensembles grew in size and complexity and their problems of coordination increased, the leader set aside his instrument and devoted his efforts to the beating of time and the communication through clear hand signals of his indications for entrances, tempo changes, dynamic accents, and the shaping of phrases. The conductor often occupies a position analogous to that of a soloist in the attention of an audience, though the conductor himself makes no musical sound. As they are chiefly responsible for the music orchestras play, both in terms of choice and execution, conductors have had considerable impact on the development of music.

Opera. Opera, the marriage of music and drama, is the most complex performance situation. It entails much more than a single performer or group of performers, their instruments, and a hall in which to play. Text, decor, costumes, histrionic projection, preparation time, as well as singers, instrumentalists, and a bevy of extramusical technicians must all be brought together and coordinated into the final production.

Artistic temperament. Many forces interact in developing those traits that distinguish various performing traditions and individual performers. Personality and temperament fundamentally affect the manner in which a performer works, as does his cultural milieu. There are performers who use music as a vehicle for display and others for whom performance is only a means to illuminate the music. Nor does performance necessarily mean public performance. For some people musical performance is essentially private, requiring no confirmation in the form of audience approval. The musical pleasure of such people rests solely on performing, either alone or with other musicians. Much chamber music is played under these private circumstances, and much music has been written for such situations. This used to be the primary realm of the musical amateur, that skilled but nonprofessional music lover, whose ranks are diminishing in the era of the phonograph.

The type of performing situation at the opposite end of the spectrum is one directed to securing audience attention and affections. The need for audience approval has led to innovations as well as some decadence in its impact on the musical scene: innovation, if the performer is led to discover imaginative and fresh means of attracting public acclaim; decadence, if the devices for audience attraction become cheap and thinly spectacular, when the performer may distract the audience from more deserving work and debase their taste.

Intuition versus intellect

Intuition and intellect figure prominently in the temperament of a musical performer. Intuition is the capacity to do the musically "right" thing without instruction or special consideration of the alternatives. Intellect is the means whereby a musician enlarges the range of his instincts through the pursuit of new information, reflection, and analysis of the musical material at hand. Each element informs and completes the other.

Many musicians depend heavily on intuition in solving performance problems. Their solutions are often imaginative and fresh and their performances exciting. Others pursue a methodical path as they examine minutely relevant musical details. They analyze thoroughly the scores they perform, comparing manuscript facsimiles and various printed editions, and attempt to discover new musical relationships, new ways of delineating these relationships in performance, and, in short, new ideas as to how the music might best be played and how it should sound. Art, poetry, biographies, cultural history, and any material relating to the period of the piece of music being studied for performance may be sources of musical insight for the performer.

National characteristics. At various times in history, national origin has been considered an important delineating characteristic in musical performance. This is partly the result of certain consistent emphases and features in the music written by composers of different nationalities. The Italian interest in the voice has evolved a special singing (bel canto) quality in their music, which has carried over into their music for instruments (the string instruments especially), and into the general texture of Italian music, which has always given melody special prominence.

The English have had a highly developed and sophisticated musical performance tradition. Amateur improvisation figured importantly in its early history. While this has perhaps tended toward a conservative musical atmosphere, it has also produced a high standard for performance. The French have maintained a strong sense of national identity in their performing arts. In music their concerns for orderly design, delicate expressiveness, simplicity, naturalness, and beauty of sound extend back for centuries. In German music, articulate philosophical and structural considerations have played important roles in developing nationalistic traits in their tradition of musical performance.

The rich folklorist traditions of Spain, Hungary, and Russia have influenced rhythm, melody, and sonority in Western musical performing traditions. The Russian schools of string and piano technique have greatly advanced the performance resources of these instruments in the past 100 years. The United States, younger and more heterogeneous, has had a shorter musical history but an abundance of great symphony orchestras and solo artists, who are in demand because of their precise execution, versatility, and breadth of repertory.

Historical stylistic developments. In antiquity the Sumerians, Egyptians, Greeks, and Romans evolved the first aesthetic theories and musical systems relevant to the music of the modern Western world. Unfortunately, few actual musical examples survive because of primitive notational practices and the gradual erosion of oral traditions. What is known is derived from the writings of the period and iconography—depictions of performing musicians, instruments, and musical events in sculpture, and in wall and vase paintings.

In the Middle Ages traditions of musical performance were kept alive by the church and in the music sung and played by wandering minstrels.

In the Renaissance, polyphony (combining several simultaneous voice parts) and the early precursors of modern tonality (organization of music around a focal tone) were developed. The smooth flow of Renaissance liturgical counterpoint (polyphony) and the perky rhythms of secular Renaissance dance music remain as models of taste and musical technique even into the 20th century.

Developments in the Baroque

The performer emerged as a central figure in the focus of musical attention and purpose during the Baroque period. The heightening of the role of the individual performing artist and the invention of increasingly dramatic gestures to demonstrate performers' skills combined with a steady refinement in the construction of musical instruments. The reduction of musical materials to two modes (scale and melody patterns), in this case the major and minor scales, and the initial efforts to compose with large musical forms (opera, oratorio, sonata, and concerto) took place in this period. It is notable that in the Baroque era the equal-temperament system for tuning the strings or pipes of keyboard instruments evolved—a development that has had profound effect on the nature of musical language.

In the Rococo or Classical period that followed, the elaborate contrapuntal texture of Baroque music gave way to music of subtle dynamic differentiation often based on simple folk materials (rhythms and melodies). The relationships between tonal materials and large musical forms achieved their highest state in the sonata and in opera.

The Romantic age was a period of refinement and intensification of Rococo principles with heavy literary overtones. It was the true age of the star virtuoso; that is, the age in which the role, person, and effect of the virtuoso was most dramatized and glamourized. The symphony orchestra in this period achieved its maximum development. Italian opera under Verdi found its noblest expression and German opera with Wagner expanded into the *Gesamtkunstwerk* ("complete art work").

Modern music dates from an era beginning roughly around World War I. Concert life has remained more or less what the 19th century established; the virtuoso conductor and performer dominate the musical establishment. More chamber, symphony, and opera programs are performed in more parts of the world than ever before, but musical progress has ceased to occur in established organizations and is found instead in the combo: jazz, rock, improvisational, experimental, live electronic, and multimedia groups that have sprung up in the 20th century.

HISTORY OF WESTERN MUSICAL PERFORMANCE

Antiquity. The civilizations of antiquity expanded the role assigned to music in primitive cultures. The Sumerians established the foundations for the tradition of liturgical music. Some of the prayers which they sang have survived. From various artifacts of this civilization something is known about Sumerian musical instruments and some of the situations in which music was played. Such instruments as the lyre, harp, sistra, pipes, timbrels, and various drums figured importantly. Particular instruments were identified as accompaniment with specific types of religious poetry, and indeed the development of different poetic genres seems to have been considerably influenced by the nature of these instruments. While its primary purpose was religious, music also had something of a secular role in Sumerian culture and was played in processions, at banquets, and during primitive sporting events. Music as a profession first developed in Sumerian culture. Both men and women participated as singers and instrumentalists and held priestlike positions with specific functions and ranks of authority.

The musical culture of ancient Egypt, which apparently emerged from the same primitive sources as Sumer, resembles that earlier culture in many aspects: the close relationship between music and religion, the presence of a musical profession, some secular musical activity, and similar musical instruments. Of special interest in Egyptian music is the development of chironomy, the use of hand signals to indicate to instrumentalists what they should play. The singer in this manner guided instrumentalists through melodies with which the singer was seemingly more familiar than the players.

In these ancient cultures there was no notational system or codified theory of musical practice. Different musical traditions were exchanged in the process of trade, migration, military conquest, and intermarriage to form that common body of practices which is the basis of Western music.

Importance of Greek music

Of the early civilizations, Greece provided the musical culture of greatest significance for the development of Western music. The system of scales and modes, as well as a large part of the general philosophy concerning the nature and effect of musical sounds, has been inherited from the Greeks. It was also the Greeks who developed the theory of ethos, which defines the character of psychological and emotional responses to different musical stimuli. Building on the ancient religions and magical accoutrements of music, the Greeks assigned specific mental and emotional states to specific pitch arrangements and instrumental combinations. Music infused with this motivating power stood at the centre of the social order.

Though a major part of Western musical terminology, basic music theory and philosophy, basic notational practices, and the foundations of acoustical physics derive from the ancient Greeks, very little of their music has survived. The great ethical significance of music in Greek society caused performing mastery to be an essential aspect of education. Everyone was taught to sing and to play instruments. For a major part of the period all music was a setting of words with instrumental accompaniment, for the most part doubling the voice at the interval of octaves, fourths, or fifths. It was only in the later part of the period, after the age of Pericles (late 5th century BC), that instruments began to play independently of singers.

Music, in the later stages of the Hellenistic period, became an increasingly important part of public spectacles. As musical performance became increasingly secularized and became the property of the masses, the upper classes withdrew to esoteric considerations of the art and reflections on its past. It was perhaps at this point that music was divided into two fairly artificial categories: the contemplation of music's nature and history and practical musical performance.

Assuming the artistic mantle of ancient Greece, the Romans disseminated Greek music throughout the known world. The essential role of music in the Roman Empire remained unchanged. Rome's principle contribution consisted in serving as a catalyst for the mixing of Hebraic and Hellenic traditions of musical performance, which, preserved by the Christian Church for a thousand years, emerged again in the Renaissance into the first flowering of modern musical practices in the West.

North Africa and the Near East. Although not in the mainstream of Western musical performance, Islāmic (North African and Near Eastern) classical music closely approaches the orchestral tradition of European music in one respect: large choruses and orchestras—consisting of tambourines, pot drums, recorder-flutes, *ʿūds* (plucked lutes), bowed lutes, and dulcimers—are assembled to perform "suites" consisting of a series of instrumental solos and orchestral selections interspersed with unison choral songs or solo recitatives based on classical poetry. But while these suites are perfectly suited to performance in formal concert halls, they may also be heard in much less regulated settings, such as cafés. Here the listener is free either to concentrate intellectually on the progress and development of the musical ideas or to converse and eat, relaxing in the beauty of the general musical design. The more "oriental" side of Islāmic musical performance is more improvisatory, either in solo performance on a recorder-flute, fiddle, *ʿūd*, or dulcimer or by any of these in combination with the voice—the instrumentalist then elaborating on the singer's improvisation. Here, too, the relationship of audience to performer is much less formal than in the performance of European music.

"Oriental" side of Islāmic musical performance

The Middle Ages. The tradition of sung prayers and psalms extends into the shadows of early civilization.

Such sacred singing was often accompanied by instruments, and its rhythmic character was marked. In the synagogue, however, the sung prayers were often unaccompanied. Ritual dance was excluded from the synagogue as the rhythmic character of sacred music surrendered its more sensual aspects. Even in the prayers themselves, rhythmic verse gave way to prose. The exclusion of women, the elevation of unison singing, and the exclusion of instruments served to establish a clear differentiation between musical performance in the synagogue and that of the street.

The musical performance tradition of the Christian Church grew out of the liturgical tradition of Judaism. The melodic formulas for the singing of psalms and the sung recitation of other scriptural passages are clearly based on Hebraic models.

Music in the Roman Catholic liturgy was performed mainly for the mass. Originally, the music was performed by the priest and the congregation, until, in time, there emerged from the congregation a special group of singers, called the choir, who assumed the musical role of answering and contrasting the solo singing of the priest. Women participated actively in musical performances in the ancient Christian Church until 578, when older Hebraic practices excluding them were restored. From that time until the 20th century, Roman Catholic Church choirs were composed solely of men and boys.

First codification of church music

The first codification of early church music was reputedly made by Pope Gregory I during his reign (590 to 604). Gregory's collection was selected from chants already in use. His codification assigned these chants to particular services in the liturgical calendar. In general it reinforced the simple, spiritual, aesthetic quality of liturgical music. The music in this collection serves as a model of melodic design even in the 20th century and is regarded as one of the monuments of Western musical literature. This school of unison liturgical singing is called Gregorian chant. Specific details concerning the manner in which chant was performed have been lost. There are speculations that the quality of sound the singers employed was somewhat thinner and more nasal than that used by contemporary singers. The authentic rhythmic style of chant cannot be ascertained. There is a theory, however, that the basic rhythmic units had the same durational value and were grouped in irregularly alternating groups of twos and threes. Pitch levels and tempos apparently varied somewhat according to the occasion. There are preserved manuscript notations reminding singers to be careful and modest in their work, indicating that temptations of inattention and excessive vocal display existed for even the earliest liturgical musicians.

While modern musical traditions in the West are based to a large extent on the principles of antiquity preserved in the notated music of the early church, a secular musical practice did exist, but because of the pervasive influence of the church, the dividing line between sacred and secular aspects was thin throughout a good part of the medieval period.

Several types of later secular song have survived. The musical notations are for the most part inadequate to give an accurate impression of the music, but it is known that it retained the essential monophonic character of liturgical music. One curious type of secular song, conductus, originated in the church itself. This song did not use traditional liturgical melodies or texts but was composed to be sung in the liturgical dramas or for processions. For this reason it dealt occasionally with subjects not religious in character. The goliard songs dating from the 11th century are among the oldest examples of secular music. They were the often bawdy Latin songs of itinerant theological students who roamed rather disreputably from school to school in the period preceding the founding of the great university centres in the 13th century.

Several other groups of medieval performers developed literary and musical genres based on vernacular texts: the jongleurs, a group of travelling entertainers in western Europe who sang, did tricks, and danced to earn their living; the troubadours in the south of France and the trouvères in the north; and the minnesingers, a class of artist-knights who wrote and sang love songs tinged with religious fervour.

Instruments, such as the *vielle*, harp, psaltery, flute, shawm, bagpipe, and drums were all used during the Middle Ages to accompany dances and singing. Trumpets and horns were used by nobility, and organs, both portative (movable) and positive (stationary), appeared in the larger churches. In general, little is known of secular instrumental music before the 13th century. It is doubtful that it had a role of any importance apart from accompaniment. Yet the possibility of accompanied liturgical music has not been eliminated by modern scholars.

Medieval polyphony

The medieval musical development with the farthest reaching consequences for musical performance was that of polyphony, a development directly related, as indicated above, to the experience of performing liturgical chant. For performers and performance, perhaps the most important developments in the wake of polyphony were refinements of rhythmic notation necessary to keep independent melodic lines synchronous. At first the obvious visual method of vertical alignment was used; later, as upper voices became more elaborate in comparison with the (chant-derived) lower ones, and writing in score thus wasted space, more symbolic methods of notating rhythm developed, most importantly in and around the new cathedral of Notre-Dame in Paris.

In the 14th century, partly because of the declining political strength of the church, the setting for new developments in music shifted from the sacred field to the secular, from the church to the court. This shift led in turn to a new emphasis on instrumental music and performance. Already the lower voices began to be performed on instruments—both because their long notes made them difficult to sing and because their texts (of only a few syllables) became senseless outside their original liturgical positions. Now, as secular princes became increasingly important patrons of composers and performers—a situation which would continue well into the 18th century—secular and instrumental music flourished. The polyphonic music of the church merged with the poetic art of the troubadours, and the two most important composers of the age were the blind Florentine organist Francesco Landini and the French poet Guillaume de Machaut, canon of Reims.

Most of the music of these composers seems to have been intended for combined vocal–instrumental performance, although this is seldom expressly indicated in the manuscripts. Medieval composers probably had no rigid expectations about performance media. Until the 17th century, and even through the 19th in the case of domestic performance, choice of instruments was likely to be dependent as much on available performers as on anything else. Many sources do, however, indicate that medieval musicians tended to separate instruments into two groups, loud and soft (*haut* and *bas*, or, very generally, wind and string), and to prefer contrasting sonorities within those groups, for maximum differentiation of the individual parts. Outdoor or ceremonial music would be performed with loud instruments (shawm, bombard, trombone, organ); room music, with soft ones (lute, viol, recorder, harp). Paintings and manuscript illuminations of the period show that much secular performance included both a wide variety of bells, drums, and other percussion instruments, and instruments with drones—bagpipes, fiddles, double recorders, hurdy-gurdies. The parts for these instruments are never found in the musical sources and must be reconstructed for modern performance.

Notation of medieval music

The notation of medieval music often is misleading for the modern performer. Accidentals (sharps and flats, called then *musica ficta*) were often omitted as being understood. Further, it seems likely that variation, embellishment, and improvisation were very important elements of medieval performance. It is known that sections of some 15th-century two-part vocal music were enhanced by an extempore third part, in a technique called fauxbourdon; the notation of the 15th-century *basse danse* consisted of only a single line of unmeasured long notes, evidently used by the performing group of three

instrumentalists for improvisation, much as a modern jazz combo's chart.

The Renaissance. The very concept of improvisation as a mere subcategory within performance practice could arise only after the invention of music printing, which had at first little discernible effect on performance. Extemporized ornamentation of polyphonic music continued and increased during the 16th century, in instrumental, vocal, and combined performance, both secular and sacred. Later in the century, liturgical music again became less extravagant in the wake of the Council of Trent (1545–63), which ordered that masses be sung "clearly and at the right speed" and that singing "be constituted not to give empty pleasure to the ear, but in such a way that the words may be clearly understood by all." Music printing was at first too expensive to alter seriously the social structure of musical performance; the traditions of ostentation and exclusiveness embodied in music written by Guillaume Dufay for the early 15th-century Burgundian court were continued in the magnificent musical establishments of the Italian Renaissance princes and popes. Detailed records exist of the elaborate musical festivities arranged for weddings and baptisms of the powerful Florentine family, the Medici. Printing increased the dissemination as well as the survival of these works but, like the earlier Burgundian chanson, and unlike the contemporary Parisian chanson, which was cast in a more popular mould, they were nonetheless primarily intended for a select group of discriminating performers.

Printing, both of music and of books, does document the ever increasing development and sophistication of instrumental music during the 16th century. Printed descriptions of instruments date from the 16th century. Their discussions of tuning and technique supplied the needs of professional and nonprofessional musicians alike. There was a growing tendency to construct instruments in families (whole consorts of homogeneous timbre, high, middle, and low), a tendency perhaps related to recent expansion at both ends of the musical scale: with more space available, contrapuntal parts no longer crossed so frequently and no longer needed the differentiation provided by the markedly contrasting timbres of the medieval "broken consort."

The 17th and 18th centuries. After printing, the next significant influence on music performance was the gradual emergence of the audience, for the relationship between participants in the musical experience—between performer and listener—became polarized. The first evidence for this shift was the rise of the professional vocal virtuoso about the last quarter of the 16th century, and this development soon had a profound influence on musical style. Italian composer–singers, such as Giulio Caccini and Jacopo Peri, reacted quickly to their audiences' desire for more expressive and passionate vocalism, and the music they wrote for themselves eventually was imitated and refined by other composers, such as the Italian Claudio Monteverdi, whose nine successive books of madrigals document the changes in style from music composed for four to six essentially equal voices to music in which the interest lay primarily at the extremes of the texture. The technical underpinning for this new monodic style was the basso continuo, or thorough bass, played by one or more polyphonic solo instruments "realizing" a "figured bass": that is to say, improvising chords above a single line of music provided with numbers and other symbols to indicate the other notes of the chords. In the 17th century a wide variety of continuo instruments was used, including lute, theorbo, harp, harpsichord, and organ. By the 18th century the practice was more standardized: the bass line would be realized on a keyboard instrument and reinforced by a monophonic bass instrument, such as a lute, viola da gamba, violoncello, or bassoon. The continuo player not only completed the harmony but could also control rhythm and tempo to suit the particular conditions of a performance.

Rise of the vocal virtuoso

The development of monody was itself a necessary precondition for that most expensive of all performance institutions, opera. Beginning in Florence at the very end of the 16th century, opera soon spread over Italy: through Rome, where its initially pastoral nature matured into full-blown spectacle, to Venice, where the first public opera theatre opened in 1637. There, although audiences were still aristocratic, opera was dependent upon the sale of admissions rather than royal patronage, and musical performance began to find an entirely new method of economic support.

In the realm of purely instrumental music, the new economy of performance was slower to emerge, but there were many other new developments. By far the most popular Renaissance instrument had been the versatile lute; it served all levels from the merchant's daughter learning the simplest dance melody to the virtuosos. In the 17th century the lute began to yield to keyboard instruments, but the intimate music of the French *clavecinistes* (harpsichordists) was still a clear outgrowth of the precious and evanescent performance style of the lutanist Denis Gaultier (1597/1603–1672). Later, keyboard ornamentation began to be codified in tables of agrément-symbols published with each new collection of music. In Italy composers also were attempting to provide performers with more explicit directions. Contemporary keyboard fingering systems, which used the thumb much less than modern ones, also served contemporary preference for subtlety and unevenness of rhythm. As the century progressed and national styles drew farther apart, there evolved a specifically French tradition of *inégalité:* performing certain evenly written notes unequally, with alternately longer and shorter values.

Predominance of keyboard instruments

A more lasting French development was the first instance of instrumental music consistently performed by more than one player to a part. In 1656, Jean-Baptiste Lully made his orchestra, the Petits Violons ("Little Violins"), abandon the old tradition of free embellishment and drilled them in a disciplined and rhythmically pointed precision which was widely imitated. Simultaneously, the violin and its family, because of their passionate brilliance and versatility, replaced viols as the standard ensemble instruments—especially quickly in Italy, where performance was less sophisticated, less mannered, and less restrained than in France.

In the 18th century national performance styles tended again to merge, except in the case of opera. French opera, which had reached its first height under Lully and had counted among its star performers Louis XIV himself, continued to emphasize ballet and correct declamation more than pure vocalism. In other areas, standardization and codification were the trend. The place of improvised embellishment and variation was further circumscribed, limited in general to such recognized spots as repeated sections in binary and da capo forms, slow movements of sonatas and concertos, and cadences. Instrumental tutors by famous performers were important and widespread.

Foundation of public concerts

The foundation of public concerts increased, and orchestras all over Europe followed the pattern set by the famous ensemble maintained by the Elector of the Palatine at Mannheim, with its standard size (about 25) and new style of performance with dramatic dynamic effects and orchestral devices (*e.g.*, crescendos, tremolos, grand pauses). The Mannheim composers also hastened the decline of the improvised thorough bass by writing out harmonic filler parts for the violas; conducting from the keyboard nevertheless remained standard practice into the 19th century. Meanwhile, entrepreneurial speculation was finally supplanting aristocratic patronage as the economic base for concert activity. Haydn, who had already spent one full career in Austria, in the service of the House of Esterházy, in 1791 began another and more lucrative one in association with the concert manager Johann Peter Salomon—conducting his London symphonies from the piano.

The 19th century. The heyday of the concert artist began before Haydn's first journey to London, and it still shows few signs of ending. It reached its zenith and was the primary factor in all music performance in the 19th century. Mozart and Beethoven were famous concert pianists before they were famous composers, and succeeding generations saw a large number of piano virtuosos travelling over Europe and, later, North and South America.

Some were composers of works for themselves; others were more important as interpreters of other composers' works. The tradition of the star singer was of course much older, and it continued; one new development was that of the claquer, paid by the star for his applause. The independent conductor, as distinct from the first violinist or the continuo player, emerged from the body of the orchestra during the first half of the 19th century, and his development toward the lionized figures of the 20th century was swift. Parallel with this rise came the establishment of many of today's major orchestras: New York Philharmonic (1842), Vienna Philharmonic (1842), Boston Symphony (1881), Berlin Philharmonic (1882), Amsterdam Concertgebouw (1883), Chicago Symphony (1891), London Symphony (1904).

Establishment of major orchestras

The result of the enormous widening of concert activity and of the increasingly international reputations of performers was an even further standardization of performance practice. Eighteenth-century concern with appropriateness and taste in embellishment yielded to emphasis on clarity and evenness of touch, purity of intonation, and accuracy of execution. As composers' scores became increasingly precise, the performers' interpretative decisions were increasingly limited to matters of technique, tempo, rhythmic and dynamic nuance, and personality—a subjectivism justified by the cult of Romantic genius prevalent in 19th-century artistic life. Real improvisation in music would not re-emerge until the 20th century—in jazz. The addition of such mechanical aids and improvements as chin rests and end pins to string instruments (which permitted a wider and more constant vibrato without tiring); valves and extra keys to brass and woodwind instruments (making scales more even and intonation more secure); and double-escapement action, iron frames, and cross-stringing to the piano (which facilitated crisper and surer attack and made both tone and tuning last longer) all had profound influence not only on performance techniques but also on the very sound of the instruments. The most successful new instrumental and vocal teaching methods emphasized virtuosity, brilliance, evenness, and wide range, reflecting a desire to make music more effective for large audiences.

Technical advances in instruments

The rise of the concert artist was seconded by the appearance of the professional music critic, whose influence on performance has been, and is, difficult to assess. At first critics tended to be primarily practicing musicians; later this was less the case. A more tangible residue of 19th-century music performance and one that illustrates how little its basic social structures have changed since then is the large number of concert halls and opera theatres which were built and are still used today.

One final development, the import of which would not be fully realized until the present century, was that of historicism: the active revival of old music. This incipient recognition of the validity of other styles of composition and performance is dated conventionally from the German composer Felix Mendelssohn's 1829 performance of parts of Bach's *St. Matthew Passion*, but it was preceded in a sense by the Concerts of Antient Music (1776–1848) in London. The stated policy of this musical group was not to perform music less than 20 years old (but they often updated the compositions with added brass parts). The revival of interest in the music of Palestrina and Bach, while at first expressed only in terms of 19th-century Romanticism, would pave the way for 20th-century advances and retrenchments in both style and performance.

The 20th century. The major performing institutions of the 19th century have continued into the 20th century with only minimum structural change, except for a rather belated movement toward unionization of personnel; this development has of course improved the performers' lot greatly, while increasing the costs of performance. Unquestionably, the major new influence on 20th-century music performance has been electronics. Broadcasting and recording have widened even further the potential audience for concert artists, at the same time as they have tended to decrease the physical necessity for large new public performance arenas. Electronic instruments have appeared, both amplified versions of older ones (guitar, piano, and even some woodwinds) and instruments with fundamentally electronic means of tone production (electronic pianos and organs, the theremin and Ondes Martenot, and sound synthesizers). Other new compositional and performance possibilities have also developed, for example, film, tape, stereophonism, and computers. Even before the phonograph (invented *c.* 1875) had begun to be regarded as more than a toy, serious research into the authentic performance of older music had produced an awareness of possibilities which pointed the way out of the seeming dead end of late Romantic gigantism and subjectivism. From the very beginning of the century, the chamber concerts given by Arnold Dolmetsch and his family, on reconstructions of old gambas and recorders, attracted attention to small ensembles and different sonorities and encouraged the activities of other artists.

Appearance of electronic instruments

The true end of the Romantic era and the beginning of the modern era can be dated from the second decade of the 20th century, the time of the composition of two masterpieces which more than any others mark the departure from 19th-century performance ideas: the German composer Arnold Schoenberg's *Pierrot Lunaire* (1912) and the Russian composer Igor Stravinsky's *Histoire du soldat* (*The Soldier's Tale*, 1918). These are chamber works, but their instrumental makeup is a unique mixture of instruments that do not necessarily blend and which seem further to repudiate the orchestra as a performing medium. *Pierrot* is a series of songs which repudiate the 19th-century lied: the voice does not sing but produces a kind of pitched speech (*Sprechstimme*). *Histoire* repudiates both orchestra and opera as previously understood: it is specifically (and inexpensively) designed for performance on a portable stage by three dancers, a narrator, and seven instrumentalists. For these works a new kind of performer is required, and these works in turn have helped to train the new performer—who might be called the group-virtuoso. Teams or groups of such performers have sprung up everywhere. Often they are centred on a living composer or the university where he teaches, where they function as partners to his composition, realizing the work rather than interpreting it. The performer is very much involved in the creative act, the product of which reflects his particular skills and personality, and the dynamics of the working situation. In the second half of the 20th century the important musical action is no longer in the recital nor in opera or symphony; it is in the combo, experimental or folk, live electronic or improvisational, or multimedia.

Prospects. The prospects for musical performance remain as rich and unpredictable as ever. Technology continues to alter the situation considerably through the invention of new sound resources and the creation of new performing situations.

The established performance situation is moving from the formal, ritualized event of the past to a more informal and spontaneous type of gathering. The interaction of various media is leading to new art forms and circumstances. Many artists are attempting to create performance situations that will actively involve as participants all those in attendance. The roles of composer, performer, and listener are consolidated in a single participant, who in interaction with others arrives at an art work, which all have invented, realized and perceived, and which can never take place again.

Increasing spontaneity of performance situations

On the other hand, the growing use of technology may intensify the problem of evaluating the meaning and effect of electronically produced and assembled performances that, in their totality, never took place at all and possibly never could.

The electronic media may continue to improve until soon anyone may be able to select chamber, concert, opera, and other new types of performance from anywhere in the world, experiencing them through nearly lifelike reproduction facilities. Vast numbers of people may study performance skills via two-way transmission with great artists. The number of actual public performing events may decrease as private musical performance increases. Already there is the phenomenon of the wide-

spread dissemination of great performers' recordings, which has forced the standards of quality for a live performance to almost inhuman heights and has increased interest in the performance of older as well as contemporary music.

HISTORY OF NON-WESTERN MUSICAL PERFORMANCE

South Asia. Although the classical South Asian or Indian musician usually performs in a concert situation quite analogous to that of a typical Western artist, his audience responds to him quite differently: he is not judged on how faithfully he reproduces the music the composer imagined but on how well he creates his own music within certain wide bounds set by the composer and by the general practice of Indian music. Since Indian musical performance is based on improvisation, Indian musical pedagogy is therefore a more personal procedure, in which an aspiring musician will "apprentice" himself to a guru, with whom he thereafter identifies himself; in the West this kind of organization is reflected in the rise of the group-virtuoso discussed above. Similarly, Western development away from large performing groups such as the full orchestra reflects—or at least parallels —the more intimate character of Indian music, the basic texture of which usually involves a quite small group of performers: one player to provide rhythm on a drum such as the double-headed, pitched *ṭablah;* one to provide a basic drone, often on the lute-like tambura; and a central performer on the sitar (technically also a plucked lute but one with melodic capability, unlike the tambura). The players often engage in a kind of competition not unlike that of Western jazz groups. If there is singing, the style of performance is low and soft, in contrast to that of Indonesian classical vocalism.

Intimate character of Indian music

Southeast Asia. The gamelan is at the center of the art-music tradition of Indonesia. It may range in size from a few to over 75 instruments. The basic melodic instrument is the *saron* (bronze xylophone), accompanied by various gongs, a kind of violin, a recorder-flute and/or a zither; the group is led by a drummer. As in medieval Western music, there are two kinds of gamelan playing, one emphasizing the bronze instruments (comparable to medieval *haut,* or loud, consorts) and the other the wind and string instruments (*bas,* or soft, groups). A similar differentiation exists in Indochinese music, in the contrast between the percussion-dominated *pi phat* band of Thailand and the string-dominated *mahori* bands of Thailand and Cambodia. Gamelan playing, particularly of the softer type, often accompanies solo and unison choral singing of classical poetry (music is connected with most of Indonesian literature). Southeast Asian vocal performance—like that of a great deal of non-Western art music—is characterized by tense, high, often nasal voice production; this is one of many alternatives being explored by the more experimental 20th-century Western composers and performers.

China and Japan. The most extensively developed and most important Chinese and Japanese traditions of musical performance are closely tied to theatrical styles and traditions. Perhaps the most spectacular of non-Western performance traditions is Chinese opera, in which singers, acrobats, costumes, scenery, and instruments are combined in the creation of a highly varied work of art. Peking opera uses two basic kinds of instrumentation: for military scenes a battery of drums, gongs, and cymbals with a kind of oboe playing the melody; for the more frequent domestic scenes a wider variety based on a drum (*pan ku*) with a peculiarly sharp, cracking sound for keeping time, and a number of two-stringed, bowed lutes played with the bow passing between the strings. Plucked lutes and flutes also appear at times. All of the melody instruments play heterophonically with the singers, whose vocal style, as in the West, is highly artificial. Heroines are usually portrayed (sometimes by female impersonators) in a high, thin voice; heroes use a raucous, rasping tone quite foreign to traditionally oriented Western ears—but, again, not unlike some of the vocal techniques required by 20th-century Western avant-garde composers.

A performance tradition peculiarly Japanese is the emphasis on the visual aspects of making music: custom directs that *gagaku* (court orchestra) instruments must be played as gracefully as possible.

BIBLIOGRAPHY. While any of the standard histories of music or of specific musical style-periods will consider performance at least implicitly, the best direct and concise account of the whole problem is THURSTON DART, *The Interpretation of Music* (1954). Other general views of the subject are FREDERICK DORIAN, *The History of Music in Performance: The Art of Musical Interpretation from the Renaissance to Our Day* (1942); and ROBERT DONINGTON, *The Interpretation of Early Music,* 2nd ed. (1965), which, like Dart, includes a bibliography of sources. What bibliographic aids to individual performers exist are given in "Dictionaries and Encyclopedias of Musical Instruments, Makers, and Performers," in VINCENT H. DUCKLES, *Music References and Research Materials: An Annotated Bibliography,* 2nd ed., pp. 40–50 (1967). Certainly the most extensive bibliography ever published on the subject is MARY VINQUIST and NEAL ZASLAW (eds.), *Performance Practice: A Bibliography* (1971).

Some book-length studies of particular aspects of musical performance are listed below: P. ALDRICH, *Rhythm in Seventeenth-Century Italian Monody* (1966); F.T. ARNOLD, *The Art of Accompaniment from a Thorough-Bass As Practised in the XVIIth and XVIIIth Centuries* (1931); J.M. BARBOUR, *Tuning and Temperament* (1951); D.D. BOYDEN, *The History of Violin Playing, from Its Origins to 1761 and Its Relationship to the Violin and Violin Music* (1965); WALTER EMERY, *Bach's Ornaments* (1953); R.E.M. HARDING, *Origins of Musical Time and Expression* (1938); WILFRID H. MELLERS, "Theory and Practice," in *François Couperin and the French Classical Tradition* (1950); FRITZ ROTHSCHILD, *The Lost Tradition in Music: Rhythm and Tempo in J.S. Bach's Time* (1953); and *Musical Performance in the Times of Mozart and Beethoven: The Lost Tradition in Music, Part II* (1961); DENIS W. STEVENS (ed.), *The Art of Ornamentation in the Renaissance and Baroque* (1967), a stereophonic record; HENRY PLEASANTS, *The Great Singers: From the Dawn of Opera to Our Own Time* (1966); and WILLIAM P. MALM, *Music Cultures of the Pacific, the Near East, and Asia* (1967).

(L.Fo./J.Th./B.A.C.)

Musical Rhythm

In its most general sense rhythm (Greek *rhythmos,* derived from *rhein,* "to flow") is an ordered alternation of contrasting elements. Though this article is confined to rhythm in music, the notion of rhythm occurs in other arts (*e.g.,* poetry, painting, sculpture, and architecture) as well as in nature (*e.g.,* biological rhythms).

Attempts to define rhythm in music have produced much disagreement, partly because rhythm has often been identified with one or more of its constituent, but not wholly separate, elements, such as accent, metre, tempo. As in the closely related subjects of verse and metre, opinions differ widely, at least among poets and linguists, on the nature and movement of rhythm. Theories requiring "periodicity" as the *sine qua non* of rhythm are opposed by theories that include in it even nonrecurrent configurations of movement, as in prose or plainsong.

Elements of rhythm. Unlike a painting or a piece of sculpture, which are compositions in space, a musical work is a composition dependent upon time. Rhythm is music's pattern in time. Whatever other elements a given piece of music may have (*e.g.,* patterns in pitch or timbre), rhythm is the one indispensable element of all music. Rhythm can exist without melody, as in the drumbeats of primitive music, but melody cannot exist without rhythm. In music that has both harmony and melody, the rhythmic structure cannot be separated from them. Plato's observation that rhythm is "an order of movement" provides a convenient analytical starting point.

Beat. The unit division of musical time is called a beat. Just as one is aware of the body's steady pulse, or heartbeat, so in composing, performing, or listening to music one is aware of a periodic succession of beats.

Tempo. The pace of the fundamental beat is called tempo (Italian "time"). The expressions slow tempo and quick tempo suggest the existence of a tempo that is neither slow nor fast. This "moderate" tempo is often assumed to be that of a natural walking pace (76 to 80

paces per minute) or of a heartbeat (72 per minute). The expression of tempo as a given number of beats to the minute resulted from the work of Johann Mälzel, who in 1816 perfected the metronome. The tempo of a piece of music indicated by a composer is, however, neither absolute nor final. In performance it is likely to vary according to the performer's interpretative ideas or to such considerations as the size and reverberation of the hall, the size of the ensemble, and, to a lesser extent, the sonority of the instruments. A change within such limits does not affect the rhythmic structure of a work.

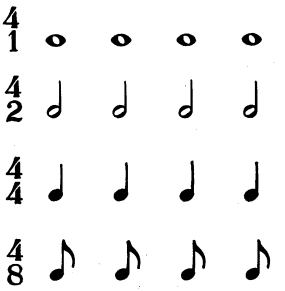
Rubato. The tempo of a work is never inflexibly mathematical. It is impossible to adhere in a musical manner to the metronomic beat for any length of time. In a loosely knit passage a tautening of tempo may be required; in a crowded passage a slackening may be needed. Such modifications of tempo, known as *tempo rubato*—*i.e.*, "robbed time"—are part of the music's character. Rubato needs the framework of an inflexible beat from which it can depart and to which it must return.

Time. The mind apparently seeks some organizing principle in the perception of music, and if a grouping of sounds is not objectively present it imposes one of its own. Experiments show that the mind instinctively groups regular and identical sounds into twos and threes, stressing every second or third beat, and thus creates from an otherwise monotonous series a succession of strong and weak beats.

In music such grouping is achieved by actual stress; *i.e.*, by periodically making one note stronger than the others. When the stress occurs at regular intervals, the beats fall into natural time measures. Although in European music the concept of time measures reaches back to a remote age, only since the 15th century have they been indicated by means of bar lines. Thus, the terms measure and bar are often used interchangeably.

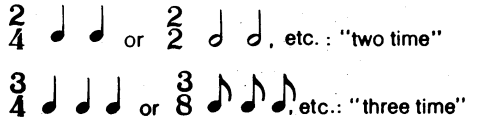
Time signatures

The time measure is indicated at the opening of a piece by a time signature; *e.g.*, $\frac{2}{4}$, $\frac{4}{8}$, $\frac{3}{4}$, $\frac{6}{8}$. The length of each beat in a measure may be a time unit of short or long duration:



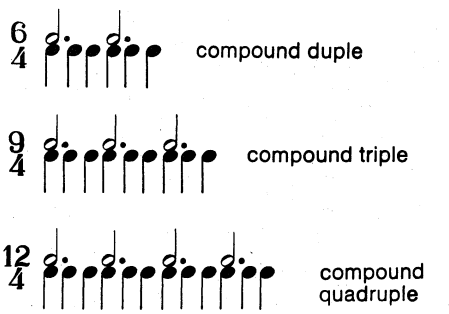
The signature $\frac{4}{1}$ (above) means that the whole note (1) is the unit in each measure, and there are four (4) of them to each measure. In the second illustration, $\frac{4}{2}$, the half note (2) is the unit of measurement, with four of them (4) to each measure, etc.

The two basic types of time measure have either two or three beats and admit of many different notations.



"Four time," or "common time," is really a species of duple time allied to "two time," as it can hardly be thought of without a subsidiary stress at the half measure; *i.e.*, on the third beat; thus:

Duple, triple, and quadruple time measures—*i.e.*, those in which there are two, three, and four beats to a measure—are known as simple time. The division of each of the component beats into three produces compound time:



More complex times, such as the quintuple, $\frac{5}{4}$, usually fall into groups of 3+2, as in "Mars" from Gustav Holst's suite *The Planets* and in the second movement of Tchaikovsky's *Sixth Symphony*. Rimsky-Korsakov, in *Sadko*, and Stravinsky, in *Le Sacre du printemps*, use 11 as a unit. Ravel's piano trio opens with a signature of $\frac{8}{8}$ with the internal organization 3+2+3. Folk song and folk dance, particularly from eastern Europe, influenced the use of asymmetrical time measures, as in the "Bulgarian Rhythm" pieces in $\frac{7}{8}$ and $\frac{5}{8}$ in Bartók's *Mikrokosmos*.

Metre. The combinations of long (—) and short (∪) syllables are known in prosody as feet. The system of notating the musical equivalents of feet derives from the application of prosody to music. The foundations for European music were laid in ancient Greece, where classical music and poetry were regarded as parts of a single art. These principles were adopted by the Romans and were transmitted, by way of Latin poetry, to medieval Europe. The feet of classical poetry and their equivalents in music are shown in Table 1. And in late antiquity St. Augustine (354–430), in *De musica*, added more.

Prosody applied to music

Table 1: Classical Poetic Metre

	metrical foot	poetic syllables	musical notation	groups of beats	equivalent time signatures
duple	dactyl	— ∪ ∪	♩ ♫	2 + 2	$\frac{2}{4}$
	anapest	∪ ∪ —	♫ | ♩	2 + 2	$\frac{2}{4}$
triple	trochee	— ∪	♩ ♪	2 + 1	$\frac{3}{8}$
	iamb	∪ —	♪ | ♩	1 + 2	$\frac{3}{8}$

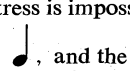
Rhythmic modes. Until the 12th century, church music was virtually limited to unadorned plainsong. The early composers found that polyphony required a rhythmical organization to keep the parts together; so rhythmic modes or metres were adopted (see Table 2).

Table 2: Classification of Rhythmic Modes

poetry	metrical foot	musical notation
— ∪	trochee	♩ ♪
∪ —	iamb	♪ ♩
— ∪ ∪	dactyl	♩. | ♪ ♩
∪ ∪ —	anapest	♪ ♩ | ♩.
— —	spondee	♩. | ♩.
∪ ∪ ∪	tribrach	♫

Compared with a hypothetical flow of beats equal in stress, metre adds significance to what was merely a forward flow in time—though the continuation of a metrical pattern may itself become monotonous. Thus, metre, though "rhythmic" by comparison with pulse, is not the whole of rhythm.

The 13th-century musicians often varied the rhythmic modes by combining several of them simultaneously in different parts of polyphonic composition.

Polyphonic metre. Theoretically, metre appears to be without stress accent, and certainly much polyphonic music of a later period, such as the masses of Palestrina, has an almost stressless flow. Yet despite the absence of strong and weak accents, these works reveal a subtle rhythmical organization. At a later period metre and time measure cannot be wholly separated. In their "purest" forms they may be extremes, but in music predominantly of one type, the other element is rarely wholly absent, though on an instrument such as the organ, actual dynamic stress is impossible. After all, metres like the spondee, , and the dispondee,  , need an accent on the first beat to keep their identity. Notwithstanding the opposite tendencies of metrical organization and stress accent, however, some metre is obviously subject to stress, so that metre and time measure become very closely linked, as in the scherzo of Beethoven's *Ninth Symphony*, where a measure has a strong first beat and at the same time follows a metre.

Organic rhythm. In broad terms, the time framework of music comprises tempo, time measure, metre, and period; and its rhythmical life hangs on rubato, musical motif (which may already include cross accent), and metrical variation, as well as on asymmetry and balance of phrase.

Rational and irrational features

Whereas the former are more or less mechanical factors, measured and rational, the latter are organically inspired and numerically irrational—the very life of the music.

Prose rhythms and plainsong. Rhythm is, therefore, not any one of these rational or formal features, nor is it composed solely of a combination of these factors. Yet rhythm requires the background of a rational framework in order that it may be fully perceived, but this framework need not embrace all the rational factors described above. Thus plainsong, as it is known in modern times, makes no use at all of measure or of regular metre but is supremely rhythmical in conception; its "free" rhythms are felt. Whereas so much music has for its framework a regular repetition of underlying accent, whether stress or durational, the framework of plainsong is irregular. Its rhythm belongs to the Latin tongue and springs from the correct accentuation of the text and the dynamic quality inherent in the word grouping.

Rhythm and melody. Thus far, music's structure in time has been examined separately from its structure in tone, but no such separation is really possible. Melody and rhythm are intimately connected. Moreover, various styles of music tend to standardize their melodic cadences and, with them, their time divisions (*e.g.*, Mozart's melodic rhythm is much more regular than Prokofiev's).

Rhythm and harmony. In music employing harmony, the rhythmic structure is inseparable from harmonic considerations. The time pattern controlling the change of harmonies is called harmonic rhythm. In 17th- and 18th-century music, harmony tends to limit rhythmic subtleties and flexibility of the melodic elements (as well as determining the basic type of melody), in regard to stress accents. It is, therefore, no accident that the polyphonic music of Indonesia and Southeast Asia, like much European music, exhibits certain four-square melodic tendencies. By contrast, the music of India and the Perso-Arab world employs a melody instrument or voice performing in a given metre offset by a drum playing cross rhythms or (in the Arab world) a quite different metre; with no harmony (except a drone) to impede its flow, the rhythm can reach a structure of great subtlety and complexity.

Diversity from conceptual differences

Rhythm, structure, and style. In European music the great variety of styles derives its relation to melody from different concepts of rhythm. They include the strict rhythmic modes of the 13th century; the free oratorical speech-rhythms of the Renaissance; the almost stressless flow of Renaissance polyphony; the strong body rhythms of the Baroque; the freedom of the late Romantics; and the primitivistic rhythms of the 20th-century with composite and ever changing time signatures

Thus, study of musical history shows a varying attitude toward rhythm, sometimes closer to strict rule, sometimes to "freedom," as the temper of the times and the relative influence of poetry, dance, and folk music decree. Plato's definition of rhythm as "an order of movement" might, therefore, be expanded. As a determining factor in the vitality of music, rhythm may be described as "an inspired, organic order of movement," communicating intelligibly to the senses. From the analytical viewpoint, it operates in the rational framework described, which it varies in terms of rubato, motif, etc. Ultimately, rhythm is the organic process of music in time; it is music's direction in time. The quality of rhythm is the quality of life; however vitally the composer conceives his music, he must depend upon the performer to recreate it rhythmically.

BIBLIOGRAPHY. CURT SACHS, *Rhythm and Tempo* (1953), the most comprehensive work on rhythm in music, ranges over many non-Western cultures as well as over the successive periods of Western musical history. Detailed rhythmical analyses of Western music since the 17th century appear in GROSVENOR W. COOPER and LEONARD B. MEYER, *The Rhythmic Structure of Music* (1960). Studies of special periods are available in CHARLES F. ABDY WILLIAMS, *The Aristoxenian Theory of Musical Rhythm* (1911); W.F. JACKSON KNIGHT, *St. Augustine's De Musica: A Synopsis* (1949); WILLIAM G. WAITE, *The Rhythm of Twelfth-Century Polyphony, Its Theory and Practice* (1954); PHILIP F. RADCLIFFE, "The Relation of Rhythm and Tonality in the Sixteenth Century," *Proc. R. Musical Assn.*, 57:73–97 (1931); and HENRY D. COWELL, *New Musical Resources* (1930, reprinted 1969). Particular applications of rhythm have been studied in CHARLES F. ABDY WILLIAMS, *The Rhythm of Song* (1925); and WILLIAM THOMSON, *The Rhythm of Speech* (1923). Aesthetic aspects are considered in MARGARET GLYN, *The Rhythmic Conception of Music* (1907); and MATHIS LUSSY, *Le Rythme musical*, 3rd ed. rev. (1897; abridged Eng. trans., *A Short Treatise on Musical Rhythm*, 1909). EMILE JACQUES-DALCROZE, *Le Rythme, la musique et l'éducation* (1920; Eng. trans., *Rhythm, Music and Education*, 1921), is the pioneer work in its aspect of the field; musical rhythm has been put in wider perspectives by ELSIE FOGERTY in *Rhythm* (1937).

(P.C–H.)

Music Theory

A theory of music comprises acoustics, aesthetics, and stylistic practice. The musical theories of a culture reflect not only its attitude to the arts but also its religious, philosophic, and scientific biases. China, India, Greece, and Islām developed musical theories long before the modern era, and western Europe has, since medieval times, contributed extensively to this multiple heritage.

Acoustics connects with instrument design and musical-scale derivation and also with number theory and metaphysical symbolism. Aesthetics derives from the practice of artists and also from religious and philosophic theories of universal and social order. Although musical practice generates theories of composition and performance, fundamentally it springs from the religious, philosophic, and scientific thought of the culture that engendered it.

Chinese music theory, based acoustically on a system of relations of the numbers 2 and 3, connects symbolically with metaphysical theories of the interaction of Yin (2) and Yang (3), the two principal contrasting elements of the Chinese cosmos; and, by dint of a theory claiming causal effects on human behaviour for musical sounds, it also connects with ethical, social, and political teachings.

Indian music, based acoustically on numbers 2, 3, and 5, yields many seven- to nine-tone scales (called *rāgas*), each with associated melodic patterns and each causing a specific emotional effect (called *rasa*). Varying interval tensions are produced against a fixed drone by the tones of the scale. The fixed pitches of the drone symbolize the *ātman* (soul; the individual's participation in the universal Brahman), and the changing pitches of the *rāga* represent life experiences. Indian music, therefore, is connected by tradition with meditation and self-transcendence. Other non-Western music is similarly based upon such religious and philosophical foundations.

Western European music is characterized by counterpoint and harmony and by its constant stylistic evolution.

Counterpoint refers to the simultaneous unfolding of several melodic lines. The combinations of tones that result are harmonies, classified as consonant (bearing harmonic repose) and dissonant (bearing harmonic tension) and arranged since the 17th century in hierarchical families called tonalities (relationships between tones). Since the Renaissance, Western music has increasingly emphasized perceptual factors other than pitch and rhythm, until, in the 20th century, types of music have emerged in which pitch and rhythm no longer predominate. The evolution of style is currently expressed in a wide diversity of kinds of music without a common practice. Also radically dissimilar are the acoustical and organizational properties in the music of contemporary composers. Hence, a theory of contemporary Western music would need to be unusually comprehensive.

ORIGINS AND HISTORY OF WESTERN EUROPEAN MUSIC THEORY

The origins of western European music lie in the music of the medieval Roman Catholic Church. The liturgical plainchant adopted by the Roman Church was codified by Pope Gregory I in the 6th century and was based upon a system of seven-tone modes (scale formations) handed down in Greek tradition from Pythagoras. The tuning of these modes was based, like traditional Chinese music, on relationships of the numbers 2 and 3, symbols of the worldly and the divine. Gregorian chant was monophonic (a single melodic line without accompaniment) until near the end of the 1st millennium AD, when a polyphonic practice (*i.e.*, combining several voice parts simultaneously), called organum, was introduced. From this early counterpoint, at first almost entirely a matter of parallel motion in fourths (*e.g.*, F above C), fifths (*e.g.*, G above C), and octaves, the unique polyphonic tradition of western European music gradually evolved.

Harmony of the universe

The Roman scholar and philosopher Boethius, writing in the 5th century, distinguished three kinds of "music," *musica mundana, musica humana,* and *musica instrumentalis.* Of these, only the last refers to sounded music, the first two indicating, respectively, the arithmetically expressed harmony of the universe (a Pythagorean idea) and the harmony of the soul. By the correspondence between these harmonies, music was believed to symbolize man's place in the universe, and the potent psychological impact of music was to some extent explained. This conception survived for a long time, especially in the tabulation of intervals (distances between notes) in terms of their relative consonance and dissonance, measured as fractions of string length on a monochord needed to produce the intervals. Considerations of this kind were especially crucial in polyphonic music, as were also the methods of rhythmic integration. Thus, medieval and Renaissance treatises usually concerned problems of consonance and dissonance, techniques for integrating polyphonic voices, and methods of notating pitch and rhythm. The most important of these for the subsequent development of music was the 14th-century treatise of a French musical theorist, Philippe de Vitry, *Ars nova,* which provided a new rhythmic notation liberating music from the rigidity of earlier methods. During the 13th and 14th centuries the tuning system shifted gradually away from the Pythagorean toward a triadic system later called "just intonation," as intervals of a third (*e.g.*, C–E) and a sixth (C–A) came to be recognized as consonances.

At the end of the Renaissance, as instrumental and operatic music began to replace choral music, the practice of accompanying instruments and voices with an improvised keyboard part, called a "figured bass," led to the organization of compositions around harmonic progressions. This had two important effects: to simplify the traditional system of modes into major and minor tonality and to make urgent the need to temper the tuning of the scale so as to provide a wide scope for modulation (change of key within a work) with a limited gamut of pitches.

Rameau and others. The French composer and theorist Jean-Philippe Rameau went far to establish tonality as a distinct technique of composition. He conceived each triad (a chord formed by two superimposed thirds, such as C–E and E–G) as representing the acoustic proportion 4:5:6 and considered a given triad to keep its identity regardless of the distribution of its three notes. Each triad thereby has a "root" position, a "first inversion," and a "second inversion," depending upon which of its tones is in the bass. Thus, the triad C–E–G can assume the forms E–G–C and G–C–E.

Another aspect of theory that gradually acquired importance during the 18th and 19th centuries was the codification of stylistic and compositional techniques. Such was *Gradus ad Parnassum* (1725), a treatise on counterpoint by the Austrian theorist Johann Joseph Fux, which had a long-standing effect upon pedagogy.

In the 19th century the emphasis in this area shifted to techniques for musical analysis. The theories of phrasing and functional harmony of the German musicologist Hugo Riemann gave tonal and rhythmic practices of the "common-practice period" increased coherence.

Then, an Austrian theorist, Heinrich Schenker, gave the process of tonality its most thoroughgoing formulation. In Schenker's view, pitch (the location of sound in the tonal scale), as both melody and harmony, is the controlling factor in tonal music. Rhythm and dynamics function secondarily to articulate the structures created by pitch design. All tonal compositions can be reduced to a melody, a bass line, and a harmonic progression. In Schenker's view, all modulation should be regarded as "tonicization" of various subordinate progressions within one overall key. Schenker's ideas, designed to apply to common-practice-period music, have been applied to earlier and more recent music with notable success, allowing for some generalizations of his basic concepts.

Modern theorists. Some 20th-century theorists have devoted themselves to the codification and teaching of compositional or theoretical techniques. A small number have provided bases in theory for the music of their times.

Schoenberg's theories

Arnold Schoenberg's 12-tone serial theory provides an alternative method of pitch organization to tonality, in which the traditional rules of tonality and harmony are replaced by new rules and principles. This system has been used widely by many 20th-century composers. Serial techniques have been applied to factors other than pitch, notably among the "post-Webern school" of composers (Anton von Webern was a pupil of Schoenberg).

The theoretic work of the German composer Paul Hindemith was directed toward the continuation of the tradition of tonality into post-triadic music. Using the concept of consonance–dissonance as expounded by a 19th-century German acoustician, Hermann von Helmholtz, Hindemith provided a method of classifying harmonies by computing the frequencies of vibration interferences produced by their constituent frequencies. His most original contribution was his use of consonance–dissonance as a controlling factor in phrase design.

The most influential of the concepts of the German composer Karlheinz Stockhausen is the breaking down of the field of musical perception into controlling and dependent factors. Beginning with the concept of total organization of all parameters by the composer (*e.g.*, pitch, rhythm, dynamics, tone colour, stress, and all other facets of sound), Stockhausen proceeded to a fluid concept of musical control, allowing for much indeterminacy.

The theories of the U.S. composer John Cage are largely aesthetic, with philosophical and religious (specifically, Zen Buddhist) overtones. Insisting that all sounds and also silence itself are materials for music, he developed a theory that places duration at the centre of musical organization. This was followed by an interest in chance and indeterminancy as valid modes of composition.

The theories of the Greek composer Yannis Xenakis concern the use of probability theory and statistical distribution in composition. His interest in applying mathematical models led him to use computers extensively.

APPLICATION OF MATHEMATICS TO MUSIC

By far the oldest and most widespread application of mathematics to music is the derivation of pitch scales and rhythmic proportions by means of whole number fractions referring either to frequency or to tempo ratios. The

Chinese, the Indian, the Islāmic, and the ancient Greek pitch systems were derived in this way. The medieval western European pitch system was similarly based upon modes derived from Pythagorean traditions. Only with the introduction of temperament (the system of tuning in which the intervals deviate from the "pure") in Europe during the Baroque era was this kind of system abandoned. Equal temperament was proposed at about the same time in China but was rejected. In spite of the adoption of equal temperament in the West in the 18th century, it was not until the 20th century that compositional techniques of pitch organization, such as serialism, which are not based on proportional tuning, made their appearance. Not only pitch scales but also laws of harmony, in particular the classification of intervals in terms of consonance and dissonance, derive from proportional tuning. Furthermore, the modern metrical system of rhythmic notation, from its inception in the 14th century, is based upon the same mathematical procedure.

Triadic tuning. The lists of intervals in their order of consonance, given by most medieval and Renaissance theorists after the 9th-century French monk-musician Hucbald, reflect the kind of tuning system favoured by the musicians of the time. The shift from Pythagorean to triadic tuning is reflected in the greater consonance allowed to thirds and sixths by later theorists. That the acceptance of new consonances was accompanied by a shift in methods of tuning can be verified by the monochord experiments detailed by theorists. From these it is clear, for instance, that the precise size and, hence, tuning of the major third dealt with by earlier theorists differed from that which later came to be preferred. Early medieval theorists, following the Pythagorean system, derived the proper tuning of the third (as well as that of the entire diatonic scale) by tuning a series of fifths: as C–G, G–D, D–A, A–E, the resulting notes C and E then forming a major third. This process yielded a sound-frequency ratio of $^{81}/_{64}$ as the relation of top to bottom pitch. Later theorists adopted a major third the frequency ratio of which was the simpler $^{5}/_{4}$, a proportion found in the harmonic series or component vibration pattern of vibrating strings and air columns. For its relation to the natural harmonic series, this later third was known as a "just" third. It differed in size only minutely from the earlier third, by a ratio of $^{81}/_{80}$ ($^{5}/_{4} = {}^{80}/_{64}$; *cf.* $^{81}/_{64}$ for the Pythagorean third), but its sound was generally regarded as more consonant. Although $^{5}/_{4}$ involves a higher prime number (81 is factorable into $3 \times 3 \times 3 \times 3$) than previously allowed in a consonance, the ear superseded previous theory (but was given legitimacy by invoking the physics of vibrating strings).

Importance of sixths and thirds

This example may shed some light upon how one musical system succeeds another. The simplest pentatonic scale, one used the world over, is obtained by tuning four successive fifths. If, in the interest of filling in the two largest intervals of this scale, two more fifths are added to the succession, the result is a seven-tone diatonic scale—*i.e.*, one consisting of five whole and two half steps. The Pythagorean thirds and sixths of this scale are dissonant. If, in order to obtain smoother thirds and sixths, the system of just intonation is adopted, an alternative tuning of the seven-tone scale becomes available. These scales received letter or syllable names in India and in Europe.

The diatonic scale

Although two chromatic inflection symbols (♯ and ♭) were introduced to take care of additional fifths or thirds added to the scale, as in forming the 12-tone chromatic scale from a combination of the major and minor seven-tone diatonic scales, no such pair of symbols was introduced to distinguish in notation between Pythagorean and just thirds and sixths. Such a pair of symbols (+ and −) would refer to the syntonic comma ($^{81}/_{80}$, or about one-tenth of a tone), an interval too small to be easily perceptible in general usage, while ♯ and ♭ refer to a much larger interval (about a half tone).

Modern practice. While the common practice period took for granted the permanence of both the tuning system of 12-tone equal temperament and the harmonic and melodic tradition of triadic tonality, the 20th century has called both of these practices into question. The condition of harmonic practice at the end of the 19th century resembled, in some ways, the condition of polyphony in the 13th century, when the use of thirds and sixths began to become widespread. During the last half of the 19th century, extended tertian chords (*i.e.*, derived by the superimposition of thirds) such as sevenths, ninths, 11ths, and 13ths (comprising three, four, five, and six superimposed thirds) became increasingly prevalent. In both cases, dissonances resulting from combinations of consonances became so prevalent as to demand acceptance as consonances. In the earlier case, a simpler tuning system was replaced by a more complex one; in the recent one, many composers abandoned harmonic organization according to the principle of consonance–dissonance. This coincided with a rejection of the tradition of tonality. In order to provide an alternative system of ordering pitch in atonal music, Schoenberg, the first to grapple with this problem, developed the tradition of thematic unity into an all-pervasive, 12-tone serialism. Serialism proved useful not only for pitch but for almost any materials capable of manipulation by the composer. As John Cage pointed out, the ultimate application of atonal thinking is the abandonment of tone as a controlling factor.

At the same time, there have been efforts to devise a new pitch system in place of the old one. The most prevalent of these new systems have been those based upon tempered subdivisions of the 12-tone equal-tempered scale. In other cases, a new pitch system has been based upon some kind of just intonation. The best known such system is that of a U.S. composer, Harry Partch, whose 43-tone scale is based upon sound frequency ratios expressible as fractions utilizing the prime numbers two, three, five, seven, and eleven. A less-complex system than Partch's is that devised by the Dutch theorist Adriaan Fokker, associated with a 31-tone tempered scale.

Much music since 1950 has dealt principally with schemes of organization based on less-precise kinds of design. For this reason, the period since 1950 has seen extensive experiment with notational methods designed to communicate these less specific modes of order. Such problems as the use of chance and indeterminacy in composition and performance or the use of statistical ordering procedures raise basic questions regarding the kind and degree of the order in music. It is because of the widespread concern about these and related matters that aspects of communications theory such as information theory and systems analysis are finding a role in mid-20th-century music theory.

Use of the chance factor

As the world's cultures become aware of each other's traditions through improved communications, the need increases for theoretical tools and concepts sufficiently versatile and nonrestrictive to show the relationship between very diverse kinds of thinking. The exploration of the nature of human intelligence affords one of the most fruitful sources for concepts useful in this endeavour. For this reason, many aesthetic theorists have sought to express their concepts mathematically, in the form of algorithms capable of producing or recognizing specific kinds of systematic order via computers. (An algorithm is a procedure by which one class of problems can be solved and in which the initial state is connected with the final state by an uninterrupted succession of intermediate states, each of which is generated by its predecessor.) The necessity to be absolutely specific about every detail of a computer program is an excellent way of discovering in any discipline exactly what assumptions have been made besides those that are explicitly recognized. Aesthetic behaviour is an especially revealing facet of human activity, connecting as it does sensory-motor, intellectual, and emotional modes of behaviour. Music especially involves activities organized by means as various as improvisation and mathematics.

BIBLIOGRAPHY. J. MURRAY BARBOUR, *Tuning and Temperament: A Historical Survey* (1951), a comprehensive historical survey of systems of tuning in Europe from Greek antiquity onward, with emphasis on Renaissance and Baroque periods; HERBERT BRUN, *Über Musik und zum Computer* (1971), on the development of a conceptual framework wherein the composition of music and the programming of computers ap-

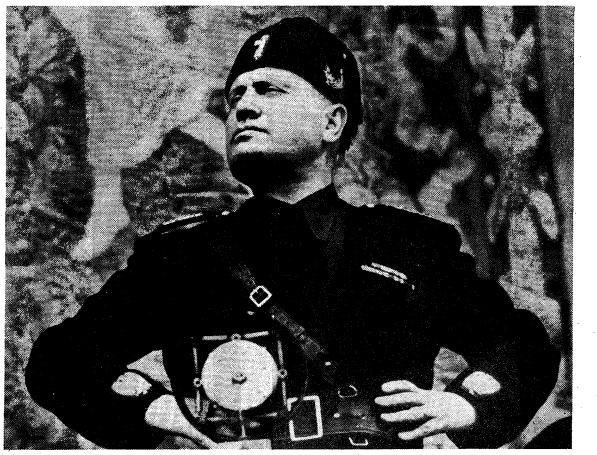
pear to be compatible activities satisfying analogous requirements; and "From Musical Ideas to Computers and Back," in HARRY B. LINCOLN (ed.), *The Computer and Music*, pp. 23–36 (1970), about the conditions under which systems of digital and analog computers would assist a composer in creating and generating music of contemporary relevance and significance; JOHN CAGE, *Silence: Lectures and Writings* (1961), *A Year From Monday: New Lectures and Writings* (1967), and *Notations* (1969), publications of one of the most influential 20th-century composers; ALAIN DANIELOU, *An Introduction to the Study of Musical Scales* (1943), deals with the problem of scales and modes on an intercultural basis, particularly in relating those of India to those of ancient Greece; HERMANN HELMHOLTZ, *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*, 3rd ed. (1870; Eng. trans., *On the Sensations of Tone As a Physiological Basis for the Theory of Music*, 1875), still an indispensable treatise for one interested in the acoustics, and to some extent, the physical aspects of music; PAUL HINDEMITH, *Unterweisung im tonsatz*, 2 vol. (1937–39; Eng. trans., *The Craft of Musical Composition*, 2 vol., 1941–42; 4th ed. reprinted 1970), presents the composer's personal approach to writing and analyzing music, and contains analyses of a number of 20th-century compositions; LEONARD B. MEYER, *Emotion and Meaning in Music* (1956), an investigation of the nature of emotional and intellectual meaning in music, the responses to musical stimuli, and the social and psychological conditions under which communication takes place in a musical experience; ABRAHAM MOLES, *Théorie de l'information et perception esthétique* (1958; Eng. trans., *Information Theory and Esthetic Perception*, 1966), the earliest and most thorough attempt to apply to sequences of musical events the mathematical theories of information developed by Shannon, Weaver, and others for linguistic codes; HARRY PARTCH, *Genesis of a Music* (1949), a personal exposition of one of the 20th-century's most individualistic aesthetics, as remarkable for its philosophy as for its detailed theories of tuning and instrumentation; LEON PLANTINGA, "Philippe de Vitry's *Ars Nova*: A Translation," *Journal of Music Theory*, 5:204–223 (1961), an epochal treatise of the early 14th century that presents a notation capable of expressing more precisely a greater variety of rhythmic relationships than theretofore possible in writing; GEORGE PERLE, *Serial Composition and Atonality: An Introduction to the Music of Schoenberg, Berg, and Webern*, 2nd ed. rev. (1968), a lucid introduction to early twelve-tone practice that emphasizes motivic manipulation and the atonal origins of serialism (many examples and analyses are included); HUGO RIEMANN, *System der musikalischen rhythmik und metrik* (1903); and *Die Natur der Harmonik* (1882; Eng. trans., *The Nature of Harmony*, 1886), views all music from the point of view of Riemann's own musical and intellectual heritage—19th-century Germany; HEINRICH SCHENKER, *Neue musikalische Theorien und Phantasien*, vol. 1 (1906; Eng. trans., *Harmony*, 1954), one of the most influential theoretical treatises of the first half of the 20th century that involves reducing harmony and melody to their skeletal and functional frameworks; ARNOLD SCHOENBERG, *Style and Idea* (1950), a collection of essays on various aspects of music that gives insights into the author's thinking about music, both technically and emotionally; KARLHEINZ STOCKHAUSEN, "How Time Passes," *Die Reihe*, no. 3, pp. 10–40 (1959), an attempt to apply serial concepts to the notation of time and to take time and its ordering to be the basis for composition (difficult and occasionally obscure, but influential and controversial); OLIVER STRUNK (ed.), *Source Readings in Music History from Classical Antiquity Through the Romantic Era* (1950), contains basic readings, in English translation, of many relevant areas of music theory, especially of Medieval and Renaissance theory.

(B.B.J.)

Mussolini, Benito

The first of Europe's Fascist dictators and the authoritarian head of the Italian state for more than 20 years, Mussolini first seized power in a coup d'etat in 1922 as the leader of a nationalist, anti-Communist political party, whose policy was based on force and supported by black-shirted military squads. Having established a totalitarian dictatorship, he imposed order upon the country and carried out extensive schemes of social reform and public works. But, impelled by territorial ambitions, he invaded Abyssinia in 1935, occupied Albania in 1939, and declared war on the Allies in 1940, leading Italy to exhaustion and defeat and his regime to extinction.

Early life. Mussolini was born near the small town of Predappio in the Romagna district of central Italy on July 29, 1883, the first child of the local blacksmith. In later years he expressed pride in his humble origins and often spoke of himself as a "man of the people." Although the Mussolini family was, in fact, less humble than he alleged—his father, a part-time Socialist journalist as well as a blacksmith, was the son of a lieutenant in the National Guard, and his mother was a schoolteacher—the Mussolinis were certainly poor. They lived in two crowded rooms on the second floor of a small, decrepit *palazzo*; and, because the blacksmith spent much of his time discussing politics in taverns and most of his money on his mistress, the meals that were eaten by his three children were often meagre.



Mussolini.
H. Roger-Viollet

Schooling and early jobs. Benito was a quarrelsome, restless child, who was disobedient, unruly, and aggressive. He was a bully at school and moody and introspective at home. "One day," he told his mother, "I shall make the whole earth tremble." Because the teachers at the village school could not control him, he was sent to board with the strict Salesian Fathers at Faenza, where he proved himself more troublesome than ever, stabbing a fellow pupil with a penknife and attacking one of the Fathers who had attempted to beat him. He was expelled and sent to the Giosuè Carducci School at Forlimpopoli, from which he was also expelled after assaulting another pupil with his penknife.

Turbulent youth

Although a violent and intractable child, he was also intelligent; and he passed his final examinations without difficulty. He obtained, also without difficulty, a teaching diploma and for a time worked as a schoolmaster. But he soon realized that he was totally unsuited for such work—too restless, too ambitious, too impatient. "I was a Bohemian in those days," he wrote, "I made my own rules and I did not keep even them." At the age of 19, a short, pale young man with a powerful jaw and enormous, dark, piercing eyes, he left Italy for Switzerland with a nickel medallion of Karl Marx in his otherwise empty pockets.

For the next few months, according to his own account, he lived from day to day, carelessly and recklessly; jumping from job to job, working as a building labourer and in a butcher's shop, as a delivery man and a factory hand; sleeping in public lavatories, under bridges, with any women who would have him and threatening to rape those who would not.

Recognition as a journalist and orator. At the same time, however, he was gaining a reputation as a young man of strange magnetism and remarkable rhetorical talents. He read widely and voraciously, if not deeply, plunging into Kant and Spinoza, Kropotkin and Nietzsche, Hegel, Kautsky, and Sorel, picking out what appealed to him and discarding the rest, forming no coherent political philosophy of his own yet impressing his companions as a potential revolutionary of uncommon personality and striking presence. While earning a reputation as a political journalist and public speaker, he undertook propaganda for a building trade union, proposing a general strike and advocating violence as a

means of enforcing the workers' demands. Repeatedly, he called for a day of vengeance; more than once he was arrested and imprisoned. When he returned to Italy in 1904, even the Roman newspapers had started to mention his name.

For some time after his return little was heard of him. He once more became a schoolmaster, this time in the Venetian Alps, north of Udine, where he lived, so he confessed, a life of "moral deterioration," neglecting his unwelcome duties, drinking too much, making flamboyant speeches to the fountain in the village square, contracting syphilis. But soon tiring of so wasteful a life, he returned to trade-union work, to journalism, and to extreme politics, which led yet again to arrest and imprisonment.

Marriage to Rachele Guidi

During a period of freedom in 1909 he fell in love with 16-year-old Rachele Guidi, the younger of the two daughters of his father's widowed mistress; she went to live with him in a damp, cramped apartment in Forlì and later married him. Soon after the marriage, Mussolini was imprisoned for the fifth time; but by then Comrade Mussolini had become recognized, both as journalist and orator, as one of the most gifted, and dangerous, of Italy's younger Socialists.

After writing in a wide variety of Socialist papers, he founded a newspaper of his own, *La Lotta di Classe* ("The Class Struggle"), all four pages of which he wrote himself. So successful was this paper that in 1912 he was appointed editor of the official Socialist newspaper, *Avanti!*, whose circulation he soon doubled; and as its anti-militarist, anti-nationalist, and anti-imperialist editor, he thunderously opposed Italy's intervention in World War I.

Soon, however, he changed his mind about intervention. Swayed by Karl Marx's aphorism that social revolution usually follows war, and persuaded that "the defeat of France would be a death-blow to liberty in Europe," he began writing articles and making speeches as violently in favour of war as previously he had condemned it. He resigned from *Avanti!* and, financed by a publisher who favoured war against Austria, he assumed the editorship of *Il Popolo d'Italia*, in which he unequivocably stated his new philosophy: "From today onwards we are all Italians and nothing but Italians. Now that steel has met steel, one single cry comes from our hearts—*Viva l'Italia!*" It was the birth cry of Fascism. To shouts of "Traitor!" "Cutthroat!" "Scoundrel!" Mussolini was expelled from the Socialist Party. He went to fight in the war.

Rise to power. Wounded while serving with the Bersaglieri (a corps of sharpshooters), he returned home a convinced anti-Socialist and a man with a sense of destiny. As early as February 1918, he was advocating the emergence of a dictator—"a man who is ruthless and energetic enough to make a clean sweep." Three months later, in a widely reported speech at Bologna, he hinted that he himself might prove to be such a man.

Advocation of a dictatorship

The birth of Fascism. The following year the nucleus of a party prepared to support his ambitious idea was formed in Milan. In an office in Piazza San Sepolcro about 200 assorted republicans and anarchists, syndicalists and discontented Socialists, restless revolutionaries and discharged soldiers met to discuss the establishment of a new force in Italian politics. Mussolini called this force a Fasci di Combattimento, a group of fighters bound together by ties as close as those that secured the *fascinae* of the lictors—the symbols of ancient Roman authority. So Fascism was created and its symbol devised; and in Mussolini Fascism had found its natural *duce* (leader).

In rally after rally, surrounded by supporters wearing black shirts, which the labourers of the Romagna had adopted as the uniform of the anarchists, Mussolini caught the imagination of the crowds. His physique was impressive; his style of oratory, staccato and tautophonic, was superb; his voice was described by an English admirer as the most beautiful she had ever heard. His attitudes were highly theatrical, his opinions were contradictory, his facts were often wrong, and his attacks were frequently malicious and misdirected; but his words were so dramatic, his metaphors so apt and striking, his vigorous, repetitive gestures so extraordinarily effective, that he rarely failed to impose his mood upon his audience.

Throughout the immediate postwar years, the Fascist *squadristi* marched through the towns and villages of Italy, singing patriotic songs, shouting nationalist slogans, waving skull-embroidered flags, attacking Socialists and Communists and all who sympathized with them, gathering more and more recruits. They provoked and intensified an anarchic situation in Italy, trying to prepare the people for the authoritarianism of the dictator that Mussolini prescribed as a cure for the nation's ills.

The march on Rome. In the summer of 1922 Mussolini's opportunity presented itself. To the dismay of the Italian people as a whole, a general strike was called. Mussolini declared that unless the government prevented the strike, the Fascists would. "Either the government will be given to us," he threatened at a gathering of 40,000 Fascists in Naples, "or we will seize it by marching on Rome. It is a matter of days, of hours. . . . I guarantee, I swear to you that the orders . . . will come." Responding to his oratory the assembled Fascists excitedly took up the cry, shouting in unison "Roma! Roma! Roma!" All appeared eager to march.

Later that day, Mussolini and other leading Fascists decided that in four days' time—following Fascist-provoked riots and the occupation of police and radio stations, prefectures, and post offices—the Fascist militia would advance upon Rome in converging columns led by four leading party members later to be known as the Quadrumviri. Mussolini himself was not one of the four. Indeed, in the opinion of one of them—Italo Balbo, a young hero of the *squadristi*—Mussolini's character was not suited to the determined leadership of a coup d'état. So hesitant, capricious, and vacillating did Mussolini prove to be at this critical period, Balbo said afterward, that he had to be firmly told that the Fascists were going to Rome, either with or without him.

In fact, even at this late stage Mussolini was still hoping for a political compromise, and he was still in touch with other parties outside the Fascist camp. He returned to Milan, determined to remain there until he could be sure that the crisis was about to be resolved, prevaricating when more zealous Fascists urged him to move to Rome and to seize the power so nearly within his grasp. He refused to move before King Victor Emmanuel III summoned him in writing. Meanwhile, all over Italy the Fascists prepared for action, and the march on Rome began. Although the march was a far less orderly affair than Fascist propaganda later suggested, it was sufficiently threatening to bring down the government. And the King, prepared to accept the Fascist alternative, dispatched the telegram for which Mussolini had been waiting.

Appointment as prime minister

Dictatorship. Mussolini's obvious pride in his achievement at becoming (October 31, 1922) the youngest prime minister in Italian history was not misplaced. He had certainly been aided by a favourable combination of circumstances, both political and economic; but his remarkable and sudden success owed quite as much to his own personality, to native instinct and shrewd calculation, to astute opportunism, and to his unique gifts as an agitator.

In the first speech he made in the Chamber after the march on Rome, Mussolini declared, "I could have transformed this grey hall into an armed camp of Blackshirts, a bivouac for corpses. I could have nailed up the doors of Parliament."

Yet, although he had achieved power by force, initially he exercised it with restraint. Anxious to demonstrate that he was not merely the leader of Fascism but also the head of the government of a united Italy, he presented to the King a list of ministers, a majority of whom were not members of his party. He made it clear, however, that he intended to govern authoritatively and personally. He demanded and obtained full dictatorial powers for a year; and in that year he pushed through a law that enabled the Fascists to secure a majority in the Chamber. The elections in 1924, though undoubtedly fraudulent, firmly secured his personal power; it was a power that he was to keep for more than 20 years.

The popular leader. Indeed, in the early days of his rule, the Italian people welcomed his authority. They were tired of strikes and riots, responsive to the flamboyant techniques and medieval trappings of Fascism, and ready to submit to dictatorship, provided the national economy was stabilized and their country restored to its dignity. Mussolini seemed to them the one man capable of bringing order out of chaos. Soon order had been restored, the conditions of the workers had begun to improve, and ambitious programs of public works had been inaugurated; and it was he who was given the credit for Fascism's apparent success. After his regime had survived the shock of the murder by Fascists of the distinguished Socialist leader, Giacomo Matteotti, Mussolini was accepted throughout Italy at his own evaluation as Italy's man of destiny.

World-wide acclaim

It was not only Italians who fell under the sway of his massive, proudly jutting jaw and black, wide-open eyes and who thrilled to the sight of him as he stood gazing down impassively upon his admirers from the balcony outside his office in the Palazzo Venezia in Rome, his hands upon his hips, his legs splayed apart, awaiting the moment when he would enthrall them with expressive gestures and deep, emotive voice. He was hailed as a genius and a superman by public figures all over Europe and in the United States, compared to Caesar, to Napoleon, and to Cromwell. The American Cardinal O'Connell of Boston said that Mussolini was a genius given to Italy by God; Winston Churchill declared that he himself would have donned the Fascist black shirt, had he been an Italian.

Mussolini's achievements were considered scarcely less than miraculous. He had transformed and reinvigorated his divided and demoralized country; he had carried out his social reforms and public works without losing the support of the industrialists and landowners; he had even succeeded in coming to terms with the papacy.

He was undoubtedly more popular in Italy than anyone had ever been before, despite his sure destruction of civil liberties. After Mussolini had survived an attempt upon his life by an unbalanced Irishwoman, the secretary of the Fascist Party announced to wildly cheering crowds, "God has put his finger on the Duce! He is Italy's greatest son, the rightful heir of Caesar!" And the crowds roared their approval and devotion.

The private man. Yet beneath the public figure, there was a different Mussolini—far less decisive, less magnanimous, less omniscient. For all his undoubted gifts, he remained essentially a journalist rather than a statesman; and his true genius lay in his ability to manipulate the masses by the written and spoken word. As the years progressed his intimates recognized that behind the bluster, the cunning, and the charm, the Duce was coarse and childish, unstable and irresolute, suspicious, deceitful, and pathologically egotistic. From Turin to Naples, walls were covered with the words that set out the fundamental tenet of Fascist doctrine: "*Il Duce ha sempre ragione*" ("The Duce is always right"). But those who knew him best knew that in truth the Duce was often wrong. He knew little of how to run a government and had no patience with difficult work; he was even known to have carelessly written "Approved" on two conflicting memorandums emanating from two different ministries before going upstairs to one or another of his mistresses. The light was left burning in his office far into the night; but there was no one there.

Foreign policies. Nevertheless, throughout the 1920s and the early 1930s Mussolini remained a hero to his own people and profoundly respected in the outside world. He might, indeed, have remained so until his death had not his callous xenophobia and arrogance, his misapprehension of Italy's fundamental necessities, and his dreams of empire led him to look about for foreign conquest.

His eye rested first upon Abyssinia, which, after ten months of preparations, rumours, threats, and hesitations, Italy invaded in October 1935. Europe expressed its horror; but, having done so, did no more. Sanctions were imposed by the League of Nations; yet care was taken to ensure that the list of prohibited exports should not include any, such as oil, that might provoke a European war. Had he been faced with oil sanctions, Mussolini said, he would have had to withdraw from Abyssinia within a week. But he was faced with no such problem; and on the night of May 9, 1936, he was able to announce to an enormous, expectant crowd of about 400,000 people standing shoulder to shoulder around Piazza Venezia, that "in the 14th year of the Fascist era" a great event had been accomplished: Italy had its empire.

Alliance with Hitler

Italy had also found a new ally. Intent upon his own ambitions in Austria, Hitler had actively encouraged Mussolini's African adventure, and, under Hitler's guidance, Germany had been the one powerful country in western Europe that had not turned against Mussolini. The way was now open for the Pact of Steel—a Rome–Berlin Axis and a brutal friendship between Hitler and Mussolini that was to ruin them both.

Hitler had long admired Mussolini, but the admiration had not been returned. The Duce had viewed Hitler with a kind of amused contempt. "He's quite mad," he had said after meeting him for the first time, "a silly little clown." As for Hitler's National Socialism, Mussolini saw it as merely the "revolt of the German tribes of the old primeval forest." But as Hitler piled triumph upon triumph, and succeeded in one aggressive gamble after another, Mussolini became increasingly respectful of German might and jealous of his rival.

Role in World War II. The outbreak of World War II in 1939 faced Mussolini, therefore, with a cruel dilemma. Although he had for years been the victim of his own propaganda, although he had grown accustomed to subordinates telling him what he wanted to be told, he recognized in his more thoughtful moments that the country was ill-prepared for war. The Abyssinian campaign, followed by Italian intervention on Francisco Franco's side in the Spanish Civil War (1936–39), and the Italian occupation of Albania (April 1939) had placed a great strain on the country's already inadequate military resources. Mussolini boasted proudly of the "8,000,000 bayonets" that could be mobilized to fight the enemy at any moment; but because he had been in charge of all three service ministries since 1925—apart from a short break between 1929 and 1933—he was well aware of the falsity of Italian military pretensions, the lack of modern tanks, artillery, and aircraft, and the weakness of the navy. Moreover, Abyssinia and the other recently acquired Italian possessions in North Africa needed a long period of peace for their stability; so did Albania. In Italy, peace was equally needed to move factories, plant, and machinery to the impoverished south from the congested industrial centres of the north and to improve Italian holdings of foreign currency.

Yet, while Mussolini understood that peace was essential to Italy's well-being, that a long war might prove disastrous, and that he must not "march blindly with the Germans," he was beset by fears that the Germans "might do good business cheaply" and that by not intervening he would lose his "part of the booty." His foreign secretary and son-in-law, Count Galeazzo Ciano, recorded that during a long, inconclusive discussion at the Palazzo Venezia, Mussolini at first agreed that Italy must not go to war, "then he said that honour compelled him to march with Germany."

Italian intervention. Mussolini watched the progress of Hitler's war with bitterness and alarm, becoming more and more bellicose with each fresh German victory, while frequently expressing hope that the Germans would be slowed down, would meet with some reverse that would satisfy his personal envy and give Italy a breathing space. When Germany marched westward, however, and France seemed on the verge of collapse, Mussolini felt he could delay no longer. To Marshal Pietro Badoglio, the army chief of staff, who urged caution and delay, he declared,

> You, Marshal, are not calm enough to judge the situation. I can tell you that everything will be over by September and that I only need a few thousand dead so that I can sit at the peace conference as a man who has fought.

So, on June 10, 1940, the fateful declaration of war was made.

From the beginning, the war went badly for Italy.

France surrendered before there had been the opportunity for even a token victory; and Mussolini left for a meeting with Hitler, sadly aware, as Ciano put it, that his opinion had "only a consultative value." Indeed, from then on Mussolini was obliged to face the fact that he was very much the junior partner in the Axis alliance.

Mussolini's subordination to Hitler

Thenceforth, every time he met Hitler throughout the war, the failure of Italian arms was uppermost in his mind; and he was forced to listen, while Hitler talked at interminable length about past German victories and future German plans. To make his meetings with Hitler more intolerable, the Germans kept the details of their plans concealed, presenting their allies with a *fait accompli* for fear lest prior discussion would destroy surprise. "All Italians chatter like gypsies," the German propaganda minister Josef Goebbels had warned; and Hitler had heeded the warning. "Our aims," Hitler had decided before the war began, "must remain hidden from Italy." And thus such moves as the occupation of Romania and the later invasion of the Soviet Union were undertaken by Germany without any prior warning being given to Mussolini.

It was to "pay back Hitler in his own coin," as Mussolini openly admitted, that he decided to attack Greece through Albania without advising the Germans of his plans. Hitler had always intended that Mussolini should concentrate his efforts upon ejecting the British from Egypt and the Middle East; exasperated by the slowness of his commanders in North Africa, however, the Duce turned upon Greece instead. The result was an extensive and ignominious defeat, and the Germans were forced unwillingly to extricate him from its consequences. Hitler also had to come to his ally's help in North Africa; and the Germans began to take precautions against the likelihood of an Italian collapse. When the Allies successfully invaded Sicily in July 1943, it was obvious that collapse was imminent.

Fall from power. For some time Italian Fascists and non-Fascists alike had been preparing Mussolini's downfall. On July 24, at a meeting of the Fascist Grand Council—the supreme constitutional authority of the state, which had not met once since the war began—a resolution was passed by an overwhelming majority in effect dismissing Mussolini from office. Disregarding the vote as a matter of little concern and refusing to admit that his minions could harm him, Mussolini appeared at his office the next morning as though nothing had happened. That afternoon, however, he was arrested by royal command on the steps of the Villa Savoia after an audience with the King.

Last years. Imprisoned first on the Island of Ponza, then on a remoter island off the coast of Sardinia, he was eventually transported to a hotel high on the Gran Sasso d'Italia in the Abruzzi Mountains, from which his rescue by the Germans was deemed impossible. Nevertheless, by crash-landing gliders on the slopes behind the hotel, German commandos effected his rescue and flew the Duce to Munich.

Rather than allow the Germans to occupy and govern Italy entirely in their own interests, Mussolini agreed to Hitler's suggestion that he establish a new republican Fascist government in the north and execute those members of the Grand Council, including his son-in-law, Ciano, who had dared to vote against him. But the Repubblica Sociale Italiana thus established at Salò was, as Mussolini himself grimly admitted to visitors, no more than a puppet government at the mercy of the German command. And there, living in dreams and "thinking only of history and how he would appear in it," as one of his ministers said, Mussolini awaited the inevitable end.

As German defenses in Italy collapsed and the Allies advanced rapidly northward, the Italian Communists of the partisan leadership determined upon Mussolini's execution. Rejecting the advice of various advisers, including the elder of his two surviving sons—his second son had been killed in the war—Mussolini refused to consider flying out of the country, and he made for the Valtellina, intending to make a final stand in the mountains; but only a handful of men could be found to follow him. He tried to cross the frontier disguised as a German soldier in a convoy of trucks retreating toward Innsbruck. But he was recognized and, together with his mistress Claretta Petacci, who had insisted on remaining with him to the end, he was shot and killed near Azzano on April 28, 1945.

Capture and execution

Mussolini's death was greeted by the great mass of the Italian people without regret. He had lived beyond his time and had dragged his country into a disastrous war, which it was unwilling and unready to fight. His achievements were forgotten; and it was not until many years later, when Fascism could be viewed objectively as a short episode in Italy's long history, that they could be truly assessed.

BIBLIOGRAPHY. All biographies of Mussolini will eventually be superseded by that of RENZO DE FELICE, whose first three volumes deal exhaustively with his career up till 1929. These are *Mussolini il rivoluzionario, 1883–1920* (1965); *Mussolini il fascista, 1921–1925* (1966); and *Mussolini il fascista, 1925–1929* (1968). At present the most comprehensive biography is GIORGIO PINI and DUILIO SUSMEL, *Mussolini, l'uomo e l'opera*, 4 vol. (1953–55). The best biography in English is IVONE KIRKPATRICK, *Mussolini* (1964). LAURA FERMI, *Mussolini* (1961), is particularly valuable for the years of power; and CHRISTOPHER HIBBERT, *Il Duce: The Life of Benito Mussolini* (1961), for the final phase. ROY MACGREGOR-HASTIE, *The Day of the Lion* (1964), is an original and controversial study of Mussolini's hold upon the Italian people. GAUDENS MEGARO, *Mussolini in the Making* (1938), provides a detailed account of his origins and early years; and PAOLO MONELLI, *Mussolini, piccolo borghese* (1950; Eng. trans., *Mussolini: An Intimate Life*, 1953), gives the best portrait of the man. There are some details not elsewhere available in RICHARD COLLIER, *Duce! The Rise and Fall of Benito Mussolini* (1971).

(C.Hi.)

Mussorgsky, Modest

A leading composer of the 19th century, Mussorgsky was one of the first to use Russian folk music as a basis for his compositions. He is chiefly noted for the realism of his music as expressed in his songs and operas, the best known of which is *Boris Godunov*. His importance and influence on later composers is out of proportion to his relatively small total output. Few composers, however, were less derivative, more bold and original in their musical language.

The Bettmann Archive

Mussorgsky, portrait by Ilya Repin, 1881. In the Gosudarstvennaya Tretyakovskaya Galeriya, Moscow.

Modest Petrovich Mussorgsky was born on March 21, (March 9, old style), 1839, at Karevo in the province of Pskov. He was the son of a landowner but had peasant blood, his father's grandmother having been a serf. According to his autobiographical sketch, written in 1881, Mussorgsky claimed that he learned about Russian fairy tales from his nurse. "This early familiarity with the spirit

of the people, with the way they lived, lent the first and greatest impetus to my musical improvisations. . . ." His mother, herself an excellent pianist, gave Modest his first piano lessons, and at seven he could play some of Franz Liszt's simpler pieces.

In August 1849 his father took Modest and also his other musical son, Filaret, to St. Petersburg (now Leningrad), where Modest attended the Peter-Paul School in preparation for a military career. At the same time, mindful of Modest's musical bent, their father entrusted the two boys to Anton Gerke, future professor of music at the St. Petersburg Conservatory.

In 1852, after a short period at another preparatory school, Mussorgsky entered the School for Cadets of the Guard. There, in his first year he composed his *Podpraporshchik* (*Porte-Enseigne Polka*), published at his father's expense. Although not the most industrious of students, he gave proof of tremendous curiosity and wide-ranging interests. His lifelong enthusiasm for literature, history, painting, philosophy, science, astronomy, and theology was already evident, to say nothing of music, as attested by his frequent and lengthy visits with his religion instructor, Father Krupsky, who taught him "the very essence of church music—the music of the Greek, Catholic, Protestant and Lutheran churches."

First published work

In 1856, by now a lieutenant, Mussorgsky joined the Preobrazhensky Guards, one of Russia's most aristocratic regiments, where he made the acquaintance of several music-loving officers who were habitués of the Italian theatre. During this same period he came to know Aleksandr Borodin, a fellow officer who was to become another important Russian composer. Borodin has provided a very vivid picture of the musician:

> There was something absolutely boyish about Mussorgsky; he looked like a real second-lieutenant of the picture books . . . a touch of foppery, unmistakable but kept well within bounds. His courtesy and good breeding were exemplary. All the women fell in love with him. . . . That same evening we were invited to dine with the head surgeon of the hospital. . . . Mussorgsky sat down to the piano and played . . . very gently and graciously, with occasional affected movements of the hands, while his listeners murmured, "charming! delicious!"

During the winter of 1856 a regimental comrade introduced Mussorgsky into the home of the Russian composer Aleksandr Dargomyzhsky. At one of the musicales there, Mussorgsky discovered the music of the seminal Russian composer Mikhail Glinka, and this quickened his own Russophile inclinations. Three years later, in June 1859, he saw the Moscow Kremlin for the first time, an important experience that represented his first "physical" communion with Russian history. Later he was to say: "To bring the past into the present—that is my task!" Through Dargomyzhsky, Mussorgsky met another composer, Mily Balakirev, who became his teacher. Since the death of their father (in 1853), the Mussorgsky brothers had seen their poorly administered patrimony decrease substantially. Now, with the freeing of the serfs, it vanished. Mussorgsky, having decided to devote himself to music, had quit the army three years earlier and, since 1863, had been working as a civil servant in the Ministry of Communications. His distressing financial troubles date from that time, and he had to seek the help of moneylenders. Mussorgsky achieved artistic maturity in 1866 with a series of remarkable songs about ordinary people; an even larger series appeared the following year. In 1868 he reached the height of his conceptual powers in composition with the first song of his incomparable cycle *Detskaya* (*The Nursery*) and a setting of the first act of Nikolay Gogol's *Zhenitba* (*The Marriage*).

Composition of *Boris Godunov*

In 1869 he settled on his great work *Boris Godunov*, to his own libretto based on a book by Aleksandr Pushkin. The first version, completed on December 27, 1869, was rejected by the advisory committee of the imperial theatres. In no way discouraged, the composer returned to the task and on August 4, 1872, put the finishing touches to the second version, adding the roles of Marina and Rangoni as well as several new episodes. The first production of *Boris* took place on February 8, 1874, at St. Petersburg.

In spite of the work's success, Mussorgsky was not happy. His solitude and bitter ordeal now commenced. Having once loved a cousin who died prematurely and having remained faithful to this memory, the composer never married. In 1865, after the death of his mother, he lived with his brother, then shared a small flat with the Russian composer Nikolay Rimsky-Korsakov until 1872, when his colleague married. Left very much alone, Mussorgsky began to drink to excess, although the composition of *Khovanshchina* perhaps offered some distraction (left unfinished at his death, this opera was completed by Rimsky-Korsakov). After long months of solitude, he found a companion in the person of a distant relative, Arseny Golenishchev Kutuzov. This impoverished 25-year-old poet inspired his two cycles of melancholy melodies, *Bez solntsa* (*Sunless*) and *Pesni i plyaski smerti* (*Songs and Dances of Death*). At that time Mussorgsky was haunted by the spectre of death—he himself had only seven more years to live. One after another his friends died, the first being the painter Viktor Gartman, who inspired the piano suite *Kartinki s vystavki* (*Pictures from an Exhibition;* orchestrated in 1922 by the French composer Maurice Ravel). In August 1875 Mussorgsky wrote to his great friend the critic Vladimir Stasov: "Arseni took our latch key with him to the country and I was unable to get into our flat, so I went to live with my venerable friend Naumov in whose home I have just finished the first act of *Khovanshchina*." Actually, Arseny had not taken the key; the landlady had put the penniless Mussorgsky's suitcase on the doorstep and locked him out.

Decline and final years

This marked the beginning, in the company of Naumov, of Mussorgsky's downfall and of a solitude made all the more painful by the total neglect of his friends, who, Balakirev included, treated him like an outcast. Nonetheless, the composer, clinging to life, began his opera *Sorochinskaya yarmarka* (unfinished; *Sorochintsy Fair*), inspired by Gogol's tale. As the accompanist of an aging singer, Darya Leonova, Mussorgsky departed on a lengthy concert tour of southern Russia and the Crimea. On his return he tried teaching at a small school of music in St. Petersburg.

On February 24, 1881, three successive attacks laid him low. His friends had him transported to a hospital where for a time his health improved sufficiently for one of the leading Russian artists of the day, Ilya Repin, to paint his portrait. But on March 28 (March 16, O.S.), the nurse attending him heard him cry out: "All is over! . . . How unfortunate I am!" His final agony lasted less than 15 minutes.

Boris Godunov and to a lesser degree *Khovanschina* display Mussorgsky's dramatic technique of setting sharply characterized individuals against the background of country and people. His original orchestration for *Boris*, as opposed to Rimsky-Korsakov's more colourful gloss on it usually heard in the opera house, is as stark and economical as his unorthodox harmony. The songs he composed, many to his own texts, describe scenes of Russian life with great vividness and insight; they also realistically reproduce the inflections of the spoken Russian language.

MAJOR WORKS

OPERAS: *Zhenitba* (1st act only, composed 1868; *The Marriage*); *Boris Godunov* (St. Petersburg, 1874); *Khovanshchina* (completed by Rimsky-Korsakov, St. Petersburg, 1886); *Sorochinskaya yarmarka* (unfinished, composed 1874–80, produced at Moscow, 1913, *Sorchintsy Fair*).

ORCHESTRAL: *Ivanova noch na Lysoy gore* (1867; *St. John's Night on the Bare Mountain*); triumphal march, *Vzyatiye Karsa* (1880; *The Capture of Kars*).

PIANO: *Intermezzo in modo classico* (1861); scherzino, Shveja (1871; *The Seamstress*); suite, *Kartinki s vystavki* (1874; *Pictures from an Exhibition*), later orchestrated by Maurice Ravel.

SONGS: More than 60, including "Noch" (1864; "Night"); "The Seminarist" (1866); "Yevreyskaya pesnya" (1867; "Hebrew Song"); "Po griby" (1867; "Gathering Mushrooms"); "Koziol" (1867; "The He-Goat: A Worldly Story");

"Klassik" (1867; "The Classicist"); "Ozoznik" (1867; "The Ragamuffin"); *Detskaya* (song cycle, 1868–1872; *The Nursery*); *Bez Solntsa* (song cycle, 1874; *Sunless*); *Pesni i plyaski smerti* (song cycle, 1875–77; *Songs and Dances of Death*); "Pesn' Mefistofelja v pogzebke Averbacha" (1879; "Mephistopheles' Song of the Flea").

BIBLIOGRAPHY. The most exhaustive works on Mussorgsky have been published in Russian. An excellent recent study is Георгии Никитич Хубов, *Мусоргский* (1969). Although it has some political overtones, it is very complete, reproducing both an extensive Russian bibliography and a complete list of the major works. Also recommended are ROBERT GODET, *En Marge de "Boris Godounov"* (1926); VLADIMIR FEDOROV, *Moussorgski* (1935); OSKAR VON RIESEMANN, *Modest Petròwitsch Mussorgski* (1923; Eng. trans., 1929); MICHEL HOFMANN, *Le Vrai Visage de Moussorgski* (1952) and *La Vie de Moussorgski* (1964); JAY LEYDA and SERGEI BERTENSSON (eds.), *The Moussorgski Reader* (1947); and MICHEL D. CALVOCORESSI, *Mussorgsky* (1962). A complete edition of Mussorgsky's works was begun in Moscow in 1928; interrupted during the war, publication is once again in progress. Certain volumes have been done in collaboration with Oxford University Press and University Editions.

(M.-R.Ho.)

Mutation

Mutation is a process by which the hereditary constitution of a cell is altered; ultimately, it results in a genetically changed organism or population of cells. Thus, a heritable change of albinism (lack of pigmentation) may appear in an individual of an inbred colony of black rats, or a heritable change to a new seed colour may appear in a true-breeding variety of bean. In bacteria, mutation may result in the emergence of strains that are resistant to the antibiotic drugs by which the original strain was killed. The term mutation was coined in this sense early in the 20th century by the Dutch botanist Hugo de Vries, one of the rediscoverers of the work of the 19th-century Austrian biologist Gregor Mendel. The term has been retained, although it is now known that the changes observed by de Vries were not mutations in the present sense. Mutations in the modern conception, far from always having striking effects, are responsible for the slight changes that play a major role in evolution. See EVOLUTION; HEREDITY; POLYMORPHISM, BIOLOGICAL.

The nature of mutation. Heritable differences in cell constituents other than the nucleus, such as plastids or mitochondria, may have arisen through mutational processes; but the present article deals only with mutations that occur within the cell nucleus. These take place in the genes, which consist of deoxyribonucleic acid (DNA) or, rarely, of ribonucleic acid (RNA). In the chromosomes of all but the most primitive organisms, DNA is combined with protein. DNA is a long molecule with a backbone of two sugar–phosphate chains to which are attached sequences formed of two purine bases, adenine (A) and guanine (G), and two pyrimidine bases, thymine (T) and cytosine (C). The base sequences in the two chains of DNA are complementary to each other; that is, adenine is always paired with thymine and guanine with cytosine. These base pairs are the smallest elements of the genetic code, and their sequence spells the information for the inherited biochemical functions of a cell as well as for the formation of identical genes at each chromosome replication. Mutations arise from substitutions of one base for another or from changes in the sequence of bases. Base substitutions are of two types: transitions, in which a purine is replaced by a purine (*e.g.*, A by G) or a pyrimidine by a pyrimidine (*e.g.*, C by T); and transversions, in which a purine is replaced by a pyrimidine (*e.g.*, A by C) or vice versa (C by A). Changes in the sequence of bases result mainly from insertion or deletion of one or a few bases. In genes specifying a protein such as an enzyme (biological catalyst) or hemoglobin (an oxygen-carrying protein in blood), the code is read in triplets of bases from one end of the gene to the other. A base substitution results in the presence of one wrong protein component (amino acid) in the protein structure. What effect this has on the functioning of the protein depends on the type and position of the wrong amino acid. Certain base substitutions result in "nonsense" triplets that code, not for an amino acid, but for premature termination of the protein structure. Such mutations result in complete loss of function, as do most insertion and deletion mutations, which change the "reading frame," or order of bases, of the gene; as a result, every amino acid coded from the mutated site to the end of the gene is incorrect. Deletions, duplications, inversions, or transpositions of whole genes or sequences of genes may result in hereditary changes that are often difficult or impossible to distinguish from gene mutations.

Origin of mutation

That mutated genes or chromosome regions form replicas of themselves just as do normal genes and chromosomes once was an intriguing aspect of mutation, but is now realized to be a consequence of the complementarity of DNA. After a mutation has occurred, therefore, it can be reversed only by a second mutation, which either restores the function of the original gene (true reverse mutation) or, by a compensatory change in the biochemical pattern of the cell, restores some degree of function even in the presence of the original (forward) mutation (suppressor mutation). Because reverse mutations must fulfill more specific requirements than forward mutations, the former are usually much rarer than the latter (see GENE; NUCLEIC ACID; and PROTEIN).

In haploid cells (which contain only one set of chromosomes) mutation immediately yields a mutant type. Thus, one mutated yeast cell can grow into many mutant cells that, for example, can ferment a sugar that the original strain could not ferment. In cells with several haploid nuclei (*e.g.*, those of many micro-organisms) or in diploid cells (which contain two chromosome sets; *e.g.*, those of most higher organisms), recessive mutations—*i.e.*, mutations to genes the effects of which are suppressed by nonmutated partner genes—remain hidden until one of the various mechanisms of hereditary transmission has produced a cell with only mutated partner genes.

In higher organisms germinal mutation in reproductive cells is distinguished from somatic mutation in other cells. Somatic mutations form the basis for such changes as a branch with white currants on a red currant bush. Germinal mutations have no effect on the individual in which they occur but may become manifest in subsequent generations. The gametes (mature reproductive cells) are haploid, but diploidy is restored at fertilization by fusion of two gametes. A dominant mutation in either gamete gives rise to mutant offspring; a recessive mutation does so only when the other gamete happens to contain the same mutated gene. This is unlikely for rare mutations, and a recessive mutation may remain hidden for many generations. A special case is sex-linked mutations,

From *Journal of Heredity* (1929)

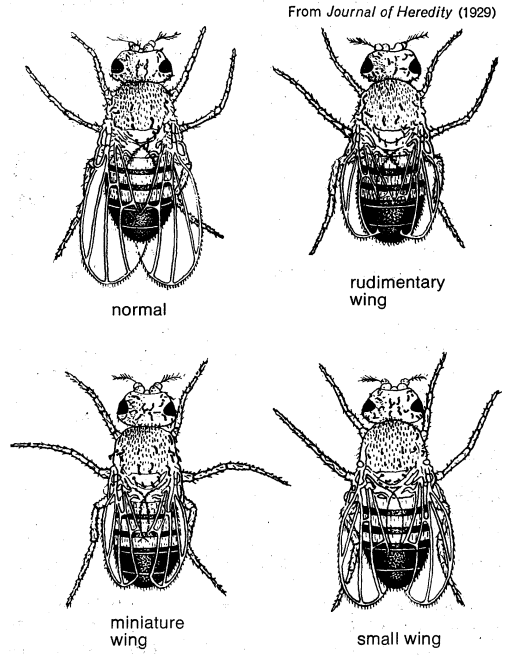


Figure 1: Mutations in the fruit fly *Drosophila*.

which, even if recessive, are always capable of expression in the sex with only one X-chromosome—the male in mammals including man. Thus, a sex-linked mutation to hemophilia expresses itself in the first male who receives the mutated gene through a gamete.

Origin of mutations. *Spontaneous mutation.* In nature, mutations occur spontaneously and unpredictably in single cells out of large populations (one in 100,000 to one in 10,000,000) of, for example, bacteria or spermatozoa.

Mutation frequency

Such a frequency seems to indicate that spontaneous mutations are exceedingly rare events in a slow-breeding species such as man. This rarity is true only when individual genes are considered. A set of human chromosomes contains at least 10,000 genes, and, if each of these mutates at a rate of one in 100,000, one out of every ten spermatozoa or ova (eggs) will carry a newly mutated gene. Because a child may receive a mutated gene from either parent, approximately every fifth child will be born with a newly mutated gene. (This is only a rough estimate, but it is not likely to be too high.) This situation does not mean that every fifth child will show some visible abnormality; the majority of mutations either are recessive or have very slight effects that are likely to be taken for normal variations in development.

The origin of spontaneous mutations is imperfectly understood. Only a fraction of them are caused by cosmic or natural mineral radiation; the remainder probably result from chemical processes in the cell nucleus or from temperature oscillations very close to a gene.

Radiation-induced mutation. The first worker to produce mutations by artificial means was the U.S. scientist H.J. Muller (1927), who exposed fruit flies (*Drosophila*) to X-rays and analyzed their progeny by ingenious methods that allowed quantitative assessment of mutation frequencies. Since then many physical and chemical agents have been shown to be mutagenic (able to cause mutation). The physical ones include all ionizing radiations (*e.g.*, X-rays), ultraviolet light, and, in cells sensitized by certain dyes, visible light. Radioactive phosphorus produces mutations not only through emitted radiation but through incorporation into the chromosomes.

The frequency of X-ray-induced mutations increases linearly with dose; *i.e.*, the output is directly proportional to the input. This finding has been interpreted to mean that every mutation is the result of one ionization event inside a gene or very close to it (target theory). The direct proportionality between X-ray dose and mutation frequency, which has been confirmed for various organisms and for doses as low as eight roentgens (units of radiation), is important in estimating genetical radiation damage. The target theory has been modified to include indirect radiation effects mediated by ionized (charged) components of water, organic substances, and peroxides. The theory is essentially unaltered by these modifications, which help to explain the influences of the physiological state of the cell and environmental conditions such as humidity and, most importantly, oxygen. Oxygen and various chemicals make cells more sensitive to the genetical action of radiation; other chemicals protect against radiation. Some chemicals act by affecting the utilization of oxygen; others, by scavenging mutagenic substances; and still others, by influencing short-lived repair processes that intervene between the original lesion in a chromosome or gene and the actual chromosome break or mutation. In actively metabolizing cells—*e.g.*, oocytes (female reproductive cells) and spermatogonia (immature male reproductive cells), but apparently not in spermatozoa (mature male reproductive cells)—the efficiency of repair seems to depend on radiation intensity; in such cells chronic irradiation produces fewer mutations than does the same dose given as acute radiation. It seems that in these cells very low radiation doses may be less effective than expected from the linear dose-effect relation.

Production and repair of dimers

While the nature of the X-ray-induced lesions in DNA has not yet been established with certainty, ultraviolet light is known to produce dimers—*i.e.*, chemically bonded pairs—between neighbouring pyrimidines, particularly thymines, on the same DNA strand. The dimers can be repaired in various ways before giving rise to mutational changes in DNA; shortwave visible light, for example, opens the chemical bonds between dimers and restores the original state (photorepair). In the dark, and in cells containing certain repair enzymes, the dimers are removed and replaced by new bases (excision repair). DNA containing dimers also may be repaired at the time of the next DNA replication by a process of reorganization involving recombination between two strands (recombination repair); this process probably results in mutant DNA.

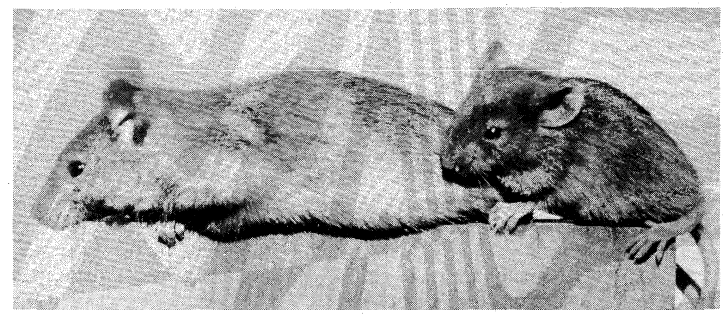


Figure 2: Two-and-a-half-month-old mouse on right, dwarfed as a result of mutation-induced pituitary deficiency, contrasted with normal littermate.
Clyde E. Keeler

In micro-organisms, mutations have been produced by exposing cells to ultraviolet-irradiated growth medium containing mutagenic peroxides. Very high doses of X-rays produce mutagenic chemicals in media used for growing bacteria and plant roots. The mutagenic effect of irradiated medium on *Drosophila*, however, is marginal. Whether radiation-sterilized food can be mutagenic for man is difficult to decide from this evidence.

Mutagenic chemicals. Chemical substances that induce mutation occur in widely divergent chemical groups and range from simple compounds such as formaldehyde to complex ones such as alkaloids. Hypotheses about the way in which chemicals produce mutations have been based mainly on studies of their reactions with DNA isolated from the cell; in some cases these hypotheses have been confirmed or supplemented by experiments with living cells containing DNA. Peroxides form a link between mutagenesis by radiation and by chemicals. Nitrous acid deaminates; *i.e.*, it removes the amino group (NH_2) from purine and pyrimidine bases. Purines such as caffeine, which has produced mutations in bacteria and fungi but not in mice, may act through their influence on nucleic-acid metabolism. Structurally similar compounds (analogues) of the naturally occurring bases in DNA—*e.g.*, 5-bromouracil or 2-aminopurine—produce mutations by first being incorporated in place of a normal base; at one of the following replications of DNA, this reaction may lead to replacement of one of the two normal purines or pyrimidines by the other—*e.g.*, of adenine by guanine. Alkylating agents such as mustard gas, epoxides, and ethylene imines remove guanine from nucleic acid in chemical tests and may do so when producing mutations. Alternatively, they may lead to replication errors via ionization of guanine. Their mutagenic ability is strongly correlated with carcinogenesis, which supports the mutation theory of cancer (see CANCER). Many chemical mutagens tend to produce delayed mutations in the descendants of treated chromosomes or instabilities that continue to give mutations over several or many generations. Some chemicals select specific chromosome regions for their action; others produce "hot spots" by attacking preferentially a few bases in a given gene. The compound hydroxylamine, for example, acts almost exclusively on cytosine; nitrosoguanidine (N-methyl-N-nitro-N-nitrosoguanidine) acts preferentially at the point of DNA replication.

Chemical mutagens have been a tool for "decoding" the information in DNA. In bacteriophage (viruses that attack bacteria), analysis of the frequencies with which certain chemicals produce mutations in opposite directions—from normal to mutant gene and back to normal—has been used to allot specific base pairs to certain

mutational sites, and results obtained with acridine dyes have helped to elucidate the triplet nature of the genetic code and the process of its translation.

Significance of mutation. Natural selection has used mutations to build well-integrated organisms. New mutations are likely to upset this balance and are therefore mostly harmful or lethal. Harmful genes are eliminated from a population by the failure of individuals carrying the mutation to reproduce. Over a long period harmful genes may do more damage than fully lethal ones, for, although a lethal gene kills the first individual in which it becomes active, a harmful gene may be carried through a series of more or less severely affected individuals before it finally kills one of them. Beneficial mutations are so rare that they can be utilized economically only in species with very high reproductive rates. Thus, better strains of *Penicillium* and other industrial fungi and of agricultural plants have been raised from artificially produced mutants. Because this work has been done at the expense of discarding thousands of inferior mutants, the method is obviously not suitable for farm animals. Nor can it be expected that an increase in mutation frequency, as a result of exposure to ionizing radiation from X-ray machines, artificial atomic disintegration, and radioactive isotopes, can have any but harmful effects on the hereditary constitution of the human race. Among chemical mutagens, only the most drastic and penetrating ones—*e.g.*, mustard gas—can be assumed to produce mutations in all cells. Most chemicals are likely to affect different cell types and organisms to different degrees; several have, in fact, been shown to do so. It is exceedingly difficult, therefore, to estimate the genetical hazards to man from chemicals in his environment on the basis of experiments with lower forms or even with other mammals. Organizations have been founded to coordinate this research and to evaluate its results.

The selective production of desirable mutations would be of obvious practical importance, but the repetitive nature of the genetic material makes it most unlikely that specific mutagens for individual genes—as distinct from individual bases—will be found. Before a chemical change in DNA becomes a mutated cell or organism with new properties, however, a series of biochemical events has to occur in the affected cell, and these may be modified by environmental or physiological conditions. Evidence is accumulating that manipulation of these secondary steps in the mutation process may specifically favour the appearance of certain mutations and inhibit that of others.

BIBLIOGRAPHY. C. AUERBACH, *Genetics in the Atomic Age*, 2nd ed. (1965), a short popular book on radiation-induced mutations, and *Mutation* (1962), strictly technical for the research worker or graduate student; J.W. DRAKE, *The Molecular Basis of Mutation* (1970), restricted to micro-organisms, a textbook for the graduate or senior undergraduate student; N.P. DUBININ, *Problems of Radiation Genetics* (1964; orig. pub. in Russian, 1961), a full compilation, for the research worker; W.G. FLAMM, L. FISHBEIN, and H.L. FALK, *Chemical Mutagens: Environmental Effects on Biological Systems* (1970), a compilation by three chemists; ALEXANDER HOLLAENDER *et al.* (eds.), *Chemical Mutagens: Principles and Methods for Their Detection*, 3 vol. (1971–1973); A. LOVELESS, *Genetic and Allied Effects of Alkylating Agents* (1966), an advanced review of the subject.

Symposium volumes (for plant breeders): NATIONAL ACADEMY OF SCIENCES, NATIONAL RESEARCH COUNCIL, *Mutation and Plant Breeding* (1961); INTERNATIONAL ATOMIC ENERGY AGENCY, *Symposium on the Nature, Induction and Utilization of Mutations in Plants* (1969).

Reports: The Hazards to Man of Nuclear and Allied Radiations (HMSO 1956 and 1960); NATIONAL ACADEMY OF SCIENCES, NATIONAL RESEARCH COUNCIL, *The Biological Effects of Atomic Radiation* (1956 and 1960).

Reviews: C. AUERBACH and B.J. KILBEY, "Mutation in Eukarytotes," *A. Rev. Genet.*, 5:163–218 (1971); J.W. DRAKE, "Mutagenic Mechanisms," *ibid.*, 3:247–268 (1969).

(C.Au.)

Mycota

The Mycota, or fungi, are plantlike organisms that lack chlorophyll and are devoid of such organized plant structures as stems, roots, and leaves. Nonetheless, they are among the most successful organisms on earth and are of the greatest importance to man. They include yeasts, rusts, smuts, mildews, molds, mushrooms, and others. The bacteria, formerly considered to be fungi of a very elementary and ambiguous kind (and still so considered by some biologists), are dealt with separately (see BACTERIA). Slime molds, straddling the animal and plant worlds, are touched on in this article but are treated fully elsewhere (see SLIME MOLD).

The mushrooms, by no means the most numerous or economically significant of the fungi, are the most conspicuous members of the group; thus the Latin word for mushroom, *fungus* (plural *fungi*), has come to stand for the whole group. Similarly, the study of fungi is known as mycology—a broad application of the Greek word for mushroom, *mykēs*. Fungi other than mushrooms are sometimes collectively called molds, although this term is better restricted to fungi of the sort represented by bread mold.

IMPORTANCE TO MAN

In 1928 a green mold accidentally grew in a culture dish of *Staphylococcus* bacteria that the bacteriologist Alexander Fleming was studying in a London hospital. The fungus colony that developed inhibited the growth of the bacteria. Such unavoidable contamination certainly had occurred many times before in laboratories throughout the world, but the people who may have seen such cultures probably regarded them as contaminated plates to be discarded as soon as possible. Fleming, however, carefully recorded his observation and in 1929 published a scientific report announcing the discovery of penicillin, the first of a series of antibiotics—many of them derived from fungi—that have revolutionized medical practice.

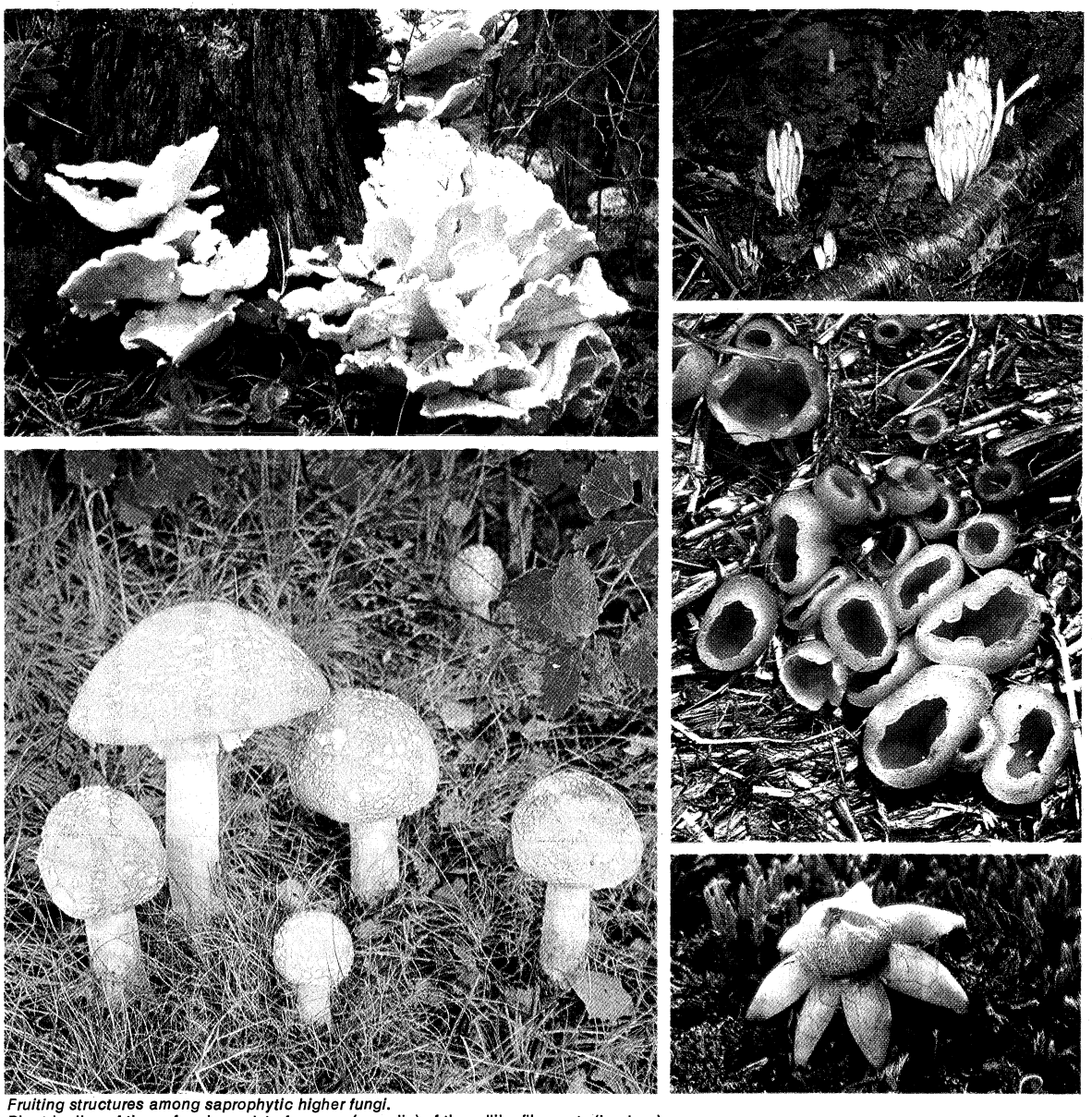
Ergotism

In 1951 a strange disease broke out in the small French village of Pont-Saint-Esprit. People became convinced that some evil spirit had entered their bodies, and several persons died. Doctors were baffled by the mysterious malady until it was recognized as a form of "St. Anthony's fire"—ergotism—that had resulted from eating bread made from contaminated flour. Ergotism was prevalent in northern Europe in the Middle Ages, particularly in regions of high rye-bread consumption; modern grain-cleaning and milling methods have practically eliminated the disease.

The cause of ergotism is ergot—a fungus. More precisely, ergot is a sclerotium (plural sclerotia), a special part of a fungus that develops on grasses and especially on rye. The wind carries the fungal spores to the flowers of the rye, where the spores germinate, infect and destroy the ovaries of the plant, and replace them with masses of microscopic threads cemented together into a hard fungal structure shaped like a rye kernel but considerably larger and darker. This is ergot—and it contains a number of poisonous organic compounds called alkaloids. A mature head of rye may carry several ergots in addition to noninfected kernels. When the grain is harvested, much of the ergot falls to the ground, but some remains on the plants and is mixed with the grain. If the ergot is not removed before milling, the ergotized flour is used to make bread and consumed; St. Anthony's fire—for which no cure is known—is the result. Nor is this the whole story. The ergot that falls to the ground may be the source of more trouble. Cattle put to graze in the rye fields after harvest are likely to consume enough ergot to bring on abortion of fetuses or death. In the spring, when the rye is in bloom, the ergot remaining on the ground produces tiny, black, mushroomlike bodies that expel large numbers of spores that start a new series of infections.

Among the many interesting chemicals in ergot is lysergic acid, the active principle of the psychedelic drug lysergic acid diethylamide (LSD). Here, then, is a single fungus that can reduce crop yields, cause abortion in cattle, sicken and sometimes kill people, and be used as a source of LSD. On the credit side, ergot provides medical science with drugs useful in inducing labour in pregnant women and in controlling hemorrhage after birth—drugs for which no good substitutes have yet been discovered.

The systematic study of fungi began 250 years ago, but



Fruiting structures among saprophytic higher fungi.
Plant bodies of these fungi consist of masses (mycelia) of threadlike filaments (hyphae) growing in soil or wood. All of the fungi shown are Basidiomycetes except *Peziza*, an ascomycete. (Top left) Bracket fungus, chicken of the woods (*Polyporus sulphureus*) growing on wood. (Top right) Club fungus (*Clavaria*) growing in soil. (Centre right) Cup fungus (*Peziza vesiculosa*) on woody forest litter. (Bottom left) Gill fungus, the poisonous fly agaric (*Amanita muscaria*), growing on soil. (Bottom right) Puffball, the earthstar (*Geastrum*), growing on moist soil among mosses.

(Top left) H.S. Knighton, (top right) Ken Brate—Photo Researchers, (centre right) H.R. Allen—EB Inc., (bottom left) Sven Samelius, (bottom right) Larry West—National Audubon Society

Universal presence of fungi

man has been indirectly aware of their activities since the first loaf of leavened bread was baked and the first tub of grape must was turned into wine. Yet, even now, few people realize that they are almost constantly either benefitted or harmed by these organisms. Fungi are everywhere—in the soil and the air; in lakes, rivers, and seas; on and within plants and animals; in food and clothing, and in the human body as well—and in very large numbers; it is this that makes them so important in the human environment. Together with bacteria, fungi are responsible for the disintegration of organic matter and the release, into the soil or atmosphere, of the carbon, oxygen, nitrogen, and phosphorus that otherwise would be forever locked up in the dead bodies of plants and animals. They are essential to many household and industrial processes, notably the making of bread, wines, beer, and certain cheeses. They are used in the production of a number of organic acids, enzymes (biological catalysts), and vitamins, and are the sources of a number of antibiotics besides penicillin. Fungi are also used as food: mushrooms, morels, and truffles are epicurean delicacies.

Fungi have had a significant role in the accumulation of biological knowledge. In 1927 the properties that make the red bread mold, *Neurospora*, an ideal organism for the study of the laws of heredity were pointed out. Geneticists and biochemists, using *Neurospora* as an experimental tool, have discovered a great deal about the manner in which genes control enzymes and about the biochemical pathways that operate in living organisms.

Although man has benefitted from the activities of certain fungi, he has also learned to look for villains among them. Numerous instances of destruction, disease, and death are directly traceable to fungi. The chestnut forests of the United States, for example, have been destroyed by the chestnut-blight fungus, and the elms in both the U.S. and Europe have been devastated by the fungus that causes Dutch-elm disease. Blights, rusts, wilts, and smuts destroy crops as man cultivates land more intensively (see DISEASES OF PLANTS)—though increasingly effective

control measures have been developed (see PEST CONTROL). Fungi cause various diseases in man—*e.g.*, athlete's foot and ringworm, candidiasis and aspergillosis, histoplasmosis and coccidioidomycosis. Mildews, molds, and rots ruin clothing and foods.

Ancient peoples were familiar with the ravages of fungi in agriculture but attributed these diseases to the wrath of the gods. The Romans even designated a particular deity, Robigus, as the god of rust and, in an effort to propitiate him, organized an annual festival, the Robigalia, in his honour.

The mushrooms, because of their size, are easily seen in the fields and forests and consequently were the only fungi known before the discovery of the microscope in the 17th century. With that discovery came the beginnings of scientific knowledge of many kinds of fungi, previously unseen, and of the obscure life processes of fungi, large and small, that had been hidden from man's eye, if not entirely from his awareness.

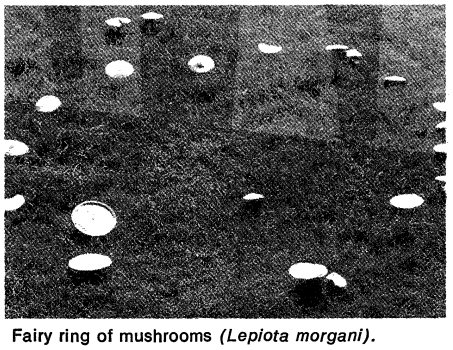
GENERAL FEATURES

Basic structure of fungi

A typical fungus consists of a mass of branched, tubular filaments that bear spores, or reproductive bodies, in various ways. The filaments, called hyphae (singular hypha), are collectively known as the mycelium, which makes up the thallus, or undifferentiated body, of the typical fungus. The mycelium grows by utilizing food materials from the environment and, upon reaching a certain stage of maturity, forms—either directly or in special fruiting bodies—spores that become detached and start new individuals. A spore thus serves the same purpose as a seed but is much simpler in structure, often consisting of a single cell. When a fungal spore germinates, it soon develops into a hypha, which grows, branches, and becomes the mycelium of the new individual.

All fungal structures are, therefore, made of hyphae or portions of hyphae; for example, a mushroom—visibly consisting of a stem supporting a cap with gills on the underside—is constructed entirely of microscopic filaments intricately associated and interwoven into specialized tissues. The mushroom is the fruiting body that bears the spores; the main body of the fungus is underground and consists of a tremendous network of hyphae—the mycelium—spread over a very large area, often several metres (yards) in diameter. This mycelium obtains food from organic matter in the soil and grows outward, just below the surface, in a circular fashion. In certain species the hyphal branches at the edge of the mycelium become organized at intervals into elaborate tissues that develop above ground into mushrooms. Such a circle of mushrooms is known as a fairy ring because, in the Middle Ages, it was believed to represent the path of dancing fairies. The ring marks the periphery of an enormous fungus colony, which, if undisturbed, continues to produce ever wider fairy rings year after year.

Walter Dawn



Fairy ring of mushrooms (*Lepiota morgani*).

Size range and diversity of structure. The part of a fungus that is generally visible is the fruiting body or sporophore. Sporophores vary greatly in size, shape, colour, and longevity. Some are microscopic and completely invisible to the unaided eye; others are no larger than a pin head; still others are gigantic structures. Among the largest sporophores are those of the mushrooms, bracket fungi, and puffballs. Some mushrooms reach a diameter of eight to ten inches and a height of ten to 12. Bracket fungi 15 inches (40 centimetres) or more in diameter are not uncommon, and puffballs often exceed that size. The largest puffballs on record measured 60 inches (150 centimetres) in diameter. The number of spores within such giants reaches several trillion.

Distribution and abundance. Fungi are found in all temperate and tropical regions of the world where there is sufficient moisture to enable them to grow. They have also been reported from the Arctic and Antarctic regions but are much more rare there, being replaced by lichens, which have a fungal component (see below *Lichen*). In general, fungi are abundant in moist habitats where organic matter is plentiful and less abundant in drier areas or in habitats with little or no organic matter. About 50,000 species of fungi have been described, but it has been estimated that there may be as many as 100,000 to 250,000 total species. Previously undescribed species are constantly being discovered and, as the tropics become more thoroughly explored, many more will undoubtedly be found.

Number of fungus species

NATURAL HISTORY

Reproduction, as contrasted with growth, is the production of new individuals resembling their parents. Characteristically, fungi reproduce by the production and dissemination of spores, a reproductive body capable of growing into a new individual. Spores may be produced either directly by asexual methods or indirectly through sexual reproduction. Sexual reproduction in fungi, as in other living organisms, involves the fusion of two nuclei that are brought together when two sex cells (gametes) unite. Asexual reproduction is simpler and more direct.

Asexual reproduction. Perhaps the simplest method of reproduction of fungi is by direct fragmentation of the thallus, the body of a fungus, which may consist of one cell or a network of filaments called a mycelium. Some yeasts, which are single-celled fungi, reproduce by simple cell division, in which one cell undergoes nuclear division and splits into two daughter cells; after some growth, these cells divide, and, eventually, a population of cells forms. In filamentous fungi the mycelium may fragment into a number of segments, each of which is capable of growing into a new individual. In the laboratory, a fragment of the mycelium is generally used to propagate a fungus. In nature, fragmentation occurs when various natural forces cause the breaking up of the mycelium in the soil or other substrata. Budding, another method of asexual reproduction, occurs in most yeasts and in some filamentous fungi. In this process, a bud develops on the surface of either the yeast cell or the hypha, its cytoplasm being continuous with that of the parent cell. The nucleus of the parent cell then divides; one of the daughter nuclei migrates into the bud, and the other remains in the parent cell. The latter is capable of producing many buds over its surface by continuous synthesis of cytoplasm and repeated nuclear divisions. After a bud develops to a certain point and even before it is severed from the parent cell, it is itself capable of budding by the same process. Thus a chain of cells may be produced. Eventually, the individual buds pinch off the parent cell and become individual yeast cells. Buds that are pinched off a hypha of a filamentous fungus behave as spores; *i.e.*, they germinate, each giving rise to a structure called a germ tube, which develops into a new hypha.

Budding

Although fragmentation, fission, and budding are methods of reproduction in a number of fungi, the majority reproduce asexually by the formation of bodies called spores. Spores produced asexually are often termed mitospores, and such spores are produced in a variety of ways. For a discussion of spores see below *Form and Function*.

Sexual reproduction. As in other living organisms, sexual reproduction in the fungi consists of three sequential stages: plasmogamy, karyogamy, and meiosis. Plasmogamy, the fusion of two protoplasts (the contents of two cells), brings together two compatible nuclei, each containing a single (haploid) set of chromosomes, in

preparation for nuclear fusion (karyogamy). In the lower fungi, karyogamy usually follows plasmogamy almost immediately. In the more evolved fungi, however, karyogamy is separated from plasmogamy. Karyogamy results in the formation of a diploid nucleus (*i.e.*, a nucleus containing two sets of chromosomes, one from each parent). The cell in which karyogamy occurs is called the zygote. In most fungi, the zygote is the only cell in the entire life cycle that is diploid. Once karyogamy has occurred, meiosis (cell division resulting in reduction of the chromosome number to one set per cell) generally follows immediately and restores the haploid phase. Either the haploid nuclei that result from meiosis or their immediate progeny are generally incorporated in spores termed meiospores.

Fungi employ a variety of methods to accomplish plasmogamy. Some produce specialized sex cells (gametes) that are released from differentiated sex organs called gametangia. In other fungi, the gametangia come in contact, and nuclei pass from the male gametangium into the female, thus assuming the function of gametes. In still other fungi, the gametangia themselves may fuse in order to bring their nuclei together. Finally, some of the most advanced fungi produce no gametangia at all; the somatic (vegetative) hyphae take over the sexual function, come in contact, fuse, and exchange nuclei.

Fungi in which a single individual bears both male and female gametangia are hermaphroditic fungi. Rarely, gametangia of different sexes are produced by separate individuals, one a male, the other a female. Such species are termed dioecious. Dioecious species usually produce sex organs only in the presence of an individual of the opposite sex.

Incompatibility. Many fungi produce clearly differentiated male and female organs on the same thallus, but the individual thalli, originating from single uninucleate spores, are sexually self-sterile because their sex organs are incompatible. Such fungi require the presence of thalli of different mating types in order for sexual fusion to take place. The simplest form of this mechanism occurs in fungi in which there are two mating types, often designated + and − (or *A* and *a*). Gametes produced by one type of thallus are compatible only with gametes produced by the other type. Such fungi are said to be heterothallic. Many fungi, however, are homothallic; *i.e.*, sex organs produced by a single thallus are self-compatible, and a second thallus is unnecessary for sexual reproduction. Some of the most complex fungi (*e.g.*, mushrooms) do not develop differentiated sex organs; rather, the sexual function is carried out by their somatic hyphae, which unite and bring together compatible nuclei in preparation for fusion. Homothallism and heterothallism are encountered in these fungi, as well as in those in which sex organs are easily distinguishable. Compatibility, therefore, refers to a physiological differentiation and sex to a morphological (structural) one; the two phenomena, although related, are not synonymous.

Sexual hormones. Various mechanisms involving organic substances called hormones induce formation of sex organs or bring together gametes for sexual union. In *Allomyces* (Chytridiomycetes) a hormone named sirenin, secreted by the female gametes, attracts the male gametes, which swim toward the former and fuse with them. In *Achlya* (Oomycetes) one or more sex hormones induce the formation of gametangia and attract the male to the female. In the Zygomycetes, in which the gametangia are usually not differentiated structurally, a hormonal mechanism brings together hyphal tips of opposite mating types. In *Ascobolus* (Ascomycetes) the receptive branch (trichogyne) of the female gametangium actively turns toward the nonmotile male gamete (spermatium), which secretes an attractant.

Attraction of sex cells to each other by hormones

Life cycle. In the life cycle of a sexually reproducing fungus, a haploid phase alternates with a diploid phase. The haploid phase ends with nuclear fusion, and the diploid phase begins with the formation of the zygote (the cell resulting from fusion of two sex cells). A special kind of nuclear division, termed meiosis or reduction division, restores the haploid number of chromosomes and initiates the haploid phase, which produces the gametes. In the majority of fungi, all structures are haploid except the zygote. Nuclear fusion takes place at the time of zygote formation, and meiosis follows immediately. Only in the water mold *Allomyces* and a few related genera, and in some yeasts is alternation of a haploid thallus with a diploid thallus definitely known. In the class Oomycetes the place of meiosis in the life cycle is in doubt. If, as is probable, the thallus in this class is diploid and meiosis takes place just before the formation of the gametes, then the gametes would be the only haploid structures in the entire life cycle, and the Oomycetes would be unique among the fungi in this respect.

In the higher fungi, a third condition is interspersed between the haploid and diploid phases of the life cycle. In these fungi, plasmogamy (fusion of the cellular contents of two hyphae) results in dikaryotic hyphae in which each cell contains two nuclei, one from each parent. Eventually, a nuclear pair fuses to form the zygote nucleus. In the Basidiomycetes, the most advanced class of fungi, the binucleate cells divide successively and give rise to a binucleate mycelium, which is the main assimilative phase of the life cycle. It is the binucleate mycelium that eventually forms the basidia in which nuclear fusion and meiosis take place prior to the formation of the basidiospores.

As stated before, fungi usually reproduce both sexually and asexually. The asexual cycle produces mitospores, and the sexual cycle produces meiospores. Even though both types of spores are produced by the same mycelium, they are very different in form and easily distinguished. The asexual phase usually precedes the sexual phase in the life cycle and may be repeated frequently before the sexual phase appears.

Some fungi differ from others in their lack of one or the

From C.J. Alexopoulos, *Introductory Mycology*, 2nd ed. Copyright © 1962 by John Wiley & Sons, Inc. Reprinted by permission

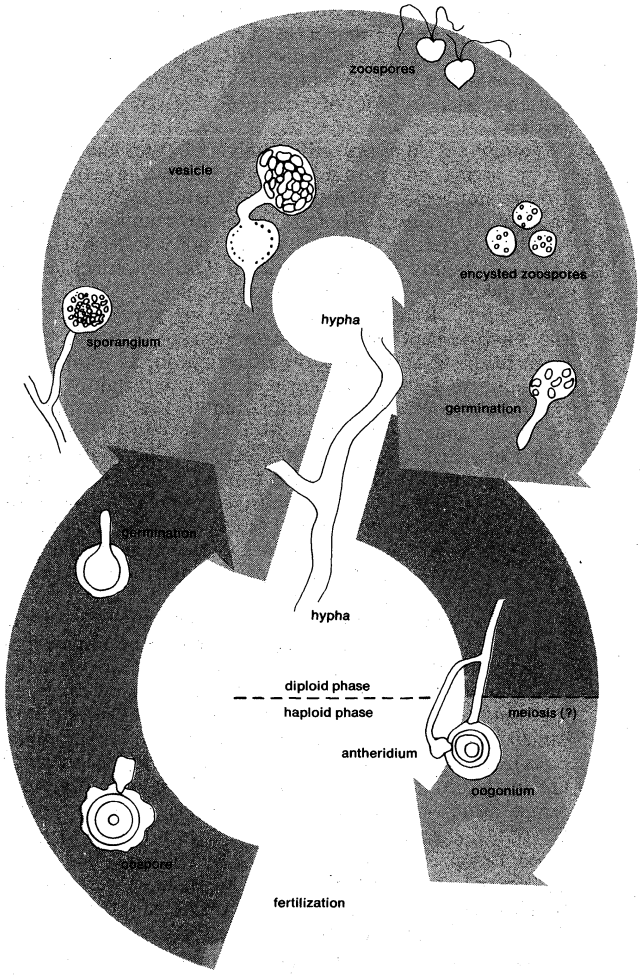


Figure 1: Life cycle of *Pythium debaryanum*, an oomycete.

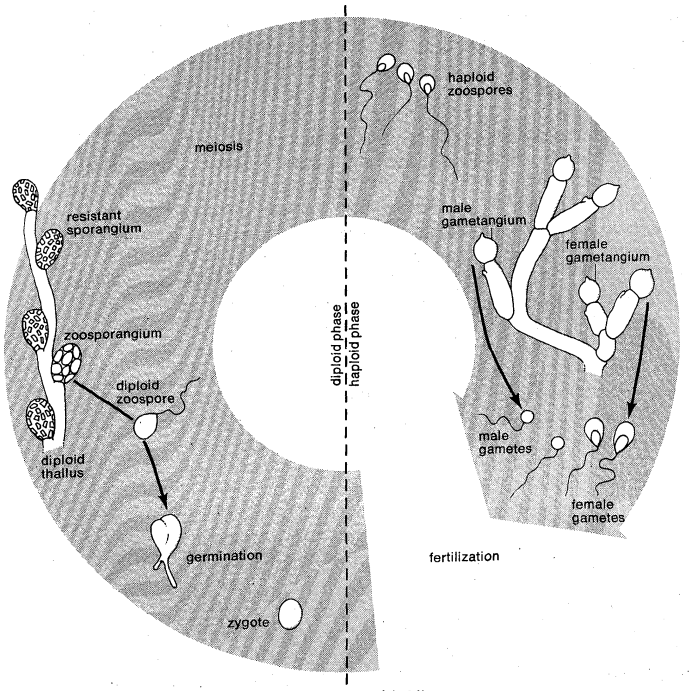


Figure 2: Life cycle of water mold *Allomyces macrogynus*, a chytridiomycete.

From C.J. Alexopoulos, *Introductory Mycology*, 2nd ed., copyright © (1962) by John Wiley and Sons, Inc.; reprinted by permission

other of these stages; some reproduce only sexually (except for fragmentation, which is common in most) and many only asexually. A number of fungi exhibit the phenomenon of parasexuality in which processes comparable to plasmogamy, karyogamy, and meiosis take place but not at a specified time or at specified points in the life cycle of the organism. Parasexuality is characterized by the prevalence of heterokaryosis in a mycelium; *i.e.*, the presence, side by side, of nuclei of different genetic composition.

Parasexuality

Ecology. Fungi are either terrestrial or aquatic, the latter living in freshwater or marine environments. Freshwater species usually do not tolerate high degrees of salinity in nature, although some are found in slightly brackish water; most prefer clean, cool waters, but a few thrive in highly polluted streams. The soil furnishes an ideal habitat for a large number of species.

Saprobes. Relatively little is known of the effects of the environment on the distribution of saprobic fungi (*i.e.*, fungi utilizing dead organic material as food). The availability of elaborated, organic food is certainly one of the factors controlling such distribution. A great number of fungi appear able to utilize most types of organic materials, such as lignin, cellulose, or other polysaccharides, which have been added to soils or waters by dead vegetation. It follows that most saprobic fungi may be expected to be cosmopolitan (widely distributed), at least in habitats with a sufficient organic content to support fungal growth. Fungi with specific nutritional requirements are, of course, more localized.

Moisture and temperature are two ecological factors, in addition to organic matter, that are of prime importance in determining distribution of fungi. Laboratory studies have shown that many, perhaps the majority, of fungi are mesophilic; *i.e.*, have an optimum growth temperature of 20–30° C (68–86° F). Some species (*i.e.*, thermophilic ones) are able to grow at 50° C (122° F) or higher but are unable to grow below 30° C. Although the optimum temperature for growth of most fungi lies at or above 20° C (68° F), a large number of species are able to grow close to or below 0° C (32° F). The so-called snow molds and the fungi that cause spoilage of refrigerated foods are examples of this group. Obviously, temperature relationships influence the distribution of various species. Certain other effects of temperature are also important determining factors in the habitats of fungi. Many coprophilous (dung-inhabiting) fungi, for example, although able to grow at a temperature of 20–30° C, require a short period at 60° C (140° F) for their spores to germinate.

Whereas a great many saprobes are cosmopolitan, some are strictly tropical, others are strictly temperate-zone forms. Whether temperature is the main factor in determining such distribution is not known.

Plant parasites. The distribution of parasitic species is closely correlated with the distribution of the plant hosts with which they are associated. Fungi that parasitize a variety of plant hosts are likely to be more widely distributed than the more specialized parasites. It is of interest to note in this connection that certain fungi common in temperate regions are also found in the tropical highlands where their hosts can grow.

A fungus that occurs both in the temperate zone and in the tropical lowlands often completes its life cycle in one of the two environments but not in the other. The powdery mildews (family Erysiphaceae), as a group, for example, go through both asexual and sexual cycles in the temperate zones, producing mitospores (conidia) and meiospores (ascospores) respectively; seldom do they form sexual stages in the tropics. Whether this is a matter of temperature or some other factor is not yet known.

Human pathogens. Some fungi that cause human diseases, such as *Candida albicans*, are universally distributed; others, like *Paracoccidioides brasiliensis*, which is confined to South and Central America, have a more limited distribution. The factors that affect the distribution of the human pathogens are unknown. The great increase in international travel, migration, and troop movements in the second half of the 20th century undoubtedly were major factors in the spread of these fungi.

Symbionts. Among symbiotic fungi, those that enter into mycorrhizal relationships (see below) and the lichen fungi, are probably the best known. Among the mycorhizal fungi are the boletes (Basidiomycetes, family Boletaceae), whose mycorrhizal relationships with larch trees (*Larix*) and other conifers have long been known; others include the truffles (Ascomycetes, order Tuberales), some of which are believed to form mycorrhizae with oak (*Quercus*) or beech (*Fagus*) trees, and various species of *Rhizoctonia* (mycelial stage of certain Basidiomycetes), which form orchid mycorrhizae.

Lichens are universally distributed, abundantly in some areas, more sparsely in others. In the Arctic regions, lichens are a dominant part of the flora. The same is true in some tropical regions. Lichens are pioneering organisms in that they are among the first to colonize lava flows and can subsist on bare rock exposed to the sun in very dry areas where few other organisms can survive. On the other hand, some are also found in the tropical rain forest together with mosses and liverworts. Lichens are particularly sensitive to changes in light conditions. A lichen community can virtually disappear or certainly change in species composition if it is shaded by abundant overgrowth of shrubs or trees. Most of the lichen fungi are Ascomycetes specialized in their association with their algal hosts. A few are Basidiomycetes.

Fungus–insect associations. Another kind of symbiotic relationship between fungi and host organisms is exemplified by that of the fungal genus *Septobasidium* and certain scale insects (order Homoptera) that feed on trees. The mycelium forms elaborate structures over colonies of insects feeding on the bark. Each insect sinks its proboscis (tubular sucking organ) into the bark and remains there the rest of its life, sucking sap. The fungus sinks haustoria (special absorbing structures) into the bodies of some of the insects and feeds on them without killing them. The parasitized insects are, however, rendered sterile.

Two-stage parasitism of scale insects

The perpetuation of the insect species and the spread of the fungus are accomplished by the uninfected members of the colony, which live in fungal "houses," safe from enemies. Newly hatched scale insects crawl over the surface of the fungus, which is at that time sporulating. Fungal spores adhere to the young insects, germinate, and infect them. As the young insects settle down in a new

place on the bark to begin feeding, they establish new fungal colonies. Thus, part of the insect colony is sacrificed to the fungus as food in return for the fungal protection provided for the rest of the insects. The insect is parasitic on the tree, and the fungus is parasitic on the insect, but the tree is the ultimate victim.

Ants of the tribe Attini feed exclusively on fungi, which they cultivate in underground gardens. The leaf-cutting ants in this group bring in sections of freshly cut leaves and flowers and chew small pieces about one millimetre (1/25 inch) in diameter into pulp. Anal droppings are then placed on the bits of pulp, which are planted in the garden in the centre of the nest. Fungal mycelium from an old garden is then transferred to the newly planted material and allowed to grow. Other fungi are kept out of the garden. The entire ant colony subsists on the cultivated fungus. Four kinds of fungi, *Leucocoprinus gongylophora, Tyridomyces formicarum,* and species of *Lepiota* and *Auricularia* are known to be so cultivated—and are not known to occur naturally other than in ant gardens.

The order Laboulbeniales includes minute Ascomycetes that are anchored on the outer skeleton of insects; they seldom penetrate beyond it and do not cause disease of the host. Specialization is extreme in some instances; a particular species of fungus may grow only on individual insects of one sex and often only on a particular organ.

The sooty molds constitute another interesting ecological group of fungi associated with insects. The majority are tropical or subtropical, but some species occur in the temperate zones. All sooty molds are epiphytic (*i.e.*, they grow on the surfaces of other plants); but only in areas where scale insects are present. The fungi parasitize neither the plants nor the insects, but obtain their nourishment exclusively from the honeydew secretions of the scale insects. Growth of the dark mycelium over the plant leaves, however, is often so dense as to reduce greatly the intensity of the light that reaches the leaf surface; this in turn significantly reduces the rate of photosynthesis.

FORM AND FUNCTION

Structure of the thallus. In almost all fungi the hypha, and therefore the thallus—the undifferentiated, nutrient-absorbing body of the fungus—has cell walls. (The thalli of the true slime molds lack cell walls and are sometimes classified as protozoans rather than fungi.) A hypha is a many-branched tube lined or filled with a jellylike substance called cytoplasm. The tube itself may be either continuous throughout or divided into compartments, or cells, by cross walls called septa (singular septum). In nonseptate (*i.e.*, coenocytic) hyphae, the nuclei are scattered throughout the cytoplasm. In septate hyphae, each cell may contain one, two, or many nuclei, depending on the type of fungus or the stage of hyphal development. The cells of fungi are similar in structure to those of many other organisms. The minute nucleus, readily seen only in young portions of the hypha, is surrounded by a double membrane and typically contains one nucleolus. In addition to the nucleus, various other components are embedded in the cytoplasm.

The wall of the hypha is complex in both composition and structure. Its exact chemical composition varies in different fungal groups. In some fungi the wall contains considerable quantities of cellulose, a complex carbohydrate that is the chief constituent of the cell walls of plants; in most fungi, however, chitosan and chitin, horny carbohydrate substances similar to cellulose, are present in the walls, to the exclusion of cellulose. In at least three fungi (*Rhizidiomyces apophysatus, Apodachlya* species, and *Ceratocystis ulmi*), and probably in others, both chitin and cellulose are components of the cell wall. Among the many other chemical substances in the walls of fungi are some that may thicken or toughen the wall of some tissues, thus imparting rigidity and strength. The wall is not a passive, dead envelope; rather there is reason to believe that it may have a regulatory function in the interchange of materials between the hypha and its external environment. The chemical composition of the wall of a particular fungus may vary at different stages of the organism's growth—a possible indication that the wall plays some part in determining the form of the fungus. In some yeasts, fusion of sexually functioning cells is brought about by the interaction of specific chemical substances on the walls of two compatible mating types.

When the mycelium grows in or on a surface—the soil, a log, a culture medium—it appears as a mass of loose, cottony threads. The richer the composition of the growth medium, the more profuse the threads and the more feltlike the mass. On the sugar-rich growth substances used in scientific laboratories, the assimilative (somatic) hyphae are so interwoven as to form a thick, almost leathery colony. On the soil, inside a leaf, in the skin of animals, or in other parasitized plant or animal tissues, the hyphae are usually spread in a loose network. The mycelium of the so-called higher fungi does, however, become organized at times into compact masses of different sizes that serve various functions. Some of these masses, called sclerotia, become extremely hard and serve to carry the fungus over periods of adverse conditions of temperature and moisture. Ergot is an example. The underground sclerotia of *Poria cocos,* an edible pore fungus known in the United States under the Indian name tuckahoe, may reach a diameter of 20 to 25 centimetres (8 to 10 inches). Various other tissues are also produced by the interweaving of the assimilative hyphae of some fungi. Stromata (singular stroma) are cushion-like tissues that bear spores in various ways. Rhizomorphs are long strands of parallel hyphae cemented together. Those of the honey mushroom *Armillariella mellea,* which are black and resemble shoestrings, are intricately constructed and are differentiated to conduct water and food materials from one part of the thallus to another.

Sporophores and spores. When the mycelium of a fungus reaches a certain stage of growth, it begins to produce spores either directly on the somatic hyphae or, more often, on special sporiferous (spore-producing) hyphae, which may be loosely arranged or grouped into intricate structures, the fruiting bodies or sporophores. The type of sporophore produced is characteristic of the group to which a fungus belongs, and the classification of fungi is based almost entirely on the characters of their sporophores and spores.

Fruiting structures as determinants of classification

The lower fungi (*i.e.*, more primitive) produce spores in sporangia, which are saclike sporophores whose entire cytoplasmic contents cleave into spores, called sporangiospores. Sporangiospores are either naked and flagellated (zoospores) or walled and nonmotile (aplanospores). The more primitive aquatic and terrestrial fungi tend to produce zoospores. The zoospores of aquatic fungi swim in the surrounding water by means of one or two variously located flagella (whiplike organs of locomotion), depending on the group to which the fungus belongs. Zoospores produced by terrestrial fungi are released after a rain from the sporangia in which they are borne and swim for a time in the rainwater between soil particles or on the wet surfaces of plants, where the sporangia are formed by parasitic fungi. After some time, the zoospores lose their flagella and surround themselves with walls; each zoospore germinates by a germ tube that grows into a system of hyphae. The bread molds (Zygomycetes), which are the most advanced of the lower fungi, produce only aplanospores (nonmotile spores) in their sporangia.

The higher fungi (Ascomycetes, Deuteromycetes, and Basidiomycetes) do not produce motile spores of any kind even though some of them are aquatic in fresh or marine waters. In these fungi, asexually produced spores (usually termed conidia) are typically formed terminally or laterally on special spore-producing hyphae, the conidiophores, which are variously arranged; *i.e.*, singly on the hyphae or grouped in special asexual fruiting bodies such as flask-shaped pycnidia, mattress-like acervuli, cushion-shaped sporodochia, or sheaflike synnemata.

Sexually produced spores of the higher fungi result from meiosis and are formed either in saclike structures (asci) or on the surface of typically club-shaped structures (basidia). Asci and basidia may be borne naked, directly on the hyphae, or in various types of sporophores, termed ascocarps or basidiocarps, depending on whether they bear asci or basidia. Well-known examples of ascocarps

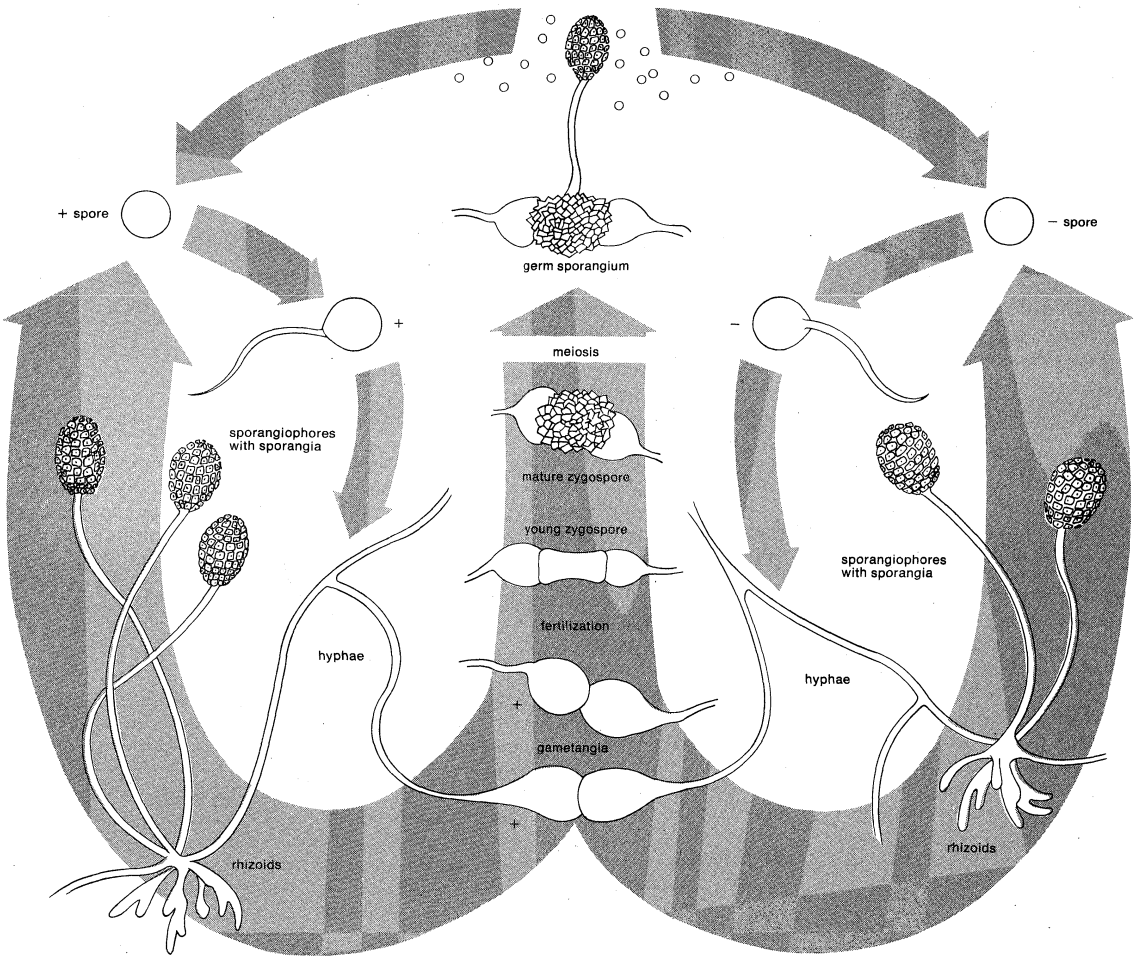


Figure 3: Life cycle of the common bread mold *Rhizopus stolonifer*, a zygomycete.
From C.J. Alexopoulos, *Introductory Mycology*, 2nd ed., copyright © (1962); John Wiley and Sons, Inc., reprinted by permission

are the morels, the cup fungi, and the truffles. Commonly encountered basidiocarps are mushrooms, puffballs, stinkhorns, and bird's-nest fungi.

Growth. Under favourable environmental conditions, fungal spores germinate and form hyphae. During this process, the spore absorbs water through its wall, the cytoplasm becomes activated, nuclear division takes place, more cytoplasm is synthesized, a break appears in the outer wall of the spore, and from the wall a germ tube bulges out, already enveloped by a wall of its own that is formed as the tube grows. The growing tip of a hypha may be roughly divided into three regions: (1) the apical zone about eight microns (.008 millimetres or .0003 inches) in length, in which all growth occurs and in which an accumulation of cytoplasmic vesicles (tiny bubbles or spheres) takes place; (2) the subapical region, extending about 40 microns back of the apical zone, which is rich in cytoplasmic components, such as dictyosomes (Golgi bodies—particulate or netlike components of the cytoplasm that function as secretory structures) and vesicles, but is devoid of vacuoles; and (3) the zone of vacuolation, which is characterized by the presence of many vacuoles and the accumulation of lipids. The growing tip eventually gives rise to a branch. This is the beginning of the branched mycelium. Growing tips that come in contact with neighbouring hyphae often fuse with them to form a hyphal net. In such a vigorously growing system, the cytoplasm is in constant motion, streaming toward the growing tips. Eventually, the older hyphae become highly vacuolated and may be stripped of most of their cytoplasm. All living portions of a thallus are potentially capable of growth. If a small piece of mycelium is placed under conditions favourable for growth, it develops into a new thallus even if no growing tips are included in the severed portion.

Regions of growth in a fungus hypha

Growth of a septate mycelium (*i.e.*, with cross walls between adjacent cells) entails the formation of new septa in the young hyphae. Septa are formed by ringlike growth from the wall of the hypha toward the centre until the septa are complete. In the higher fungi, the septum stops growing before it is complete; the result is a central pore through which the cytoplasm flows, thus establishing organic connection throughout the thallus. In the mycelium of most Basidiomycetes (*e.g.*, mushrooms, puffballs), the septa are much more elaborate, the central pore being bordered by a barrel-shaped, inflated structure that appears to function as a valve.

The individual fungus is potentially immortal, for it continues to grow at the hyphal tips as long as conditions remain favourable. It is possible that, in undisturbed places, mycelia exist that have grown continuously for many thousands of years. The older parts of the hyphae die and decompose, releasing nitrogen and other nutrients into the soil. Thus, if a fairy ring develops in a pasture, city lawn, or golf course, the grass just inside the circle of mushrooms, fertilized by dead hyphae, may be noticeably greener than the grass outside the ring.

Optimum conditions for hyphal growth vary with the species. Generally speaking, acid growth substances and temperatures near 20° to 30° C (68° to 86° F) are most favourable. Some species are thermophilic (*i.e.*, grow best in elevated temperatures) and grow little, if at all, at temperatures below 40° C (104° F). Temperatures slightly above freezing (32° F) inhibit but do not altogether prevent fungal growth, as is shown by the spoilage of foods in a refrigerator. High humidity enhances the growth of the hyphae. Light seems to have little effect on growth but is required by some fungi for sporulation.

Nutrition. Since fungi possess no chlorophyll—the pigment molecules of green plants—they are unable to

carry out photosynthesis and must obtain carbohydrates —the products of photosynthesis—from materials in which they are stored in the form of sugars, starches, celluloses, hemicelluloses, and lignins. Fungi therefore lead either a saprobic or a parasitic existence; in the former case they obtain their food from dead organic material, in the latter by feeding on living organisms (usually plants), thus causing disease.

External digestion of food

Fungi secure food through the action of enzymes (biological catalysts) secreted into the surface on which they are growing; the enzymes digest the food, which then is absorbed directly through the hyphal walls. The rotting of fruits, such as peaches and citrus fruits in storage, demonstrates this phenomenon, in which the infected parts are softened by the action of the fungal enzymes. In brown rot of peaches, the softened area is somewhat larger than the actual area invaded by the hyphae: the periphery of the brown spot has been softened by enzymes that act ahead of the invading mycelium. Food must enter the hyphae in solution, and, since most fungi have no special absorbing organs, the entire mycelial surface is capable of taking in materials dissolved in water. Some fungi, however, produce special rootlike hyphae, called rhizoids, which anchor the thallus to the growth surface and probably also absorb food. Many parasitic fungi are even more specialized in this respect, producing special absorptive organs called haustoria.

Saprobiosis. Together with the bacteria, saprobic fungi are to a large extent responsible for the decomposition of organic matter. They are also responsible for the decay and decomposition of foodstuffs. Among other destructive saprobes are fungi that destroy timber and timber products as their mycelia invade and digest the wood; many of these produce their spores in large, woody, fruiting bodies—*e.g.*, bracket or shelf fungi. Paper, textiles, and leather are often attacked and destroyed by fungi. This is particularly true in tropical regions, where temperature and humidity are often very high. Photographic film used in the tropics is specially packaged in an effort to exclude fungi. Fungal growth on lenses in binoculars, cameras, telescopes, and other optical equipment is a serious problem in regions where humidity is high.

The nutritional requirements of saprobes (and of some parasites that can be cultivated artificially) have been determined by growing fungi experimentally on various synthetic substances of known chemical composition. The fungi seem to exhibit the same morphological characteristics in these culture media as they do in nature. Carbon is supplied in the form of sugars or starch; the majority of fungi thrive on such sugars as glucose, fructose, mannose, maltose, and to a lesser extent, sucrose. Decomposition products of proteins—*e.g.*, proteoses, peptones, and amino acids—can be used by most fungi as nitrogen sources; ammonium compounds and nitrates also serve for many species. It is doubtful, however, that any fungus can combine, or fix, atmospheric nitrogen into usable compounds. Chemical elements such as phosphorus, sulfur, potassium, magnesium, and small quantities of iron, zinc, manganese, and copper are needed by most fungi for vigorous growth; elements such as calcium, molybdenum, and gallium are required by at least some species. Oxygen and hydrogen are absolute requirements; they are supplied in the form of water or are obtained from carbohydrates. Many fungi, deficient in thiamine and biotin, must obtain these vitamins from the environment; most fungi appear able to synthesize all other vitamins necessary for their growth and reproduction. As a rule, fungi are aerobic organisms; *i.e.*, they require free oxygen. Fermentations, however, take place under anaerobic conditions as well. Knowledge of the physiology of saprobic fungi has enabled industry to use several species for fermentation purposes.

Parasitism. In contrast with the saprobic fungi, parasitic fungi attack living organisms, penetrate their outer defenses, invade them, and obtain nourishment from living cytoplasm, causing disease and sometimes the death of the host. Most pathogenic (disease-causing) fungi are parasites of plants, but several are known to cause diseases of the lower animals and man. Most parasites enter the host through a natural opening, such as a stomate (microscopic air pore) in a leaf, a lenticel (small opening through bark) in a stem, a broken plant hair or a hair socket in a fruit, or a wound in the plant or animal epidermis (skin). Such wounds may be insect punctures or accidentally inflicted scratches, cuts, or bruises.

Fungus-caused diseases

Among the most common and widespread diseases of plants caused by fungi are the various downy mildews—*e.g.*, of grape, onion, tobacco; the powdery mildews—*e.g.*, of grape, cherry, apple, peach, rose, lilac; the smuts—*e.g.*, of corn, wheat, onion; the rusts—*e.g.*, of wheat, oats, beans, asparagus, snapdragon, hollyhock; apple scab; brown rot of stone fruits; and various leaf spots, blights, and wilts. These diseases cause millions of dollars' worth of damage annually all over the world.

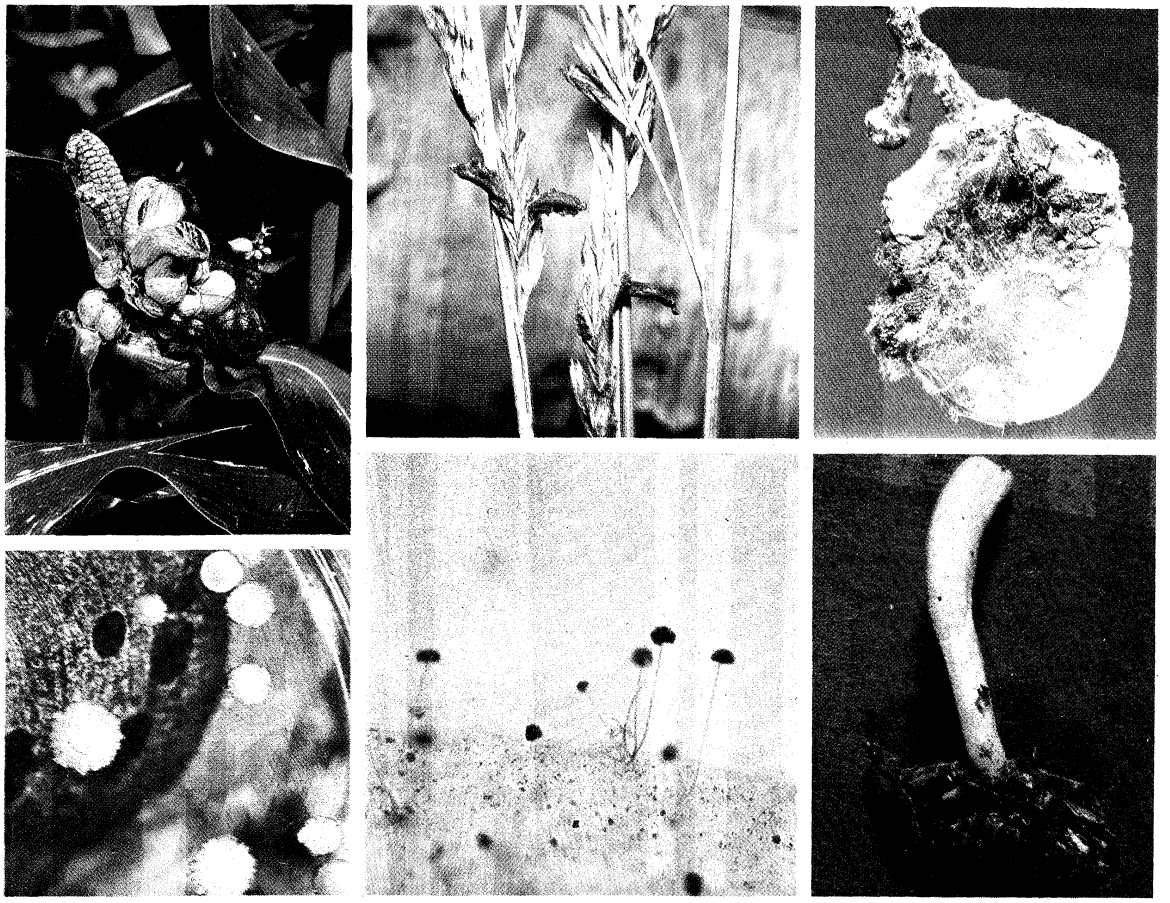
Infection of a plant takes place when the spores of a pathogenic fungus fall on the leaves or the stem of a susceptible host and germinate, each producing a germ tube. The tube grows on the surface of the host until it finds an opening; then the tube enters the host, puts out branches between the cells of the host, and forms a mycelial network within the invaded tissue. The germ tubes of some fungi produce special pressing organs called appressoria, from which a microscopic, needlelike peg presses against and punctures the epidermis of the host; after penetration, a mycelium develops in the usual manner. Many parasitic fungi absorb food from the host cells through the hyphal walls appressed against the cell walls of the host's internal tissues. Others produce haustoria (special absorbing structures) that branch off from the intercellular hyphae and penetrate the cells themselves. Haustoria, which may be short, bulbous protrusions or large branched systems filling the whole cell, are characteristically produced by obligate (*i.e.*, invariably parasitic) parasites; some facultative (*i.e.*, occasionally parasitic) parasites such as the potato blight, *Phytophthora infestans*, also produce them. Obligate parasites, which require living cytoplasm and have extremely specialized nutritional requirements, cannot be grown in artificial culture under presently known conditions; their physiology, therefore, is not yet well understood. Examples of obligate parasites are the downy mildews and the powdery mildews.

Predation. A number of fungi have developed ingenious mechanisms for trapping animals of microscopic size, such as amoebas, roundworms (nematodes), and rotifers. After an animal is captured, the fungus penetrates its body and quickly destroys it. Many of these fungi secrete adhesive substances over the surface of their hyphae, so that a passing animal that touches any portion of the mycelium adheres firmly to it. Soon a penetration tube grows out of a hypha and penetrates the host's soft body. This haustorium grows and branches, and enzymes secreted by it quickly kill the animal, whose cytoplasm serves as food for the fungus.

Animal-trapping fungi

Other fungi produce hyphal loops in which small animals become ensnared until the fungus is able to send haustoria into their bodies and kill them. Perhaps the most amazing of these fungal traps are the so-called constricting rings of some species of *Arthrobotrys*, *Dactylella*, and *Dactylaria*—soil-inhabiting fungi easily grown under laboratory conditions. In the presence of nematodes, the mycelium produces large numbers of rings through which the average nematode is barely able to pass. When a nematode rubs the inner wall of a ring, which usually consists of three cells with touch-sensitive inner surfaces, the cells of the ring swell rapidly, and the resulting constriction holds the worm tightly. All efforts of the nematode to free itself fail, and a hypha, which grows out of one of the swollen ring cells at its point of contact with the worm, penetrates the animal's body, branches therein, and kills the host, which is then used for food by the fungus. In the absence of nematodes, these fungi do not usually produce rings in appreciable quantities. A substance (nemin) of largely unknown chemical composition is secreted by the nematodes and stimulates the fungus to form the mycelial rings.

Mycorrhiza. A large number of fungi "infect" the roots of plants forming an association with them called my-



Fungi of economic or scientific interest.
(Top left) Corn smut (*Ustilago*) on *Zea mays* yearly destroys a considerable percentage of the corn (maize) crop. (Top centre) Ergot (*Claviceps purpurea*), the original source of LSD-25 as well as a drug used in obstetrics, attacks rye and other cereal grains and causes the disease ergotism when eaten. (Top right) Blue-green mold (*Penicillium*), shown on grape, attacks many fruits and is the source of the drug penicillin. (Bottom left) Brewers' yeast colonies (*Saccharomyces cerevisiae*) growing on agar medium. (Bottom centre) Bread mold (*Rhizopus stolonifer*) showing the sporangia that bear conidia (asexual spores). (Bottom right) Parasitic *Cordyceps militaris* attacking insect pupa.
(Top left) J.C. Allen and Son, (top centre, top right) Walter Dawn, (bottom left) A to Z Botanical Collection—EB Inc., (bottom centre) Grant Heilman—EB Inc., (bottom right) F.K. Anderson—EB Inc.

corrhiza or mycorhiza. This association differs markedly from ordinary root infection, which is responsible for root-rot diseases. Mycorrhizal fungi establish a mild form of parasitism that in many instances verges on mutualism; *i.e.*, it is beneficial to the plant as well as to the fungus.

Lichens. A lichen is a "composite plant" consisting of an alga and a fungus in intimate association. The fungus forms a characteristic stroma (the body of the lichen) in which the algal cells are embedded. The fungal hyphae penetrate many of the algal cells with haustoria and therefore "parasitize" them; but, as in mycorrhizae, the relationship is benign in that the algal cells are protected by the fungal mantle from too much light, in exposed situations, and perhaps also from extreme desiccation (see LICHEN).

EVOLUTION AND PHYLOGENY

The origin of the fungi is obscure, the fossil record being scanty and virtually meaningless. The older theory supposed the fungi to have originated by loss of chlorophyll from one or two groups of algae, one school favouring the development of the fungi monophyletically (*i.e.*, from a single ancestor) from the green algae, the other postulating that the lower fungi originated from the green algae but that the Ascomycetes came from the class Florideae of the red algae and later gave rise to the Basidiomycetes. Most present-day mycologists derive the fungi from ancestral flagellates (algal or protozoan organisms bearing flagella, whiplike swimming organs) but yield that the Oomycetes may have had a separate origin because of the supposed absence of chitin from, and the known presence of cellulose in, their cell walls; their probably diploid thallus; and their sexual reproduction through oogamy (production of differentiated egg cells).

As for the interrelationships among fungal groups, there is much controversy. The modern tendency is to emphasize flagellation as an important phylogenetic criterion in the lower fungi. Thus, all the posteriorly uniflagellate fungi (*i.e.*, those with a single flagellum located at the rear end of the organism) are brought together in one class, Chytridiomycetes; all the anteriorly uniflagellate in the class Hyphochytridiomycetes, and so on, as may be seen in the section on classification below. Whether the Plasmodiophoromycetes should be grouped with the Myxomycetes (slime molds) and removed from the fungi because of their plasmodial phase and their swarm cells, which bear two anterior whiplash flagella, is a moot question.

Those who favour a monophyletic development derive the Zygomycetes from the Chytridiomycetes through *Zygochytrium*, in which sexual reproduction is said to be zygomycetous (*i.e.*, of the same type as is found in the class Zygomycetes), and the Ascomycetes from the Zygomycetes), through the genus *Dipodascus*, which has a many-spored "ascosporangium." The yeasts are considered to be either primitive or advanced (hence simple by reduction), depending on how one wishes to derive them.

Virtually all mycologists agree that the Basidiomycetes have been derived from the Ascomycetes. This opinion is based on the similarity of the nuclear cycle of the ascus

and basidium, the supposed homology of the clamp connection (a structure joining two adjacent cells in the Basidiomycetes) with the crozier (a hook-shaped terminal cell found in Ascomycetes), and the similarity of the binucleate mycelium of Basidomycetes with the ascogenous (ascus-producing) hyphae of Ascomycetes. Not all, however, are agreed as to which group of Ascomycetes gave rise to the Basidiomycetes, nor indeed as to which Basidiomycetes are primitive and which are advanced. Whether the holobasidium (simple, club-shaped basidium) or the heterobasidium (septate or deeply divided basidium) came first is a highly controversial question that has a great bearing on the origin of the Basidiomycetes. Also, the fact that the dolipore (inflated) septum has not been found in the rusts or smuts or in any clearly nonbasidiomycetous group poses an interesting question on the origin of the Basidiomycetes.

CLASSIFICATION

Distinguishing taxonomic features. The fungi as a group are distinguished from other organisms by the nature of their somatic (body) and reproductive structures, and by the mode of nutrition they employ. Within the division Mycota, the classes are further distinguished by variations of these characteristics, particularly those involving reproductive stages.

Annotated classification. The following classification is adapted from G.W. Martin in Ainsworth's *Dictionary of the Fungi*, 5th ed. (1961).

DIVISION MYCOTA (fungi)

Eukaryotic (with true nuclei), achlorophyllous (without chlorophyll), acellular, unicellular, or multicellular organisms; microscopic or macroscopic in size; usually with cell walls and filaments; typically reproducing by spores produced asexually or sexually; walls containing chitin, cellulose, or both, among other substances. About 50,000 living species; fewer than 500 fossil species known.

Subdivision Myxomycotina (true slime molds)

Assimilative stage a multinucleate, free living, essentially naked, acellular, mobile plasmodium (jellylike mass), which becomes converted into one or more characteristic sporocarps; holozoic (feeding like an animal; *i.e.*, by engulfing solid food particles); sexual reproduction by fusion of compatible myxamoebae (motile amoeba-like cells) or flagellated cells; spores enveloped in definite cell walls containing cellulose.

Class Myxomycetes (slime molds)

Characters of the subdivision. Cosmopolitan, terrestrial, sporulating on any substratum on which the plasmodium settles, such as living plants, plant debris, stones, etc. (see SLIME MOLD).

Subclass Ceratiomyxomycetidae. Fructification exosporous (spores borne externally), with spores on individual stalks, each spore giving rise to a protoplast that cleaves into 8 swarm cells.

Order Ceratiomyxales. Characters of the subclass. One genus (*Ceratiomyxa*) with 3 species, 1 of which (*C. fruticulosa*) is ubiquitous, the other 2 tropical.

Subclass Myxogastromycetidae. Spores borne internally in characteristic sessile (nonstalked) or stipitate (stalked) often brilliantly coloured sporophores ranging in size from 50 microns to 6 centimetres or more, characteristically containing a capillitium (tangled mass of threadlike structures) interspersed with the spores. Spores upon germination giving rise to 1 to 4 creeping myxamoebae or flagellated swarm cells.

Order Liceales. Microscopic (*Licea*) to large (*Reticularia*) sporophores devoid of capillitium but which may contain pseudocapillitium; lime absent; typically on dead wood or on bark of dead or living trees, rarely on dead leaves; spores in mass pallid, bright-coloured, or dingy olivaceous, hyaline (transparent, but still characteristically coloured) or subhyaline by transmitted light. About 70 species.

Order Echinosteliales. Microscopic, stalked; lime never present; capillitium, when present, scanty, arising from the tip of the columella; (central internal stalk of sporophore); spores white, pale pink, pale yellow, or brown in mass, typically with conspicuous, evenly distributed thickened areas on the walls; assimilative stage a protoplasmodium. Ten species.

Order Trichiales. Stalked or sessile with persistent peridium but without columella; lime rarely present and then scanty and on the surface only; capillitium abundant, consisting of a network or of separate elaters (filaments of the capillitium); spores bright-coloured in mass. About 150 species. *Arcyria, Hemitrichia,* and *Trichia* are the largest genera.

Order Stemonitales. Usually stalked, sometimes sessile; lime rarely present and then confined to the hypothallus (a filmlike residue that remains after the formation of sporangia), stalk, and columella; spores in mass black or deep violaceous or ferruginous (rust-coloured); assimilative stage an aphanoplasmodium (*i.e.*, a nongranular, transparent, thin plasmodium showing little or no differentiation into endoplasm and ectoplasm). About 90 species. *Stemonitis, Comatricha,* and *Lamproderma* are the most common genera.

Order Physarales. Stalked or sessile; lime characteristically present on the peridium and the capillitium; spores in mass black, deep violaceous, or ferruginous; assimilative stage a phaneroplasmodium (*i.e.*, a granular, thick, opaque plasmodium whose veins have clearly differentiated endoplasm and ectoplasm). About 190 species. *Physarum* is the largest genus with *Physarella, Fuligo, Diderma,* and *Didymium.*

Subdivision Eumycotina (true fungi)

Assimilative stage walled, typically filamentous (a mycelium), sometimes unicellular, usually eucarpic (having only part of the thallus forming a fruiting structure); asexual reproduction by fission, budding, fragmentation, or, more typically, by spores; sexual reproduction by various means, usually resulting in the formation of resting structures or meiospores.

Class Chytridiomycetes

Unicellular or filamentous, holocarpic (having all of the thallus involved in the formation of the fruiting body) or eucarpic; motile cells (zoospores or planogametes) with a single, posterior, whiplash flagellum. Mostly aquatic fungi saprobic or parasitic on algae, fungi, or less often, on flowering plants.

Order Chytridiales. Mycelium lacking but rhizoids (short absorbing filaments) or rhizomycelium often present. Chiefly freshwater saprobes or parasites of algae and fungi. Some terrestrial species, such as *Olpidium brassicae* and *Synchytrium endobioticum* cause plant disease. About 550 species.

Order Blastocladiales. Water molds with a restricted thallus, characterized by the production of thick-walled, pitted, resistant sporangia; sexual reproduction by isogamous (equal in size and alike in form) or anisogamous (unequal in size but still similar in form) planogametes; *Allomyces* exhibits an alternation of two equal generations. Most are saprobes but various species of *Coelomomyces* are parasitic in mosquito larvae; uniquely, their hyphae are devoid of cell walls. More than 50 species.

Order Monoblepharidales. Water molds with an extensive, foamy mycelium; sexual reproduction by a motile male gamete (antherozoid) fertilizing a nonmotile differentiated egg, resulting in a thick-walled oospore. About 20 species.

Class Hyphochytridiomycetes

A small group of mostly marine fungi very similar to the order Chytridiales but with motile cells bearing a single tinsel flagellum (*i.e.*, a flagellum with short side branches along the central axis, comblike).

Order Hyphochytriales. Characters of the class. About 15 species.

Class Plasmodiophoromycetes

Endoparasites (internal parasites) of fungi or plants often causing hypertrophy (excessive abnormal growth). Assimilative stage an endophytic (living within plant tissue) plasmodium that becomes converted into a group of zoosporangia (structures producing motile asexual spores) or a large number of small, walled spores; motile cells with 2 unequal, anterior, whiplash flagella.

Order Plasmodiophorales. Characters of the class. *Plasmodiophora* and *Spongospora* cause serious plant diseases. About 35 species (see also SLIME MOLD).

Class Oomycetes

Aquatic, amphibious, or terrestrial fungi; saprobic, facultatively (occasionally) or obligately (invariably) parasitic on plants, a few on fish. Asexual reproduction typically by zoospores with 2 anterior or lateral flagella, 1 whiplash, 1 tinsel; sexual reproduction usually by contact of differentiated gametangia (gamete-or sex-cell-producing structures) with nuclei from the male fertilizing differentiated eggs and resulting in thick-walled oospores; thallus probably diploid with meiosis occurring in the gametangia.

Order Lagenidiales. Holocarpic, unicellular or filamentous water molds, parasitic on algae and fungi or saprobic; oogonium (egg-producing structure) typically containing a single egg. About 85 species.

Order Saprolegniales. Mostly eucarpic, filamentous water molds or soil fungi; saprobic or parasitic; hyphae without constrictions or cellulin plugs; oogonia containing 1 to many free eggs. Some species are diplanetic, *i.e.*, they produce 2 types of zoospores, primary (pear-shaped with anterior flagel-

la) and secondary (kidney-shaped with lateral flagella). Some (*Aphanomyces*) cause root rots; others (*Saprolegnia*) infect fish and fish eggs. About 200 species.

Order Leptomitales. Aquatic saprobes found often in polluted waters; eucarpic; hyphae constricted, with cellulin plugs, arising from a well-defined basal cell; oogonium typically containing a single egg, which may be free or embedded in periplasm (a peripheral layer of protoplasm). Twenty species.

Order Peronosporales. Aquatic or terrestrial; parasitic on algae or vascular plants, the latter mostly obligate parasites causing downy mildews; zoosporangia, in advanced species, borne on well-differentiated sporangiophores, deciduous and behaving as conidia (asexually produced spores). About 250 species.

Class Zygomycetes

Terrestrial saprobes or parasites of plants, animals, or humans; asexual reproduction by aplanospores (nonmotile spores) in sporangia, or by conidia; sexual reproduction by fusion of morphologically similar gametangia, sometimes differing in size, resulting in thick-walled zygospores.

Order Mucorales. Saprobic, weakly parasitic on plants, or parasitic on humans and then causing mucormycosis (a pulmonary infection); asexual reproduction by sporangiospores, one-spored sporangiola (a small deciduous sporangium), or conidia. In the genus *Pilobolus* the heavily cutinized sporangium is forcibly discharged. These fungi are often called the bread molds. About 360 species.

Order Entomophthorales. Insect parasites or saprobes, some implicated in animal or human diseases; asexual reproduction by modified sporangia functioning as conidia, forcibly bread molds. About 360 species.

Order Zoopagales. Parasitic on amoebae, rotifers, nematodes, or other small animals, which they trap by various specialized mechanisms; asexual reproduction by conidia borne singly or in chains, not forcibly discharged. About 60 species.

Class Trichomycetes

Commensals (organisms living parasitically on another organism, but conferring some benefit in return, or at least not harming the host) with a filamentous thallus attached by a holdfast or basal cell to the digestive tract or external cuticle of living arthropods; asexual reproduction by sporangiospores (a spore borne within a sporangium), trichospores (zoospores or ciliated spores), arthrospores, (a spore resulting from fragmentation of a hypha), or amoeboid cells; sexual reproduction, where known, zygomycetous.

Order Amoebidiales. Thallus coenocytic (without cross walls, with numerous freely distributed nuclei) arising from a holdfast; amoeboid cells formed. About 12 species.

Order Eccrinales. Thallus coenocytic, attached by a holdfast to the digestive tract of arthropods; aplanosporangia produced in succession. More than 50 species.

Order Asellariales. Thallus branched, septate, attached by a basal coenocytic cell; asexual reproduction by arthrospores. Six species.

Order Harpellales. Thallus simple or branched, septate. Asexual reproduction by trichospores; sexual reproduction zygomycetous. About 35 species.

Class Ascomycetes

Saprobic or parasitic on plants, animals, or humans; some are unicellular but most are filamentous, the hyphae septate with 1, rarely more, perforations in the septa; cells uninucleate or multinucleate; asexual reproduction by fission, budding, fragmentation, or, more typically, by conidia usually produced on special sporiferous (spore-producing) hyphae, the conidiophores, which are borne loosely on somatic (main-body) hyphae or variously assembled in asexual fruiting bodies; sexual reproduction by various means resulting in the production of meiosphores (ascospores) formed by free-cell formation in saclike structures (asci), which are produced naked or, more typically, are assembled in characteristic open or closed fruiting bodies (ascocarps). Among the largest and most commonly known ascomycetes are the morels, cup fungi, saddle fungi, and truffles.

Subclass Hemiascomycetidae. Asci naked, formed from single cells or on hyphae; no ascocarps or ascogenous hyphae produced; saprobic or parasitic.

Order Protomycetales. Spore-sac compound (a synascus). A poorly known small group of plant parasitic ascomycetes. Twenty or more species.

Order Endomycetales. Mostly saprobic, a few parasitic; zygote or single cell transformed directly into the ascus; mycelium sometimes lacking. This group includes the yeasts and their relatives.

Order Taphrinales. Parasites on vascular plants. Asci produced from binucleate ascogenous (ascus-producing) cells formed from the hyphae in the manner of chlamydospores (thick-walled spores). Ninety or more species.

Subclass Euascomycetidae. Asci unitunicate (ascus wall single-layered), borne in various types of ascocarps; saprobic or parasitic on plants, animals, or humans.

Order Eurotiales. Asci globose to broadly oval, typically borne at different levels in cleistothecia (completely closed ascocarp or fruiting structure). Most of the human and animal dermatophytes belong here. Also many saprobic soil or coprophilous fungi. Possibly up to 150 species.

Order Microascales. Asci evanescent (quickly deteriorating), borne at different levels in perithecia (closed ascocarps with a pore in the top) with ostioles (the opening of the perithecium), or sometimes a long necklike structure terminating in a pore. Some serious plant parasites such as *Ceratocystis ulmi* (Dutch elm disease) and *C. fagacearum* (oak wilt) belong here. About 100 species.

Order Onygenales. Asci formed in a mazaedium (a fruiting body consisting of a powdery mass of free spores interspersed with sterile threads, enclosed in a peridium or wall structure), evanescent, and liberating the ascospores as a powdery mass among sterile threads. About 25 species.

Order Erysiphales. Obligate parasites on flowering plants causing powdery mildews; mycelium white, superficial in most, feeding by means of haustoria sunken into the epidermal cells of the host; 1 to several asci in a cleistothecium, if more than 1, in a basal layer at maturity; asci globose to broadly oval; cleistothecia with appendages. About 150 species.

Order Meliolales. Mycelium dark, superficial on leaves and stems of vascular plants, typically bearing appendages (termed hyphopodia or setae); asci in basal layers in ostiolate perithecia without appendages; mostly tropical fungi. More than 1,000 species.

Order Chaetomiales. Asci in basal layers in superficial perithecia that bear conspicuous, straight or curly, simple or branched, hairs on the surface; asci evanescent. About 110 species.

Order Xylariales. Perithecia with dark, membranous or carbonous (appearing as black burned wood) walls, with or without a stroma (a compact structure on or in which fructifications are formed); asci persistent, borne in a basal layer among paraphyses (elongate structures resembling asci but sterile), which may ultimately gelatinize and disappear. A rather large group of fungi one of which, *Neurospora crassa*, has been used extensively in genetic and biochemical studies. Approximately 4,500 species.

Order Diaporthales. Perithecia immersed in plant tissue or in a stroma with their long ostioles protruding; ascal stalks gelatinizing, freeing the asci from their basal attachment; paraphyses lacking. The chestnut blight fungus (*Endothia parasitica*) belongs here. Close to 500 species.

Order Hypocreales. Perithecia and stromata when present, brightly coloured, soft, fleshy, or waxy, when fresh; asci borne in a basal layer among apical paraphyses. About 800 species.

Order Clavicipitales. Perithecia immersed in a stroma that issues from a sclerotium (a hard-resting body resistant to unfavourable environmental conditions); asci with a thick apex penetrated by a central canal through which the septate, threadlike ascospores are ejected. The ergot fungus (*Claviceps purpurea*), cause of ergotism in plants, animals and humans, and the original source of LSD, belongs to this order. Some other Clavicipitales parasitize insect larvae. About 170 species.

Order Coryneliales. Asci in ascostromata with funnel-shaped ostioles at maturity. About 20 species.

Order Coronophorales. Asci in ascostromata with irregular or round, never funnel-shaped openings. About 30 species.

Order Laboulbeniales. Ascomycetes of uncertain affinity; minute parasites of insects and arachnids with mycelium represented only by haustoria and stalks. About 1,635 species.

Order Ostropales. Ascocarp a locule-like apothecium (an open, often cuplike, ascocarp); asci inoperculate (without a terminal pore) constructed as in the Clavicipitales; ascospores septate, threadlike. About 80 species.

Order Phacidiales. Ascocarp an apothecium immersed in a black stroma, the upper covering of which splits in stellate (star-shaped) or irregular fashion when ascospores mature. About 150 species.

Order Helotiales. Ascocarp an apothecium bearing inoperculate asci exposed from an early stage. Some important plant diseases are caused by members of this group. *Monilinia fructicola* causes brown rot of stone fruits. The earth tongues (*Geoglossaceae*) also belong here. More than 1,500 species.

Order Pezizales. Ascocarp an apothecium bearing operculate (with a hinged cap) asci above the ground; apothecia often large, cup- or saucer-shaped, spongy, brainlike, saddle-shaped, etc. This group includes the morels, the false morels, and the saddle fungi among others. About 700 species.

Order Tuberales (truffles). The ascocarps, mostly closed and borne below the ground, are considered to be modified apothecia. The asci are globose, broadly oval, or club-shaped. About 230 species.

Subclass Loculoascomycetidae. Asci bitunicate, borne in ascostromata; saprobic or parasitic on plants.

Order Myriangiales. Asci borne singly in locules arranged at various levels in a more or less globose stroma. About 100 species.

Order Dothideales. Asci borne in fascicles (clusters) in a locule devoid of sterile elements. About 600 species.

Order Pleosporales. Asci borne in a basal layer among pseudoparaphyses. More than 2,000 species.

Order Microthyriales. Stroma flattened, hemispherical, opening by a pore or tear; base usually lacking; asci borne among pseudoparaphyses. Mostly tropical fungi. About 1,200 species.

Order Hysteriales. Stroma boat-shaped, opening by a longitudinal slit, which renders it apothecium-like; asci borne among pseudoparaphyses. About 110 species.

Class Basidiomycetes

Saprobic or parasitic on plants or insects; filamentous; the hyphae septate, the septa typically inflated (dolipore) and centrally perforated; mycelium of 2 types, primary of uninucleate cells, succeeded by secondary, consisting of dikaryotic cells, this often bearing bridgelike clamp connections over the septa; asexual reproduction by fragmentation, oidia (thin-walled, free, hyphal cells behaving as spores), or conidia; sexual reproduction by fusion of hyphae (somatogamy), fusion of an oidium with a hypha (oidization), or fusion of a spermatium (a nonmotile male structure that empties its contents into a receptive female structure during plasmogamy —a kind of gamete) with a specialized receptive hypha (spermatization), resulting in dikaryotic hyphae that eventually give rise to basidia, either singly on the hyphae or in variously shaped basidiocarps; meiospores (basidiospores) borne on basidia. In the rusts (Uredinales) and smuts (Ustilaginales), the dikaryotic hyphae produce teleutospores (thick-walled resting spores), which are a part of the basidial apparatus. This is a large class of fungi containing the rusts, smuts, jelly fungi, club, coral and shelf fungi, mushrooms, puffballs, stinkhorns, and bird's-nest fungi.

Subclass Heterobasidiomycetidae. Basidia septate or deeply divided, or arising from a teleutospore or cyst; basidiospores often germinating by repetition, budding, or production of conidia. The subclass includes the jelly fungi, the rusts, and the smuts.

Order Tremellales (jelly fungi). Fruiting bodies (basidiocarps) well formed, appearing as inconspicuous horny crusts when dry but usually bright-coloured to black gelatinous masses after a rain. A few are parasitic on mosses, vascular plants, or insects; most are saprobes. About 500 species.

Order Uredinales (rusts). Parasitic on vascular plants. The basidial apparatus consists of a thick-walled teleutospore (probasidium) which either gives rise to a 4-celled tube (metabasidium) on which basidiospores are borne, or which itself becomes 4-celled and produces basidiospores directly; basidiospores forcibly discharged. Many rusts are heteroecious; *i.e.*, they require 2 species of host to complete their life cycle. Rusts are among the fungi most destructive to agriculture. About 4,600 species.

Order Ustilaginales (smuts). Called smuts because the masses of spores (sori) are usually black and dusty. Basidial apparatus consisting of a thick-walled teleutospore (probasidium), which, upon germination, gives rise to a septate or nonseptate tube (metabasidium), which bears the basidiospores. Basidiospores not forcibly discharged, germinating usually by budding, or fusing and then producing a mycelial germ tube. The various cereal smuts are of great economic importance. About 700 species.

Subclass Homobasidiomycetidae. This subclass includes the great majority of the Basidiomycetes. Most produce conspicuous, large-fruiting bodies, which bear the spores on basidia. The basidia are simple, cylindrical, or club-shaped. The basidiospores, which may or may not be forcibly discharged, germinate directly into a mycelium.

Order Exobasidiales. Basidiocarps lacking; basidia produced in a layer on the surface of parasitized vascular plants. Fifteen species.

Order Polyporales. Basidiocarps present. A large and probably heterogeneous order of fungi in which the basidia are borne in various ways but rarely on gills. It includes the coral fungi, the club fungi, the chanterelles, and the pore (shelf or bracket) fungi among others. Common genera include *Stereum, Clavaria, Hydnum, Cantharellus, Polyporus, Fomes. Schizophyllum* has been used extensively for genetic research. Perhaps 2,500 species.

Order Agaricales (mushrooms and boletes). Basidia produced in layers (hymenia) on the underside of fleshy fruiting bodies (basidiocarps), in tubes (boletes) or on gills (mushrooms). Some of these fungi form mycorrhizae, some are parasitic and cause root rots. The majority are saprobic. 4,000 to 5,000 species.

Order Hymenogastrales. Basidiocarps underground or on the surface but usually buried in humus, remaining closed, the interior (gleba) disintegrating into a slimy mass containing the spores. About 225 species.

Order Lycoperdales (puffballs). Gleba dry and powdery at maturity; consisting of small, pale spores and well-developed capillitium. About 160 species.

Order Sclerodermatales. These are puffballs with a hard peridium enclosing a dry, powdery gleba consisting of large, dark spores and some capillitium. About 120 species.

Order Phallales (stinkhorns). Gleba slimy and fetid at maturity; exposed on an elongated or net-shaped receptacle. *Phallus, Mutinus, Dictyophora, Simblum, Clathrus* are temperate-zone genera. About 70 species.

Order Nidulariales (bird's-nest fungi). The gleba separates into chambers, which become thick-walled, waxy, and hard. These are the peridioles ("eggs"), which are evident within a cup-like or goblet-like basidiocarp, the whole resembling a bird's nest at maturity. *Cyathus* and *Crucibulum* are the 2 most widely distributed genera. About 60 species.

Form-class Deuteromycetes

Fungi with septate mycelium reproducing only asexually and resembling asexual stages of Ascomycetes and Basidiomycetes.

Form-order Sphaeropsidales. Conidia borne in pycnidia. About 5,500 species.

Form-order Melanconiales. Conidia borne in acervuli. About 1,000 species.

Form-order Moniliales. Conidia borne on variously assembled conidiophores but never in pycnidia or acervuli. 10,000 or more species.

Form-order Mycelia Sterilia. No conidia produced. Probably mycelial stages of Basidiomycetes. About 200 species.

Critical appraisal. Although the fungi are traditionally classified in the plant kingdom, some of their characteristics argue strongly against such affinities. Chief among them are their heterotrophy (*i.e.*, the obtaining of food from an outside source rather than producing it as do green plants), and the almost universal presence of chitin, which is also found in the skin of many invertebrate animals such as insects, in their walls. Although, with the exception of the slime molds (Myxomycetes, or Mycetozoa), the fungi have never actually been included in the animal kingdom, some taxonomists have proposed their removal from the plant kingdom and their inclusion either in the kingdom Protista or in a separate kingdom —Fungi (see CLASSIFICATION, BIOLOGICAL).

Within the division Mycota, older classification systems applied the name Phycomycetes (algal fungi) to the six classes of "lower fungi" found in present-day systems (Chytridiomycetes, Hyphochytridiomycetes, Plasmodiophoromycetes, Oomycetes, Zygomycetes, and Trichomycetes). Although no longer recognized as a formal taxonomic category, the name Phycomycetes is still used informally by many mycologists as a useful term for designating those fungi that produce their spores in sporangia and have, in most cases, coenocytic hyphae—as is true of the presently recognized six classes.

The groups that are designated as the form-class Deuteromycetes (also sometimes called "Fungi Imperfecti") consist of organisms whose sexual stages either have not been found, or do not exist, together with the asexual stages of some sexually reproducing species. It has happened that sexual stages have been found for some species in this group and whenever this occurs, the species is immediately given a name appropriate to its proper taxonomic group. Sometimes, however, it has been found advisable to maintain the form-category name even though sexual stages are known and this has led to the unusual condition of one species having been assigned two scientific names, only one of which, of course, is valid—the one applying to the sexual stage. The practice is justified mainly for the convenience of having a classification system, artificial though it is, into which conidial (asexual) forms may be grouped and studied. It has even been found useful to extend the concept of form

categories to include the conidial stages of known sexually reproducing fungi, mainly Ascomycetes. Since many of these fungi, particularly the parasitic ones, are usually encountered only in the conidial stage, they can be easily identified by making use of the form-category classification. The convenience is great and the practice so widespread that in 1950 the International Botanical Congress legalized the use of form-names for conidial stages, recognizing, of course, the name of the perfect (sexual) stage as the official name of the whole organism.

BIBLIOGRAPHY

Popular, historical, and ethnomycological works: E.C. LARGE, *The Advance of the Fungi* (1940, reprinted 1962), a classic book on plant disease fungi and their impact on history; JOHN RAMSBOTTOM, *Mushrooms and Toadstools* (1953), a fascinating, beautifully illustrated book on the occurrence and activities of fungi; LUCY KAVALER, *Mushrooms, Molds and Miracles* (1965), a popular discussion of various discoveries of fungal products and their effects on man; V.P. and R.G. WASSON, *Mushrooms, Russia, and History*, 2 vol. (1957), an ethnomycological classic, now a collector's item; C.M. CHRISTENSEN, *The Molds and Man*, 3rd ed. (1965), a very interestingly written discussion of the relationships of fungi to other organisms and their value to man.

General works: G.C. AINSWORTH, *A Dictionary of the Fungi*, 6th ed. (1971), the standard reference for terminology and definitions—includes all taxa above species level with approximate number of species in each; W.H. SNELL and E.A. DICK, *A Glossary of Mycology*, rev. ed. (1971), an excellent dictionary of mycological terms; ELIZABETH MOORE-LANDECKER, *Fundamentals of the Fungi* (1972), a discussion of the biology of fungi at a very elemental level; C.J. ALEXOPOULOS, *Introductory Mycology*, 2nd ed. (1962), a textbook of mycology written in simple style for the student who knows nothing about fungi; L.E. HAWKER, *Fungi* (1966), a short discussion of fungi written at the introductory level; JOHN WEBSTER, *Introduction to Fungi* (1970), an introductory mycology textbook; J.H. BURNETT, *Fundamentals of Mycology* (1968), a discussion of fungal biology somewhat difficult for the novice; F.A. and F.T. WOLF, *The Fungi*, 2 vol. (1947), vol. 1 deals with morphology and taxonomy, vol. 2 with physiology, genetics, ecology, medical, and industrial mycology; E.A. BESSEY, *Morphology and Taxonomy of Fungi* (1950, reprinted 1964), a fine reference strong on phylogeny and bibliography of classification; E.A. GAUMANN, *Die Pilze*, 2nd ed. (1966), an excellent treatment, in German, of comparative morphology of the fungi with a lengthy bibliography and with special emphasis on phylogeny; W.D. GRAY, *The Relation of Fungi to Human Affairs* (1959), a discussion of useful and destructive fungi with emphasis on the application of mycology to industry; G.C. AINSWORTH and A.S. SUSSMAN (eds.), *The Fungi*, 3 vol. (1965–1968; vol. 4 in prep.), an advanced treatise written by specialists in various fields that discusses fungi in terms of cells, the organism, populations, and classification; A.H.R. BULLER, *Researches on Fungi*, 7 vol. (1909–50), a classic collection of studies on various aspects of fungi, particularly strong on spore dispersal, development, and sexual reproduction of fungi; G.W. MARTIN, *Outline of the Fungi* (1950), an outline of fungus classification.

Special subjects: V.W. COCHRANE, *Physiology of Fungi* (1958), a very comprehensive treatment of the subject; L.E. HAWKER, *The Physiology of Reproduction in Fungi* (1957), an authoritative treatment of the subject; C.T. INGOLD, *Dispersal in Fungi* (1953) and *Spore Liberation* (1965), two thorough discussions of the methods by means of which fungi liberate their spores; KARL ESSER and RUDOLF KUENEN, *Genetik der Pilze* (1965; Eng. trans., *Genetics of Fungi*, 1967), a general treatment of the subject; J.R. RAPER, *Genetics of Sexuality in Higher Fungi* (1966), a concise discussion of life cycles and sexual mechanisms in Ascomycetes and Basidiomycetes; D. PARKINSON and J.S. WAID (eds.), *The Ecology of Soil Fungi* (1960), a series of essays on soil fungus ecology; T.W. JOHNSON and F.K. SPARROW, *Fungi in Oceans and Estuaries* (1961), the most important reference on this group of fungi; W.B. COOKE, *Our Mouldy Earth* (1969), a study of the fungi of our environment with emphasis on water; C.W. EMMONS, C.H. BINFORD, and J.P. UTZ, *Medical Mycology*, 2nd ed. (1970), an up-to-date reference on fungi pathogenic to humans; J.W. WILSON and O.A. PLUNKETT, *The Fungous Diseases of Man* (1965), a medical treatise.

Special groups of fungi: W.D. GRAY and C.J. ALEXOPOULOS, *Biology of the Myxomycetes* (1968), a comprehensive summary of knowledge about the plasmodial slime molds, not including classification; G.W. MARTIN and C.J. ALEXOPOULOS, *The Myxomycetes* (1969), a monograph of the world species of Myxomycetes with technical descriptions of all known species, notes on distribution, and commentary on each species; H.M. FITZPATRICK, *Lower Fungi Phycomycetes* (1930, reprinted 1964), a taxonomic treatise of the Phycomycetes with keys to important taxa; F.K. SPARROW, *Aquatic Phycomycetes*, 2nd rev. ed. (1960), a monographic treatment of worldwide scope; J.S. KARLING, *The Plasmodiophorales*, 2nd ed. (1968), a monographic study including life cycles, cytology, morphology, and taxonomy; C.L. DUDDINGTON, *The Friendly Fungi* (1957), a fascinating book about fungal predators of small animals; J.N. COUCH, *The Genus Septobasidium* (1938), a monographic study of the genus, including a detailed account of the fungus–insect relationship; F.J. SEAVER, *The North America Cup-Fungi*, vol. 1, *Operculates* (1942), vol. 2, *Inoperculates* (1951; both vol. reprinted 1961), a beautifully illustrated taxonomic treatment; R.W.G. DENNIS, *British Ascomycetes* (1968), probably the best all-around taxonomic treatment of Ascomycetes but limited to species occurring in Britain; G.B. CUMMINS, *Illustrated Genera of Rust Fungi* (1959), a very useful book for identification of the genera of plant rusts; J.C. ARTHUR, *Manual of the Rusts in the United States and Canada* (1934, reprinted 1962), a monographic study of the plant rusts of North America; W.B. GROVE, *The British Rust Fungi (Uredinales)* (1913), a taxonomic treatment of the British plant rusts; G.C. AINSWORTH and KATHLEEN SAMPSON, *The British Smut Fungi (Ustilaginales)* (1950), a taxonomic treatment of the Ustilaginales of Britain; G.W. FISCHER, *Manual of the North American Smut Fungi* (1953), an indispensable work for the identification of the Ustilaginales; L.O. OVERHOLTS, *The Polyporaceae of the United States, Alaska, and Canada* (1953), the most comprehensive taxonomic treatment of wood-rotting and bracket fungi of North America; E.J.H. CORNER, *A Monograph of Clavaria and Allied Genera* (1950, reprinted 1966), a comprehensive treatment, definitely not for the beginner; ROLF SINGER, *The Agaricales in Modern Taxonomy*, 2nd rev. ed. (1962), a technical treatment of the gill fungi; C.M. CHRISTENSEN, *Common Fleshy Fungi*, 2nd ed. (1955), a small, spiral-bound, easy-to-use manual for the determination of many common mushrooms; J.W. GROVES, *Edible and Poisonous Mushrooms of Canada* (1962), identification of mushrooms based on field characters, illustrated with kodachromes; L.R. HESLER, *Mushrooms of the Great Smokies* (1960), illustrated with excellent black and white photographs; RENE POMERLEAU and H.A.C. JACKSON, *Champignons de l'Est du Canada et des États-Unis* (1951; Eng. trans., *Mushrooms of Eastern Canada and the United States*, 1951), a good mushroom manual; A.H. SMITH, *Mushrooms in Their Natural Habitats*, 2 vol. (1949), superbly illustrated with stereochromes, particularly useful for the Pacific Northwest region; *The Mushroom Hunter's Field Guide*, rev. ed. (1963), one of the easiest manuals to use, well illustrated in black and white and colour; W.C. COKER and J.N. COUCH, *The Gasteromycetes of the Eastern United States and Canada* (1928), a taxonomic treatment of the puffballs, stinkhorns, earthstars, and related fungi; H.L. BARNETT, *Illustrated Genera of Imperfect Fungi*, 2nd ed. (1960), a relatively simple manual for the identification of the genera of the fungi imperfecti; G.L. BARRON, *The Genera of Hyphomycetes from Soil* (1968), a well-illustrated manual for the identification of the Moniliales; K.B. RAPER and CHARLES THOM, *A Manual of the Penicillia* (1949, reprinted 1968), indispensable for the determination of species of *Penicillium;* K.B. RAPER and D.I. FENNELL, *The Genus Aspergillus* (1965), indispensable for the identification of species of *Aspergillus.*

(C.J.Al.)

Myriapod

Myriapod is a general term applied to any member of several closely related groups of the invertebrate phylum Arthropoda—the extinct Archipolypoda; the Diplopoda, or millipedes; the Chilopoda, or centipedes; the Pauropoda; and the Symphyla. The myriapods are a little-known group, although some 11,000 living species have been recognized.

General features. *Importance.* Most myriapods are secretive and seldom seen. Some attract attention by spectacular mass migrations; others are found occasionally in dark corners of houses and other buildings. Certain tropical species inflict painful bites on man. The primary significance of myriapods, however, is the role they play in the ecological balance of forested regions. In addition, their limited ability to migrate, their dependence upon stable conditions of moisture and shelter, and their general intolerance of seawater, together with the fact that they appeared rather early in geological history and, since then, have evolved little, make the myriapods im-

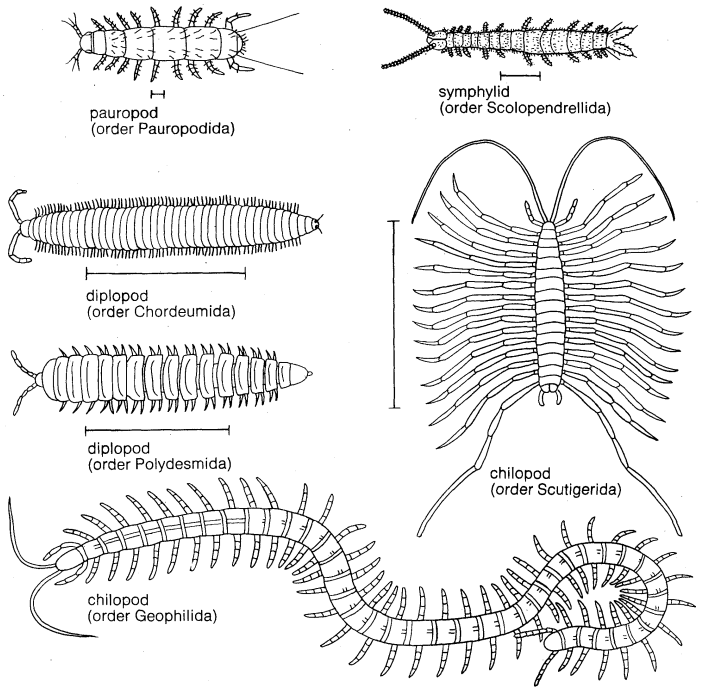
portant indicators of land–water relationships; these organisms thus provide useful information for understanding evolution and geographic dispersal.

Size range, distribution, and abundance. For the most part, myriapods are relatively small. Pauropods do not exceed two millimetres (about 1/12 inch), symphylids are rarely more than four millimetres. The smallest chilopods (centipedes) are about 5 mm; the largest may reach about 27 centimetres (about 10½ inches), but the average is about 30 mm (about 1⅕ inches). Diplopods (millipedes) range in length from about 3 mm to about 28 cm; the average size for the group, however, is less than 50 mm (about two inches).

The four living myriapod classes are distributed widely throughout the tropical and temperate regions; in some areas myriapods are conspicuous in the fauna of the soil, including its organic portion (humus). Because they are largely phytosaprophagous (subsisting on dead plants), they play an important role in the breakdown of dead vegetable material. Some species are primarily carnivorous, however.

Greater numbers of individuals of tiny myriapod species occur than do those of large ones. Myriapods are most abundant and diverse in forests, although some species of diplopods and an even greater number of chilopods thrive in grassland or in semi-arid habitats; still others live in desert conditions.

From *Invertebrate Identification Manual* by Richard A. Pimentel
© 1967 by Litton Educational Publishing, Inc.; reprinted
by permission of Van Nostrand Reinhold Company



Representative myriapods. Line scales indicate approximate size of each myriapod.

Natural history. *Reproduction and life cycle.* Internal fertilization occurs in all myriapods; however, the method of sperm transfer from male to female is indirect and highly variable. Paired gonads occur in both sexes, and maturation, with production of viable gametes (reproductive cells), begins at the time of the final molt (shedding of skin). In diplopods, symphylids, and pauropods, ducts from the gonads open near the anterior (front) end of the body, usually just behind the base of the second pair of legs; in chilopods the reproductive systems of both sexes open at the posterior (hind) end of the body.

Sperm transfer

Only in the diplopods does sexual contact occur. In diplopods except those of the order Polyxenida, the male produces a spermatophore, or sperm-bearing capsule, either during or prior to clasping the female. Subsequently he transfers the capsule to specialized receptacles at the openings of the oviducts (tubes that carry the eggs) by mouth or by means of modified appendages (gonopods) on the seventh body segment. There is little courtship activity, the male usually finding the female by chemical sensors and then seizing her.

In diplopods of the order Polyxenida and in other myriapod groups, sperm transfer is much less direct and usually entails deposition of a spermatophore by the male and its subsequent recovery by the female. In Pauropoda and Polyxenida, the male constructs a small web of silken threads, in which he deposits several drops of spermatic fluid; the female subsequently picks up the sperm near her oviduct openings. Sperm transfer in the symphylids is more complex. The female recovers spermatophores with her mouthparts, and the sperm remain in special receptacles inside the mouth until deposition of the eggs. Fertilization is accomplished by the female's transferring some of the stored sperm from the oral cavity to the nearby opening of the oviduct on the second body segment, by means of a chewing-like activity.

Chilopods practice a form of courtship that involves a mutual caressing of the antennae, followed by the construction of an irregular web and deposition of a spermatophore by the male. In some species the female picks up the sperm mass almost immediately; in others she may leave the area after courtship and not return to the spermatophore for several hours. A male of the order Scutigerida has been observed transferring the spermatophore from its web to the female gonopore with his mouth. Following transfer of the spermatophore, fertilization occurs after the sperm have moved along the oviduct.

The number of eggs produced varies among different groups. Usually diplopods of the orders Spirobolida and Spirostreptida produce a few relatively large eggs; those of the order Polydesmida produce as many as 2,000 relatively small ones.

Care of the eggs varies among diplopods. Spirobolids coat each egg with a layer of undigested material from the mouth, then pass it backward by the legs and place it into the rectum, where most of the moisture is removed. The completed egg capsule closely resembles a fecal pellet of the species. Members of the orders Polydesmida and Julida construct an earthen "igloo"; the female deposits her eggs through the top before sealing it. The young, after hatching inside the capsule, apparently derive some nourishment from the capsular material as they gnaw their way out. Species of the order Glomerida also build earthen capsules, enclosing only one or two eggs in each. Diplopods of the order Chordeumida enclose their eggs in a cocoon, the threads of which are produced by spinnerets, special organs at the posterior end of the body.

Chilopod females of certain members of the orders Scutigerida and Lithobiida deposit the eggs singly, roll them in dirt for camouflage, then place them in a concealed place. Chilopods of the orders Geophilida and Scolopendrida, accumulate the eggs into a cluster, which is brooded and protected by the female from desiccation, fungal infestation, and predation. Development is largely completed prior to hatching, and the newly emerged animal is similar to the adult except for its smaller size, lack of pigmentation, and sexual immaturity. The young chilopods sometimes remain with the mother for several days after hatching.

Molting and growth

In diplopods, pauropods, symphylids, and two orders of chilopods (Lithobiida and Scutigerida), segmentation of the embryo is not completed before the eggs hatch; the immature animal thus has fewer segments than the adult and must pass through anamorphosis—a sequence of growth stages during which segments and appendages are added. The young hatchlings usually begin life with three pairs of legs at the anterior end of the body. Subsequent growth is characterized by periods of active feeding alternating with quiescent phases during which the animal sheds its exoskeleton (hard external covering) and forms a new one; forms several new legless segments at the end of the body; and develops appendages for the legless segments of the preceding molt.

The number of immature stages (stadia) required to reach maturity varies among the different groups. Matur-

ation of the reproductive system usually occurs at the final molt, after which no further growth occurs. In some diplopods and chilopods, however, molting and growth may continue after the attainment of sexual maturity. During the molting period the animals are soft and defenseless, and many diplopods construct a round molting chamber consisting of mud and excrement in which to pass this time safely.

Some myriapods attain maturity within a few months after hatching; others may require a year or more, during which time the body increases in size and length. If the species typically has a short period of immaturity, adulthood also is brief. Slower growing forms, on the other hand, may live several years after completing their development.

Behaviour. Myriapods, generally secretive animals, pass much of their life hidden in the soil or in leaf mold; thus they have not developed the complex patterns of behaviour that occur in many other arthropod groups—*e.g.*, insects, spiders, and crabs. Most diplopods, symphylids, and pauropods are primarily phytosaprophagous; only rarely do they eat healthy living plants. Some diplopods accept animal tissue as food, and a few species appear to be preferentially carnivorous. All chilopods, however, are apparently predators that feed upon small organisms.

Most myriapod species are negatively phototropic; *i.e.*, they avoid light and are active at night or in conditions of shade or darkness; sudden exposure to sunlight results in either active evasive responses or, particularly in diplopods, a coiling of the body.

Defensive behaviour

Myriapods respond to human handling by vigorously writhing and twisting their bodies, by releasing a defensive secretion, and sometimes by discharging their intestinal contents. Most chilopods attempt to bite when picked up by man, particularly the large scolopendrids. Some chilopods have strong, thickened rear legs capable of inflicting a sharp pinch.

The predatory chilopods are basically solitary animals; the herbivorous pauropods and diplopods, however, often gather in large groups in a particularly favourable habitat. Under exceptionally favourable conditions, populations of diplopods may increase until a mass migration takes place. Sometimes the purpose of the migration may be to escape unfavourable conditions. "Armies" of migrating diplopods occasionally invade homes or contaminate water supplies in rural areas by falling into wells and dying there. Locomotives have been stopped by the slippery crushed bodies of great numbers of diplopods covering the rails.

Locomotion. The methods of locomotion among myriapods vary, as would be expected in a group with such diversity of body form. Some myriapods are long and cylindrical with rigid body segments, others relatively short and flexible; some are adapted to surface running, others to burrowing in the soil. The overall characteristics of each myriapod group presumably reflect adaptations of locomotory mechanisms appropriate to particular environmental niches.

A basically elongated, cylindrical body form with short legs can, in general, be associated with the need for a powerful forward (pushing or burrowing) motion. The mechanism characteristic of the rigid-bodied diplopods, however, is quite different from the burrowing mechanism of similarly elongated, short-legged chilopods of the order Geophilida. The latter have a highly flexible body capable of little direct pushing motion, but they have the ability to wedge through small crevices.

The nerves that lead to the leg muscles come from the ganglia, or nerve centres, of the ventral nerve cord; in a slowly moving diplopod, leg movement begins near the head and passes backward in a series of wavelike motions, suggesting the passage of a primary impulse along the entire ventral nerve.

Myriapods with relatively short bodies often have long legs and the ability to run rapidly. Variations in the internal musculature and the mechanism of leg attachment to the body reflect functional requirements for different gaits among the various groups.

Ecology. Small diplopods and chilopods, as well as all pauropods and symphylids, usually live in conditions of sustained coolness and moisture. Some species may be active during the day, crawling on the forest floor or climbing trees, but such activity is usually confined to the larger species capable of resisting dehydration. Only a few diplopods (order Spirostreptida) and chilopods (order Scolopendrida) have been able to adapt to arid or semidesert conditions; even these forms do not tolerate prolonged exposure to full sunlight.

Myriapods, however, are found in a wide variety of habitats, ranging from holes and crevices in the soil (including the burrows of earthworms, insects, and small mammals) to caves, leaf litter, loose bark, rotting logs, and even on plants some distance above the ground.

Many species adapt readily to human intrusion into their environment. Under such conditions they may thrive and sometimes are accidentally transported great distances (*e.g.*, on agricultural products). The common house chilopod (*Scutigera coleoptrata*) occurs widely in the Northern Hemisphere in places such as basements and has become adapted to outdoor conditions in areas far beyond its original range. Many tropical diplopods now are found in plantations and gardens throughout the warmer parts of the world, to the extent that their native distribution can no longer be ascertained.

Adaptation to environment

Although most diplopods are unable to endure being exposed to seawater, many chilopods live in the intertidal zone along seacoasts, where they hide beneath stones or among algae and prey upon worms and other small organisms.

Diplopods, pauropods, and symphylids are primary consumers in the ecological food chain, deriving their nourishment chiefly from dead and decomposing organic material. Chilopods, which are mainly predators, are secondary consumers in the food chain and thus usually occur in much smaller numbers.

Myriapods are parasitized internally by various protozoans, nematodes (a minute type of worm), and certain fungi. A few species of flies act as external parasites; mites live on the legs and in depressions of the body wall. Predators include certain insects of the order Heteroptera, the larval stages of some beetles (Coleoptera), scorpions, spiders, chilopods, salamanders and toads, various birds, and small mammals such as shrews, moles, and mice.

Form and function. Myriapods are characterized by the presence of mandibles, one pair of antennae, tracheae (respiratory tubules), and little or no tendency for modification of the body into thoracic and abdominal regions. These features, except for the last, are shared with insects, and it is the opinion of some authorities that the transition between symphylids and true winged insects has been so gradual that the insects should be considered myriapods. The five myriapod classes are sharply differentiated. Except on the basis of individual characters of questionable significance, it is not possible to arrange the four living classes into groups.

Members of the class Diplopoda (millipedes), perhaps the most specialized myriapods, are characterized by the fusion of the original body somites, or segments, into pairs, forming diplosomites, or paired somites, each equipped with two pairs of legs (except where secondarily lost) and two ganglia on the ventral nerve cord. In most millipedes the legs of the original anterior somite have migrated posteriorly, permitting each diplosomite to be telescoped into the preceding one. Diplosomites occur only in millipedes, and this character is the most significant in their taxonomic ranking as a class. At the anterior end of the body several segments may be provided either with a single pair of legs or none. This reduction is commonly associated with the ability of most species to roll the head region backward in a protective posture.

Specializations of millipedes

The mouthparts of the millipede also are distinctive, consisting of one pair of mandibles—usually massive—and a flattened platelike structure, the gnathochilarium, which forms the floor of the front part of the mouth. This structure, once thought to correspond to the two original pairs of maxillae (external mouthparts), is now regarded

as being formed by the fusion of the two appendages of the fourth head somite of the embryo. The form and arrangement of the small sclerites, or sections, of the gnathochilarium differ from group to group and are important in classification; in some millipede orders this structure is reduced to a small fragile plate with no suggestion of individual sclerites.

The body wall of most millipedes is calcified. The presence of calcium carbonate in addition to the basic chitin material (chitin is a complex carbohydrate) results in a rigid, brittle skeletal structure found in no other group. In what is presumed to be the primitive arrangement, each diplosomite consists of a dorsal surface (the tergum), two lateral surfaces (the pleura), and a ventral surface (the sternum) connected by hardened tissue. Various modifications of this arrangement occur—in some cases, the tergal and pleural sclerites are fused, and in others all four elements are fused into a ring.

Most diplopods have a nearly circular body with short legs. In a few groups, the posterior part of each segment has broad lateral extensions (paranota), which, when horizontally directed, give the effect of a broad, flat body. In most diplopod orders, some species have glands that secrete a volatile fluid capable of repelling predators and killing fungi. Comparable glands are unknown in other myriapod classes.

The class Pauropoda includes a number of extremely small myriapods, remarkable in lacking well-developed tracheal and circulatory systems. The thin flexible body wall contains setae (bristles) in regularly arranged patterns. The body consists of 12 segments, including a pygidium, or terminal segment, and there are nine pair of legs. The dorsal, or back, side is modified to form from six to ten tergal plates. The head contains a pair of mandibles and a structure formed from a pair of fused maxillae, perhaps analogous to the millipede gnathochilarium. The most characteristic pauropod structures, however, are the paired antennae, composed of a four-jointed stalk with two branches, one long and one short.

Symphylid structure

Symphylid species are small, fragile, and lacking pigmentation. The head bears a pair of simple, multisegmented moniliform (beardlike) antennae, a pair of mandibles, and two pairs of maxillae similar to those of insects. As in the insects, the second pair of maxillae are fused into a so-called labium. The long slender body is composed of 12 leg-bearing segments provided with 15 or 22 tergites and a legless terminal segment bearing a pair of large cerci—sensory structures perhaps homologous with walking legs. On the legs, close to the juncture with the body, is a small projection called a stylus and a small saclike structure—characters presumably homologous with similar structures on the abdominal appendages of certain wingless insects. No other myriapods have these structures. A tracheal system and a simple dorsal circulatory vessel are present.

The external structure of members of the class Chilopoda (centipedes) approximates the ancestral, or least specialized, myriapod. The centipedes are characterized by one pair of legs on each body segment, by modification of the appendages of the first trunk segment into a pair of forceps-like fangs, and by the presence of two pairs of maxillae behind the small mandibles. The antennae are short, simple, beadlike, unbranched, and composed of a number of short unmodified segments. The head is flattened and covered by a single plate. The centipede body is typically elongated and more or less flattened. The body wall is chitinous and flexible and lacks calcium carbonate. Each body segment corresponds to an original embryonic somite and is composed of at least a dorsal tergum and a ventral sternum, with a variable number of smaller pleural sclerites in different patterns according to the taxonomic group. The body shows no differentiation into thorax and abdomen. The hindmost legs are often elongated or otherwise modified, particularly in the male. In some groups, the segment of the leg nearest the body may have glands opening on the lower side. In other groups, the sternal, or lower, body plates may likewise be perforated with minute ducts of internal glands.

In most chilopods, the mouthparts, located ventrally and well back from the front of the head, are relatively small and often quite fragile. What often appear to be functional mouthparts are the modified legs of the first body segment, turned forward beneath the head (often extending in front of it) and provided with poison glands opening through the tip. All species of centipedes thus are poisonous, a character not found in any other myriapod class. The tracheal system opens through spiracles, or pores, usually located on the sides of the body. The spiracles may form a continuous series or be absent from specific body segments.

Poison gland in centipedes

Evolution and paleontology. Information about the evolution of myriapods is deduced largely from the characters of living species. There is no fossil record for pauropods, and that for diplopods, symphylids, and chilopods is of little value. Apparently, however, all of these groups originated in the Devonian Period (about 345,000,000–395,000,000 years ago) during the great evolutionary radiation of terrestrial arthropods. The earliest diplopod fossils are from middle Pennsylvanian beds of the United States and central Europe. Some of these fossils represent living orders but have not been identified with certainty. The next oldest fossil diplopods are of the Oligocene Epoch (about 26,000,000–38,000,000 years ago), chiefly from northern Europe. All of the recognizable species from this period belong to modern families and genera. Chilopods and symphylids are not known with certainty earlier than the Oligocene. The fossils from that period represent modern genera.

There is general agreement that myriapods and insects share a number of common features, not only of morphology but also of embryological patterns and reproductive biology. It is generally believed that these groups originated in the Early Devonian or Late Silurian Period (about 395,000,000–400,000,000 years ago) from a common ancestral type probably bearing a close resemblance to modern symphylids.

Classification. *Distinguishing taxonomic features.* Myriapods may be identified on the basis of subtle, often concealed, characteristics—mouthparts, genitalia, pore and seta distribution. Perhaps less than 20 percent of the myriapod species have been discovered and classified. New families continue to appear, and the status of several orders remains uncertain. The classes comprising the myriapods are very different, although most live in temperate or tropical regions, have separate sexes, are mostly relatively small in size, and require moisture and shelter to survive.

Annotated classification. The classification of myriapods is not settled. Many categories are involved; the scheme below represents a brief summary of one of several possible schemes. The groups indicated with a † are extinct and known only from fossils.

MYRIAPOD

Myriapod is a general term used to describe members of 5 arthropod classes, consisting of about 11,000 known extant species; size range varies from 2 to 280 mm; important in maintaining ecological balance of tropical and temperate regions; most feed on dead plants, some (centipedes) are carnivorous.

† **Class Archipolypoda**

Apparently similar to diplopods, but with segmental elements less coalesced; head larger than body segments and provided with large compound eyes; possibly aquatic in habit; some species probably several feet in length; lived during Silurian and Pennsylvanian periods (about 280,000,000–430,000,000 years ago).

Order Euphoberiida. Characteristics same as those for class.

Class Diplopoda (millipedes)

Elongated myriapods; abdomen consisting of an indefinite number of double segments (diplosomites), most with 2 pair of legs; mouthparts consisting of mandibles and maxillae fused into a platelike structure, a gnathocilarium; trachea present; thorax with gonopores in the 3rd segment (progoneate); length 3 mm–28 cm; about 8,000 known species; Pennsylvanian to Recent.

Subclass Penicillata

Minute; 11–13 distinct body segments; both wall thin, with series and tufts of specialized bristles, without carbonate impregnation; sperm transfer to female indirect, by a spermato-

phore, or sperm-bearing, web constructed by male; about 80 species known; Oligocene to Recent.

Order Polyxenida. Three families, chiefly tropical; genus *Polyxenus* in North America and Eurasia.

Subclass Pentazonia

Small to large in size; body short and robust, segments consist of arched tergum, 2 pleura, 2 paramedian sterna separated by coxae; last pair of legs of male modified into clasping organ to hold female during sperm transfer; about 200 species; Pennsylvanian to Recent.

Order Glomeridesmida. One family in Indonesia, India, West Indies.

Order Glomerida. Two suborders: Sphaeriotheridea distributed throughout southern Africa, India, and Australasia (about 100 species); Glomeridideа distributed throughout the Northern Hemisphere and in Southeast Asia (about 150 species).

†*Order Amnilyspedida.* Flourished during the Pennsylvanian Period (about 280,000,000–325,000,000 years ago) in North America and Europe.

Subclass Helminthomorpha

"Typical" diplopods, usually elongated and slender; body segments usually cylindrical; sternum never divided; most sclerites coalesced; 7th pair of legs modified into gonopods for sperm transfer; about 7,600 species; Pennsylvanian to Recent.

Order Spirobolida. Chiefly tropical; 12 families; includes some of the most common and conspicuous diplopods.

Order Stemmiulida. One family with species in Central America, the West Indies, West Africa, South India, Ceylon, and New Guinea.

Order Polyzoniida. Two families: Polyzoniidae, native to the Northern Hemisphere; Siphonotidae, native to the Southern Hemisphere; species small, poorly known.

Order Siphonophorida. One family in tropical America and Southeast Asia.

Order Julida. Mainly in Northern Hemisphere; several hundred known species in 4 families; many species common in gardens.

Order Spirostreptida. Eight families; chiefly tropical; includes the largest known diplopod species, up to 28 cm (nearly 1 ft) long; other species less than ½ in.

Order Callipodida. Small group (3–4 families) chiefly in the Mediterranean region and in North America; agile, largely predatory species found in dry, rocky places and in caves.

Order Platydesmida. Two families; in the Northern Hemisphere and Southeast Asia; often colonial in habit; males of some species brood the eggs.

Order Chordeumida (Craspedosomatida). Many species in 20–30 families; occur mainly in the Northern Hemisphere; produce a silken thread to construct egg cocoons and molting chambers.

Order Polydesmida. A large, worldwide group of 40–50 families; all species eyeless; most with 20 body segments; body length from 3 mm to 13 cm (about ⅛–5 in.); often brightly coloured.

Class Pauropoda

Very small, length up to 1.9 mm; progoneate; 2 pairs of appendages transformed into mouthparts; usually 9 pairs of walking legs; branched antennae; circulatory and tracheal systems much reduced or lacking; about 180 living species known; worldwide distribution; no fossil record.

Order Pauropodida (Tetramerocerata). Three families; worldwide distribution.

Order Millotauropodida (Hexamerocerata). One family only in Madagascar.

Class Symphyla

Small, soft-bodied; 3 pairs of mouthparts, 12 leg-bearing segments; terminal segment with a pair of cerci (sensory structures); elongated moniliform antennae; gonopore usually on 4th trunk segment; length to 8 mm; about 120 living species known; in tropical regions; Oligocene to Recent.

Order Scolopendrellida. Characteristics same as class; 3 families.

Class Chilopoda (centipedes)

Elongated with many distinct abdominal segments, each with one pair of legs; head with moniliform antennae, 3 pairs of mouthparts; limbs of first abdominal segment modified as poison claws; gonopore on last segment (opisthogoneate); length about 5 mm–26.5 cm; about 2,800 living species known worldwide; Cretaceous to Recent.

Order Scolopendrida (Scolopendromorpha). Largest known chilopods, up to 26.5 cm (about 10 in.) but others less than an inch in length; large forms capable of inflicting severe bites; 2 families.

Order Geophilida (Geophilomorpha). Slender, extremely elongated; mostly burrowers, but some occur in caves and intertidal zones; nearly 200 pairs of legs in some species; 8 families.

Order Lithobiida (Lithobiomorpha). Short-bodied, fast-running species; body with 15 segments; found in North and South Temperate zones; about 6 families.

Order Scutigerida (Scutigeromorpha). Short-bodied species with extremely long legs; mouthparts leg-like in appearance; stigmata opening middorsally; found chiefly in tropical regions; one family.

Critical appraisal. Previous classificiation of myriapods have been to a large extent, based on the writing of C. Attems, one of the great authorities on these animals; his classification was summarized in 1926 (see below *Bibliography*). Many of Attems' concepts originated from those of earlier workers; he was conservative both in defining groups and in naming them. The arrangement of diplopod groups has undergone many changes and is still being modified. A revised list of higher categories was published by Ralph Chamberlin and Richard Hoffman in 1958 and another by the latter in 1969.

Because several myriapod features are found also in insects, and because the transition between symphylids and true winged insects is so gradual, some authorities would include the insects with the other myriapod groups in a subphylum Myriapoda. Other authorities (see bibliography under Tiegs and Manton) defend the traditional viewpoint that insects should be maintained as a category with rank equivalent to the myriapod classes; the latter is the arrangement followed in the classification scheme above.

The features of the various myriapod classes are sharply differentiated. An earlier arrangement of grouping the Diplopoda, Pauropoda, and Symphyla into a category called Progoneata and of grouping the Chilopoda and Insecta into an equivalent category called Opisthogoneata has been discredited as having been based upon a character—position of the gonopore of the male—of only secondary importance.

Formerly, the millipedes and centipedes were grouped together as the class Myriapoda, but their features differ to such an extent that each is now given class status.

BIBLIOGRAPHY. CARL GRAF VON ATTEMS, "Myriopoda," in W. KUKENTHAL and T. KRUMBACH (eds.), *Handbuch der Zoologie*, vol. 4 (1926), the best general, illustrated reference (in German) to myriapods, although the classification is out of date; R.V. CHAMBERLIN and R.L. HOFFMAN, *Checklist of the Millipeds of North America* (1958), the only recent complete list of diplopods, now outdated; ALFRED KAESTNER, *Invertebrate Zoology*, vol. 2, trans. and adapted by H.W. and L.R. LEVI (1968), contains a good chapter on the myriapods; A.G. SHAROV, *Basic Arthropodan Stock* (1966), the view of an eminent Russian paleontologist, highly controversial in many areas, but with interesting insights and valuable information; R.E. SNODGRASS, *A Textbook of Arthropod Anatomy* (1952), an excellent introductory account of the morphology of all arthropod groups, beautifully illustrated; O.W. TIEGS and S.M. MANTON, "The Evolution of the Arthropoda," *Biol. Rev.*, 33:255–337 (1958), a detailed and comprehensive review of the subject, including an historical survey of earlier works.

(R.L.H.)

Myrtales

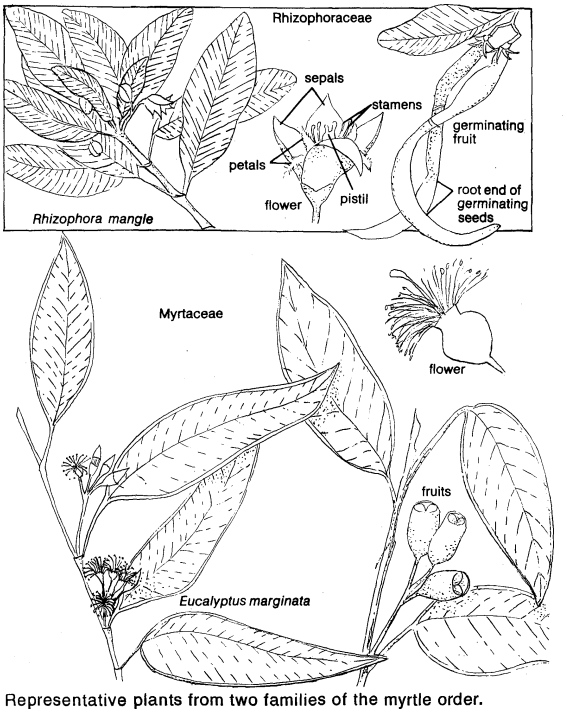
The Myrtle order (Myrtales) consists of over 400 genera and nearly 10,000 species of flowering plants, and its representatives grow in a variety of habitats throughout the world. Members of the family Melastomataceae are particularly abundant in Brazil, where they form a characteristic component of the vegetation. In many parts of Australia, Myrtaceae is the most prominent plant family, with the genus *Eucalyptus* forming large forests. These two families are the largest in the order, and together they account for approximately three-quarters of the total genera and species.

GENERAL FEATURES

Predominance of the woody habit, either as shrubs or trees, is a feature of the order, and some members of the genus *Eucalyptus* are among the world's tallest trees. *E. regnans*, the mountain ash of Victoria and Tasmania, commonly attains a height of over 300 feet (90 metres), and an exceptional specimen was once measured at 370 feet (113 metres). Some herbaceous (nonwoody) species, however, occur in the families Lythraceae, Melastomataceae, and Onagraceae, while the family Trapaceae consists entirely of herbaceous annuals (plants entirely completing a life cycle in one growing season). In the Lythraceae family, there is an association between habit and climate; in temperate regions, the species are annual or perennial herbs, whereas in warmer areas they are often woody shrubs and, sometimes, trees. In the family Penaeaceae, which is confined to rather dry areas of southwestern and southern Africa, the plants are small shrubs of heathlike habit adapted to dry-wasteland habitats. Diversity in size and habit is exhibited by members of the family Melastomataceae, which includes, as well as shrubs and trees, root climbers, epiphytes (plants that are not rooted in soil but live among the branches of trees), and herbaceous marsh and water plants. A similar range of variation is seen in the family Onagraceae which is found in the subtropics and temperate zones.

Diversity of growth forms

Drawing by R. Findahl



Representative plants from two families of the myrtle order.

Distribution and abundance. Members of this order are to be found chiefly in the tropics and subtropics, but most families have some representatives in temperate regions. Apart from the families Oliniaceae and Penaeaceae, which are restricted to the southern half of Africa, all families are widely distributed either as native plants or, as in the families Punicaceae and Myrtaceae, as introduced species through extensive cultivation. The widest natural distribution is that of the Onagraceae family, which has its chief development in the temperate zones of the Americas, particularly in the western regions. Its largest genus, *Epilobium*, is composed of some 160 species and is cosmopolitan except in the tropics. The family Lecythidaceae is found throughout the tropics, but the family Combretaceae, although mainly tropical, extends into the subtropics. The Lythraceae family has representatives in a variety of climatic zones, but it is most abundant in tropical America. The number of its species declines rapidly toward the temperate regions, and it is absent from the cold zones. *Lythrum salicaria* is one of the few plant species found in both Old and New Worlds, and its range covers Europe, through Asia to Japan, as well as North America and southeastern Australia. Melastomataceae, the largest family, is also mainly tropical, but its members extend into the subtropics and more temperate regions. About 2,000 of its species are native to the American continent. Two main centres of development of the family Myrtaceae are recognized: tropical America, for the berry-fruited members, and Australia, where the dry-fruited species predominate. The 600 species of *Eucalyptus* are distributed throughout Australia from the semi-arid regions to the coastal forests. The most widely spread genus is *Myrtus*, which grows in the warmer regions of both hemispheres; and the myrtle (*M. communis*) is either long naturalized in Europe or the only native *Myrtus* member in that continent. Mangrove species of the Sonneratiaceae and Rhizophoraceae families are found along tropical seacoasts, although the latter family is mainly Old World. *Rhizophora mangle* is the only native mangrove species in America, and it is restricted to peninsular Florida. The related family Anisophylleaceae is exclusively tropical, and the family Trapaceae is found from central and southern Europe to eastern Asia and from tropical to subtropical Africa.

Importance. Economically, the most important single genus is *Eucalyptus*, many of whose species are a valuable source of hardwoods. Because of their rapid growth rate, they are now grown widely throughout the temperate regions of the world as shade trees or in forestry plantations. In addition, oils with fever-controlling and germicidal properties, gums, and resins are extracted from their leaves and wood, while some species are of importance as a source of nectar and pollen for hive bees. Other members of the Myrtaceae family used for timber are *Metrosideros* and *Angophora*, and the pungent, medicinal cajeput oil is extracted from the paperbark tree, *Melaleuca leucadendron*. Two members of the Lythraceae family used as a source of good timber are the tulipwood tree of Brazil, *Physocalymma scaberrima*, and *Lagerstroemia flos-reginae* of eastern India, while *Terminalia* of the family Combretaceae and *Lecythis* and *Couroupita* of the Lecythidaceae family are also valuable timber trees. For underwater constructions such as piers and jetties, whole trunks of the mangrove *Rhizophora* are widely used because their tannin content inhibits attacks from marine wood-boring animals. Although the order is not of great importance as a source of food, edible fruits are produced by some members of most families. The most widely known is probably the pomegranate (*Punica granatum*), which is now cultivated in the warmer regions of the world for its fruit and also as an ornamental shrub. In the Myrtaceae family the guava (*Psidium guajava*), strawberry guava (*P. cattleianum*), and many species of *Eugenia* are of local value, such as the Brazil cherry (*E. michelii*), the Surinam cherry (*E. uniflora*), the Otaheite apple (*E. malaccensis*) and the rose apple (*E. jambos*). The dried flower buds of *Syzygium aromaticum* are the spice cloves of commerce, and the powdered unripe fruit of *Pimenta dioica* is allspice. The water chestnut (*Trapa natans*) is eaten in China, and the fruits of *T. bispinosa* are the source of the singhara flour of northwestern India. Other edible fruits are found in the genus *Maietia* of the Melastomataceae family and in some species of *Fuchsia*, while in tropical Africa a butter-like substance, chiquito, is obtained from the fruits of *Combretum butyrosum*. The family Lecythidaceae is notable for the Brazil nut (*Bertholletia excelsa*) and sapucaia nuts (*Lecythis*), for the oil extracted from the seeds of *Barringtonia*, and for the anchovy pear (*Grias cauliflora*), cultivated in the West Indies. The family Combretaceae supplies the Indian, or country, almond (*Terminalia catappa*) and the myrobalan nuts (*T. chebula*), but the only edible products of the family Anisophylleaceae are the poga nuts (*Poga oleosa*). Other products from members of this order are a red cosmetic dye from *Lawsonia inermis* and *Woodfordia floribunda* of the family Lythraceae, fibre for textiles from the soft bark of *Couratari* (family Lecythidaceae), colouring matter from some of the members of Melasto-

Edible fruits and other useful products

mataceae family, and tannins for tanning hides from the bark and foliage of *Rhizophora* and *Terminalia*. Many species of the Lythraceae family are horticulturally important as shade trees or as garden plants, such as crape myrtle (*Lagerstroemia*), mignonette or henna tree (*Lawsonia*), loosestrife (*Lythrum*), and cigar flower (*Cuphea*). Horticulturally important plants in the Myrtaceae family are *Eucalyptus*, myrtle (*Myrtus communis*), geraldton waxflower (*Chamaelaucium*), bottlebrush (*Callistemon*), and many others, including beautiful horticultural hybrids of *Leptospermum*. The family Melastomataceae provides the garden shrubs *Medinilla*, *Heterocentron*, and *Tibouchina* and the small vine *Schizocentron*, which is grown in hanging baskets. The herbaceous annuals or perennials of the Onagraceae family, such as those found in the genera *Oenothera*, *Clarkia*, and *Godetia*, as well as the shrubby *Fuchsia*, are familiar cultivated plants throughout the world.

NATURAL HISTORY

Life cycle. The flowers of most species have both male and female parts in the same flower (perfect, or bisexual, flowers), but those of the Sonneratiaceae, Anisophylleaceae, and Combretaceae families may have flowers of only one sex; structures of the opposite sex, though originally present, fail to mature in these cases. At the completion of meiosis (nuclear reduction division leading to gamete, or sex-cell, formation) in the anthers, the microspores remain joined together for a short time in tetrads, or clusters of four, before they separate as pollen grains. In the genus *Epilobium*, however, they remain together as permanent tetrads and are liberated as such when the anthers split open. In most members of the Onagraceae family, the pollen grains are connected by viscid threads that stick to a visiting insect; thus, hundreds of pollen grains are drawn out of the anther together. When mature, the pollen is usually two celled; *i.e.*, it consists of a large vegetative or pollen tube cell and a smaller generative cell. In the Lecythidaceae and Melastomataceae families, however, the generative cell has already divided to form the two male gametes, or sperm cells, when the anthers ripen and split. The pollen grains in these families are therefore three celled, while in the Combretaceae family both two- and three-celled conditions are shown by different species.

Family differences

Within the ovary are one or more future seeds, or ovules, each of which is provided with two seed coats, or integuments. In the centre of each ovule is the embryo sac, or female gametophyte, one of whose cells, the egg, develops into the embryo after it has been fertilized by a male gamete from a pollen grain.

In *Eucalyptus*, only relatively few of the ovules present in each ovary develop into seeds, and the majority are sterile because they lack embryo sacs. These are the ovulodes, which later shrivel and form the chaff that is shaken out of the mature capsule, together with the few viable seeds.

Pollination. Although normally the flowers open when mature, the opposite condition, of mature flowers remaining closed (cleistogamy), occurs in *Ammania* and *Rotala* of the family Lythraceae and in *Jussiaea acuminata*. The last mentioned plants grow near the Sahara in locales where the ground temperatures of 70°–80° C (158°–176° F) are too hot for insects, and there the mature flowers remain closed. Usually when mature flowers remain closed it implies self-pollination, but, in *Purpurella cleistoflora* of the family Melastomataceae, robber bumblebees effect cross-pollination when they force an entrance to the flowers in search of pollen. The normal condition, of mature flowers unfolding, however, does not necessarily mean that cross-pollination takes place, and the British *Epilobium* species form a series from the small-flowered, largely self-pollinated species to those with large, insect-pollinated flowers. Pollination by wind occurs in some of the fuchsias and perhaps in *Eucalyptus*, but usually the transfer of pollen is accomplished by insects, chiefly bees or moths and butterflies. In certain species of the Lecythidaceae and Melastomataceae families, however, hummingbirds are the pollinators, and honey eaters are the pollinators in the larger flowered members of the Myrtaceae family. *Metrosideros excelsa* of New Zealand relies for its pollination on a forest bird, the tui, which migrates in early summer to the coast of the North Island in search of nectar. In Brazil, flowers of the genus *Feijoa* are visited regularly by small birds that eat the sweet, fleshy petals. The most complex cross-pollination mechanism in the order is that of *Lythrum salicaria*, which produces three forms of flowers on different plants. The stamens, pollen grains, styles, and stigmas likewise are of three different shapes, and only certain crosses are effective in seed production.

Postfertilization events. Following fertilization, nuclear endosperm, a special nourishing tissue for the developing plant embryo, is formed in the embryo sac. This tissue subsequently becomes cellular, except in the Melastomataceae family, where it is digested by the growing embryo while it is still in the free-nuclear condition. In all other families it is the cellular endosperm that is digested, and the mature seeds are nonendospermic. Extrusion of the endosperm through the broad micropyle opening of the ovule to form haustoria, special absorbing organs, is a feature of the family Rhizophoraceae, and these may play a supplementary role in the nutrition of the embryo. The genus *Trapa* is unique in the Myrtales order in that the primary endosperm nucleus degenerates, and so no endosperm is formed, but the lack of this source of nourishment for the embryo is compensated for by its formation of a haustorial suspensor, which invades the tissues of the ovule outside the embryo sac.

Early development of the seed

The embryo is dicotyledonous (*i.e.*, it has two cotyledons, or "seed leaves"), but in the Penaeaceae family the cotyledons are minute, and in the Anisophylleaceae family they are either very small or absent. In the Melastomataceae and Trapaceae families, one cotyledon is much smaller than the other and is sometimes reduced to a scale. Polyembryony, the presence of several embryos in each seed, has been reported in several families of the order and is very common in *Myrtus communis* and *Eugenia*. In *E. paniculata* up to 21 embryos have been found in a seed, and it is unusual for there to be fewer than two.

The fruit varies from dry to fleshy, these extremes often occurring within the same family. In the genus *Lecythis* (the monkey pot), the tissues giving rise to the stamens in the flower persist and enlarge to form a lid, which separates when the fruit is ripe, and in *Bertholletia excelsa* the hardened persistent sepals (green, leaflike structures next to the petals in a flower) form a plug that is later dislodged during seed germination, and the seedlings escape through the opening.

Vivipary is a feature of many of the mangroves; that is, the seed germinates while the fruit is still retained on the tree, and, following growth of the seed leaves, the root end of the embryo elongates and passes through the micropyle opening in the ovule and then through the ovary wall. In some instances it may reach three feet in length before the seedling drops into the mud, where it establishes itself at once. In *Couroupita guianensis* and *Grias cauliflora* (family Lecythidaceae) and *Myrciaria* (family Myrtaceae) the flowers and, later, the fruits are borne directly on the old wood of the trunk. This is known as cauliflory, and in *Couroupita* the large, round, woody fruits attached to the trunk are responsible for the name cannonball tree.

Unusual responses to the environment. Those species of *Eucalyptus* that grow on poor soil in low-rainfall areas show a prominent development of the lignotuber, a woody rounded knot or lump surrounding the base of the tree trunk and usually partly or entirely underground. Such trees have a multistemmed type of growth and remain small, both in height and in stem diameter. A similar environmental response may be exhibited in medium- to high-rainfall localities in individuals growing on very shallow soil on the tops of ridges, while members of the same species growing on deeper soil nearby show no lignotuber development and have normal growth. In the Onagraceae family, temporary adaptation to an oxygen-deficient environment is seen in the genus *Jussiaea*. When plants grow in water, air spaces develop in the tissues of

the roots and submerged stems, and in *J. repens* some special spongy-tissued respiratory roots grow to the surface of the water. The same plants under ordinary dryland conditions do not exhibit these features.

Ecology. Certain members of the order are adapted to special ecological situations. For example, *Trapa* and some species of the Melastomataceae and Onagraceae families are aquatic or marsh plants, while the mangroves require habitats found only along tropical muddy seashores or estuaries. Both epiphytes and root climbers of the Melastomataceae family are confined to tropical forests, but the genus *Penaea*, with its low, shrubby habit, is adapted to life in dry parts of southwestern and southern Africa. The fireweed *Epilobium angustifolium* is one of the first plants to appear after fire devastation or other disturbances, such as bulldozing, and it achieved fame by its rapid colonization of bomb sites in Great Britain during World War II.

FORM AND FUNCTION

No structural characters are common to all families of this order. Vegetatively, many of the plants are woody perennials, and most bear opposite leaves, but variation both between and within the families is considerable, and the concept of the order is based on flower structures. Of these, the most important is the form of the receptacle, which forms a floral tube of varying length and degree of attachment to the wall of the ovary (female seed-bearing structure).

Growth peculiarities are apparent in some members of the order. For example, the alternate adult leaves of *Eucalyptus* assume a vertical position by twisting of the petiole (leafstalk) during development, but this feature is not exhibited by those species that bear permanently juvenile foliage, in which the leaves are opposite instead of alternate. In the family Combretaceae, two types of leaf arrangement occur in one genus, *Quisqualis*. Alternate leaves are borne on the stem, but in the flowering shoots the leaves are opposite, and, in some climbing species, the petiole bases persist as hooks after the leaves have dropped. Probably the most unusual growth form is that exhibited by the mangroves of the Sonneratiaceae and Rhizophoraceae families, in which portions of some roots grow upward through the mud until they project some inches above low-tide level. These "breathing roots" bear small openings in their bark and act as aerating organs. Aerial prop roots are also a feature of most mangroves, and these persist, forming a tangled jungle, even after the main roots and stem bases of the trees have decayed. The genus *Terminalia* also possesses breathing roots, and loose internal cell tissue containing air spaces occurs in the submerged portions of the stems and roots of marsh-inhabiting species of the family Melastomataceae. *Trapa*, a floating annual, is supported by the inflated petioles of its floating leaves, while its submerged leaves are finely divided and resemble roots. An interesting association with ants is exhibited by some members of the Melastomataceae family, such as *Maietia* and *Tococa*. In these, saclike outgrowths are formed on the upper surfaces of the leaves and sometimes on the petioles, and they serve as shelters (*dormatia*) for ants, which enter them from the lower surface. The ants protect the flowers from all but the legitimate flower pollinators by swarming out of hiding at any unusual disturbance of the branch.

The mangrove habit of growth

EVOLUTION

The fossil record has not contributed to an understanding of the origin of the Myrtales order, but its families, as now known, were probably all defined by Tertiary times (*i.e.*, by about 65,000,000 years ago). Theories advanced as to their evolution rest largely on deductions from comparative structural studies of living species and are, therefore, subjective. Most such schemes are based on details of floral structure. Accordingly, the perigyny–epigyny (*i.e.*, position of the flower petals and other flower parts relative to the ovary, being either "around" or "above" it) shown by the typical flower of the Myrtales order is seen as the expression of an evolutionary tendency initiated in the Saxifragales order. Because of its perigynous–epigynous flowers, the Saxifragaceae family of that order has been favoured by many as an ancestral family. It is considered by some authorities, on the other hand, that vegetative features are a more reliable guide to ancestry than those associated with reproduction (*i.e.*, the flower). On this basis, they recognize in flowering plants two independent but parallel lines of descent: the fundamentally woody families and the fundamentally herbaceous families. According to this concept, the families Lythraceae, Onagraceae, and Trapaceae in the Myrtales order may then be grouped into an order, the Lythrales, that can be thought to have arisen through the order Caryophyllales from the most primitive herbaceous flowering plants, and the remaining families can be traced ultimately from the most primitive woody flowering plants of the order Magnoliales.

CLASSIFICATION

Distinguishing taxonomic features. The chief botanical feature of importance in classification in the Myrtales order is the nature of the floral tube (receptacle, or hypanthium), whose form ranges from the free tubular condition of varying lengths to that of its partial or complete fusion with the ovary wall. This is the most important taxonomic character, and many taxonomists consider that its variation in expression between—and sometimes within—the component families represents an evolutionary sequence that reaches its final expression in epigynous flowers (*i.e.*, "inferior-ovary" flowers, those with flower parts appearing to arise at the top of the ovary rather than at its base). The number of component families in the Myrtales order varies in different classifications according to the authority. The form of the flower, the shape of the floral tube or receptacle, and the degree of its fusion with the ovary wall are characters of primary importance, but the number, arrangement, and points of origin of the stamens, as well as details of the female flower parts, are also taken into account. Vegetative features such as the form and arrangement of the leaves, the presence or absence of stipules (small, leaflike appendages at the bases of leaves), and the occurrence of internal phloem (an additional layer of special conducting cells inside the stem, varying from the usual presence of only one layer just beneath the bark) are of secondary taxonomic value.

Criteria used to define members of the order

Annotated classification. Although there are some differences in the families included in the Myrtales order by different authorities, the following is a good recent treatment. Because the number of genera and species of each family also varies among authorities, the numbers given below should be considered to be no more than a general guide.

ORDER MYRTALES

Flowers usually regular, bisexual, and 4- or 5-merous (lobed or parted), with a floral tube and showing various degrees of perigyny leading to epigyny (fusion of the floral tube around and above the ovary). Stamens are usually free, but the ovary is syncarpous (of fused sections), and the style (the slender upper part of the ovary) is simple. Fruits are either dry or succulent, with nonendospermic seeds. The order consists of mostly woody shrubs or trees with usually opposite, simple, entire leaves. Internal phloem is present in the stems of all except the Rhizophoraceae and Lecythidaceae families, but this information is not available for the Punicaceae, Anisophylleaceae, or Oliniaceae families.

Family Lythraceae

Sepals 4–6, petals 4–6 or none, stamens in 2 series of either 4+4 or 6+6, carpels (ovary sections) 2–6 fused. Flowers perigynous, rarely bilaterally symmetrical, with a spurred floral tube in the genera *Cuphea* and *Pleurophora*, and borne in racemes, panicles, or cymes. Appendages sometimes present between the sepals; petals crumpled in the bud, commonly red or purple. Fruit a capsule. Leaves opposite or whorled, stipules minute or none. Chiefly tropical, but found in all zones except frigid. About 25 genera, 470 species.

Family Sonneratiaceae

Sepals 4–8, petals 4–8 or none, stamens numerous, carpels 4–15 fused. Flowers sometimes unisexual, being either male or female, perigynous, solitary, or in 3-flowered cymes or corymbs. Calyx lobes persistent; petals often crumpled in the bud. Fruit a capsule or berry. Leaves opposite or whorled, without

stipules. Tropical coasts from eastern Africa to Australia, Polynesia, and Indomalaya. About 3 genera, 12 species.

Family Punicaceae
Sepals 5–8, petals 5–8, stamens numerous, carpels 3–7, fused. Flowers epigynous, solitary or in cymes. Sepals fleshy, persistent; petals crumpled in the bud, scarlet. Fruit berrylike. Leaves usually opposite or whorled and without stipules. Balkans to Himalayas; Socotra, an island off Yemen. One genus, *Punica,* with 2 species, one of which is the cultivated pomegranate, *P. granatum.*

Family Rhizophoraceae
Sepals 4–8, petals 4–8, stamens numerous, carpels 2–5 fused. Flowers slightly perigynous to epigynous, solitary or in cymes. Sepals fleshy and persistent, petals lobed or with lacerate margins, convoluted in the bud. Fruit a berry, rarely a capsule or drupe. Leaves opposite, often leathery, with early-deciduous stipules and branches swollen at the nodes. Tropical, mostly Old World, also found in Australia, Polynesia, and Florida. About 16 genera, 120 species.

Family Anisophylleaceae
Sepals 4, petals 4, stamens 8, carpels 4 fused. Flowers often unisexual, epigynous, borne in spikes. Sepals with erect lobes; petal margins entire, lobed or lacerate. Fruit dry, rarely a drupe or winged capsule. Leaves alternate, with 3–5 main veins, without stipules, often in 2 rows; alternate leaves may be small or minute. Tropical America, Asia and Africa. About 4 genera, 36 species.

Family Combretaceae
Sepals 5, petals 5 or none, stamens in two series of 5+5, carpels 1. Flowers epigynous, usually small, occasionally zygomorphic (bilaterally symmetric) in spikes or racemes. Sepals persistent; petals absent in the genus *Terminalia.* Fruit dry or drupelike, occasionally winged. Leaves alternate or sometimes opposite, without stipules. Found throughout the tropics, rare in subtropics. About 18 genera, 600 species.

Family Lecythidaceae
Sepals 4–6, petals 4–6 or none, stamens numerous and joined, carpels 2–6 fused, the ovary with as many chambers as carpels. Flowers sometimes zygomorphic, perigynous to epigynous, solitary or in racemes. Sepals persistent; petals lacking in the genus *Napoleona,* in which a corona is formed of petallike sterile stamens. Fruit a berry or capsule, sometimes very large, woody, and not splitting open at maturity. Leaves alternate with marginal glands, often very large, without stipules. Tropical America, west tropical Africa, northern Australia, Indomalaya. About 20 genera, 150 species.

Family Myrtaceae
Sepals 4–5, often small or even absent, petals 4–5, fused in *Eucalyptus* to form a cap that is shed when the flower opens, stamens numerous and often conspicuous, carpels one or many but fused. Flowers perigynous to epigynous, in cymes or rarely racemes or spikes. Sepals often persistent; petals usually nearly circular. Leaves usually opposite, leathery, gland-dotted, without stipules. Those with fruit a berry, rarely a drupe, are placed in the subfamily Myrtoideae. Those with leaves often alternate and fruit a capsule comprise the subfamily Leptospermoideae. The remainder belong to the subfamily Chamaelaucoideae, which is defined by dry fruits that are usually 1-seeded. Tropics and subtropics. About 100 genera, 3000 species.

Family Melastomataceae
Sepals 4–5, petals 4–5, stamens 8–10, carpels 4–5, fused. Flowers perigynous to epigynous, sometimes slightly zygomorphic, in cymes. Sepals may be reduced to a rim or united into a cap. The anther connective often extended into a conspicuous appendage. Fruit a capsule, but a berry in epiphytes. Leaves usually opposite, and each pair is arranged at right angles to the next pair on the stem, with one member of each pair often small and falling with age, the 3–9 longitudinal leaf veins connected by side branches, leaves without stipules. Found throughout the tropics, mainly in South America, less common in subtropics. About 200 genera, 4,000 species.

Family Oliniaceae
Sepals 4–5, petals 4–5, stamens 4–5, carpels 3–5 fused. Flowers epigynous, in cymes. Calyx teeth minute and deciduous or absent; petals alternating with small coloured scales. Fruit a drupe. Leaves opposite, with comblike or feather-like "nerves," or veins, without stipules; branches 4-angled. Southern and eastern tropical Africa. 1 genus, 10 species.

Family Penaeaceae
Sepals 4, petals absent, stamens 4, carpels 4, fused. Flowers perigynous, solitary or in pairs. Sepals persistent. Fruit a capsule. Leaves small, opposite, adjacent pairs at right angles to each other on the stem, mostly without petioles, stipules minute, glandlike, or absent. About 5 genera, 25 species.

Family Onagraceae
Sepals 4, petals 4, stamens in two series of 4, carpels 4, fused. Flowers epigynous, occasionally zygomorphic, solitary or in spikes, racemes, or panicles. Sepals bent downward or backward in the open flower; petals occasionally absent. Fruit a capsule, sometimes a berry or a nut. Leaves alternate, opposite, or whorled, rarely with stipules. Subtropics and temperate zones. About 20 genera, 650 species.

Family Trapaceae
Sepals 4, petals 4, stamens 4, carpels 2, fused. Aquatic plants with perigynous, solitary flowers. Sepals persistent, becoming hard and spinelike. Fruit a 1-seeded drupe with fleshy, early-deciduous ovary wall. Floating leaves in a rosette arrangement, each roughly diamond-shaped with toothed margins, without stipules; submerged leaves opposite, finely divided. Central and southeastern Europe, temperate and tropical Asia and Africa. One genus, *Trapa,* with either 30 species or, according to some authorities, 1 polymorphic (*i.e.,* having many forms) species.

Critical appraisal. Earlier taxonomic treatments of this order recognize 23 families in four suborders. The more recent view presented here, however, defines the order to include only 12 of the former families, and it recognizes the family Anisophylleaceae as a 13th independent, though closely related, family near the family Rhizophoraceae.

BIBLIOGRAPHY. G.H.M. LAWRENCE, *Taxonomy of Vascular Plants* (1951), a reference book that covers most families of flowering plants as well as gymnosperms and ferns; A.B. RENDLE, *The Classification of Flowering Plants,* vol. 2, *Dicotyledons* (1956), a student text and useful reference book containing descriptions of families and related information such as pollination and anatomical characters; J. HUTCHINSON, *Families of Flowering Plants,* vol. 1, *Dicotyledons,* 2nd ed. (1959), a reference book containing descriptions of families and some keys to genera as well as to the families, with a scheme of classification based primarily on vegetative features; L.H. BAILEY, *Manual of Cultivated Plants,* rev. ed. (1960), a useful book containing keys to cultivated plants and a descriptive paragraph about each species; J.C. WILLIS, *A Dictionary of the Flowering Plants and Ferns,* 7th rev. ed. (1966), a reference work of value to students and botanists, giving descriptions of all families and some genera and listing the distribution and number of species of the remaining genera; G.L. DAVIS, *Systematic Embryology of the Angiosperms* (1966), a reference work summarizing the reproductive processes in each family; N. HALL, R.D. JOHNSTON, and G.M. CHIPPENDALE, *Forest Trees of Australia.* Australian Forestry and Timber Bureau, 3rd ed. (1968), a text containing good keys, descriptions, and photographs of 118 eucalypts and other myrtaceous trees; A.R. PENFOLD and J.L. WILLIS, *The Eucalypts: Botany, Cultivation, Chemistry and Utilization* (1961), a very useful discussion of the economic products of eucalypts containing a good summary of growth habits, floral morphology, and leaf types.

(G.L.D.)

Mysore

The state of Mysore, situated on the west coast of India, is the seventh largest in area and the eighth largest in population of the 21 constituent states of the Republic of India. With an area of 74,037 square miles (191,757 square kilometres) and a population of about 29,000,000, it coincides approximately with the area in which the Kanarese, or Kannada, language is spoken. It is bounded by the territory of Goa and the state of Mahārāshtra to the north, by the state of Andhra Pradesh to the east, the state of Tamil Nadu to the southeast, the state of Kerala to the south, and the Arabian Sea to the west. The state extends for about 420 miles from north to south and for about 300 miles from east to west. It has a coastline 200 miles long. The state capital is the city of Bangalore (population about 1,600,000), near the southeastern border.

While it forms a coherent linguistic unit, Mysore—which includes a coastal region, part of the hill ranges known as the Western Ghāts, and the Mysore Plateau in the interior—does not coincide with natural physical boundaries. Before the independence of India in 1947, Mysore was a prosperous and progressive but landlocked princely state, with an area of less than 30,000 square miles and a population of about 7,000,000 situated on the plateau. The transfer of additional territories to the state

in 1953 and 1956 united the Kanarese-speaking peoples, gave the state an outlet to the sea, and greatly extended its boundaries. Mysore has abundant hydroelectric power and extensive forests. It is also India's chief source of gold, produces nearly two-thirds of its coffee, and also provides the bulk of the world's supply of sandalwood.

Economic importance

History. The name Mysore (from the Hindu word for "buffalo town") derives from the destruction of Mahiṣāsura by the goddess Cāmuṇḍā. The prehistory of Mysore is lost in legends that concern the struggles that took place in southern India between invading Aryan peoples and the original inhabitants; in legendary form this is represented as a conflict between devils and demons on the one hand and gods and goddesses on the other. The subsequent history of Mysore deals mainly with the state as it was before its expansion in 1953 and 1956, for no dynasty succeeded in ruling the whole region occupied by the Kanarese-speaking peoples, which is called the Karnatak (Kanara), sometimes referred to by Westerners as the Carnatic.

After the reign of Aśoka, the Mauryan emperor of India, in the 3rd century BC, the principal dynasties in the area of Mysore were the Kadambas of Banavasi, the Western Gaṅgas (who were in power from the 3rd to the 11th centuries AD), the Bānas, and other feudatories of the Pallava dynasty, which ruled from about the 4th to the 9th centuries. The rich lands of the upper Tungabhadra region and the land between that river and the Krishna River were taken from the Kadambas in the 6th century by a central Karnatak family, the Cālukyas. Their efforts, and those of their competitors of the Rāṣṭrakūṭa dynasty, to unite the plateau and exploit the softer lands of the coastal plains enriched Mysore but led to reprisals from the Tamils to the east and south. By the 12th century the Hoysaḷa dynasty had absorbed Gaṅgavāḍi (as the state of Mysore was then called); but after the Hoysaḷas had been obliged to submit to the sultan of Delhi, Mysore gradually came under the sway of the state of Vijayanagar, whose capital of the same name stood on the site now partly occupied by the village of Hampi on the Tungabhadra River in contemporary Mysore. In the latter part of the 16th century the Vijayanagar Empire faded, giving place to Mughal power north of the Tungabhadra River and to the rajahs of Mysore in the south. In the 17th century the *wadiyars* ("rulers") of Mysore profited from the conflict between the Mughal Empire and the Marāthās in western India, as well as from the internal power struggles that occurred in the Mughal Empire after the death of Aurangzeb, to expand their rule. In 1610 the *wadiyar* of Mysore seized Srirangapatnam (Seringapatam); later, Bangalore was also acquired and *wadiyar* power consolidated. Later, maladministration at home and interference in wars of succession in the plains led to the usurpation of power in 1761 by the military adventurer Hyder Ali. His invasions of Malabar and the Karnatak plains extended Mysore's dominion but eventually led, after the resulting Mysore Wars, to the death of his colourful and energetic son Tippu Sultan in 1799 and to the conquest of Mysore by the British, who sponsored the restoration of *wadiyar* rule. From 1831 to 1881 Mysore was governed by a British commissioner, after which administration was once again restored to the *wadiyars*. The last of the *wadiyars* became governor of the state after the territorial reorganizations of 1953 and 1956.

The Hyder Ali episode

Physical and natural landscape. Physiographically, Mysore is divided into four distinct regions—the coastal plain, the hill ranges (the Western Ghāts), the Mysore Plateau to the east, and the black cotton soil tract to the northwest. The coastal plain represents a continuation of the Malabar Coast and experiences heavy rainfall from June to September, when the southwest monsoon winds blow. Coastal sand dunes give place inland to small alluvial plains with coconut-fringed lagoons. The coast is difficult of access, except by sea. To the east the land rises sharply to the slopes of the Western Ghāts, which have an average height of about 3,500 feet above sea level. Covered with forest, the Ghāts form a region known as the *malnad*; it is also a watershed, and from its crest numerous swift streams, including the Sharavati River, which is associated with the Gersoppa Falls, flow to the plains. Other streams flow over the undulating Mysore Plateau, which slopes gently eastward. The plateau region has an average height of about 1,500 feet above sea level; its soils are generally porous and infertile. The basins of its rivers, however, which include the Cauvery and the Tungabhadra, which is a tributary of the Krishna River to the north, are loamy and of some fertility. From early times the smaller streams have been dammed to form irrigation reservoirs; in recent times much hydroelectric power development has taken place. In the northwest part of the state, underlying volcanic rock produces a soil known as regur, the humus-rich black cotton-growing soil of India.

Passing from west to east, the *malnad* gives way on the plateau to more open plain country, known as *maidan*, where rainfall is less and where monsoon forests are replaced by scrub forests and scrubland. The monsoon forests are rich in wildlife, which includes tigers, elephants, bison, and deer. In the *maidan*, wild boar, bears, panthers, and leopards may be encountered. Among the common birds is the peacock.

Population. Mysore had a population of about 29,200,000 in 1971, at which time Bangalore, the capital, had a population of about 1,650,000, and the city of Mysore 356,000. Other important towns are Hubli-Dhārwār 380,000; Mangalore 214,000; Belgaum 214,000; Gulbarga 146,000; Bellary 125,000; Dāvangere 121,000; Bijāpur 103,000; Shimoga 103,000; and Bhadravāti 101,000.

Ethnologically, the population of Mysore is predominantly Dravidian. The Dravidians were the original inhabitants of India and are related to the peoples of the Malay Peninsula, Sumatra, and Madagascar. Over the centuries, however, much intermixture with other races has occurred. In the north of the state, Aryan traits, mixed with Dravidian, occur. Some Anglo-Indians, representing a mixture of European and Dravidian, are occasionally encountered, particularly in Bangalore.

The Dravidian population

Kannada is the language spoken by more than 65 percent of the population, although Hindustani (a mixture of Hindi and Urdu) is often used in trade and business. Toward the borders of the state other languages, such as Tamil, Telugu, Marathi, and Konkani, are also spoken. Konkani is associated particularly with the city of Mangalore.

The major religion is Hinduism, but Jainism and Buddhism—once widespread—still survive. About 10 percent of the population follows Islām, and 2 or 3 percent is Christian.

While about 76 percent of the population is rural and only about 24 percent urban, the trend toward urbanization is becoming more marked as industrialization increases.

Administration. The state is divided into 19 districts, which are grouped into the four administrative divisions of Bangalore, Mysore, Belgaum, and Gulbarga. The head of state is the governor, who is appointed by the president of India. The bicameral legislature consists of a legislative council of 63 indirectly elected members and a legislative assembly of 208 directly elected members. The chief minister is assisted by a council composed of ministers of cabinet rank. The state high court is subordinate to the Supreme Court in New Delhi; it consists of 13 judges, who are appointed by the president of India. There are also district and subordinate courts. A public service commission, whose members are appointed by the governor, functions in an advisory capacity.

Social conditions. Mysore, which at the time of the 1971 census had a population that was more than 31 percent literate, is one of the educationally advanced states of India. There are about 34,000 schools and educational institutions, including the University of Mysore at Mysore, the Karnatak University of Dhārwār, and Bangalore University and the University of Agricultural Sciences, both at Bangalore. Nearly half of the schools and institutions are managed by the government, and the remainder of local boards and private bodies. Compulsory free primary education is provided in about 4,000

towns and villages out of the 25,000 towns and villages in the state.

By the early 1970s there were about 1,000 hospitals and dispensaries and a number of health centres in both urban and rural areas. The Employees State Insurance Scheme covers sickness, maternity, and employment injury risks and provides free medical treatment to 48,000 factory workers and their families in Bangalore. Welfare schemes are run by the government for tribal and other groups, such as the Harijan group (named "Harijan," or "people of God," by Gandhi, in place of the controversial term untouchable), and the beggars. (Beggary is discouraged by legislation.) The State Social Welfare Board also provides educational, cultural, and welfare facilities for women and children through more than 400 rural centres.

The economy. The economic resources of Mysore are constituted by its forests, its oils, its rivers, and its mineral wealth.

The forests of the *malnad* region in the west produce teak, sandalwood, bamboo, and such other products as tanning dyes, gums, and lac (a resinous substance used in the manufacture of varnishes). Oil processed from sandalwood at Mysore city is a leading state export.

Agriculture is the occupation of 80 percent of the population. The coastal plain is intensively cultivated, with rice being grown as a food crop and sugarcane as a cash crop. In the Western Ghāts coffee and tea plantations are established. More than 90 percent of India's coffee is produced there, especially in the southern part; tea is grown on the higher hills. In the eastern region, where irrigation makes cultivation possible, sugarcane is grown and a little rubber, as well as such fruits as bananas and oranges. The black cotton soil of the northwest is devoted to the cultivation of cotton, oilseeds, and peanuts (groundnuts).

Tea and coffee plantations

Mysore has sufficient hydroelectric power to supply surplus energy to neighbouring states. The Sharavati project near Jog is the largest of several hydroelectric plants that provide power to Mysore's industries.

In places the Precambrian rocks of Mysore, which are from 4,600,000,000 to 570,000,000 years old, are highly mineralized; manganese ore, iron ores, mica, chromite, copper ore, and bauxite are obtained in small quantities. From the region of the Kolār Gold Fields, 85 percent of India's gold is obtained. The principal gold mines, now nationalized, have workings up to 9,000 feet deep.

The mineral resources of the state provide the basis for the iron and steel industry at Bhadrāvati and heavy engineering works at Bangalore. Other industries in the state include cotton milling, sugar processing, and the manufacture of cement and paper. Both Mysore city and Bangalore have old established silk industries.

Isolation caused by physical features

Transportation. The obstacle formed by the Western Ghāts has prevented the building of railroads linking the ports to the plateau in the interior. There are 1,700 miles of track, most of which has a one-meter gauge. There are 95 miles of narrow gauge track. Bangalore in the southeast is the main focus of rail transportation. The port of Mangalore in the southwest, which is connected with the state of Kerala by tracks running parallel to the coast, is being improved.

Of the 24,000 miles of surfaced roads in the state, more than half is situated in the original state of Mysore on the plateau. The state as a whole compares favourably with the rest of India in road development, having 32 miles of surfaced road for every 100 square miles of territory, as compared with the national average of 16 miles per 100 square miles. The import and export trade relies primarily on road transport, but many roads in the western part of the state become impassable during the rainy season. Good roads run from Bangalore east to Madras, north to Hyderābād, northwest to Bombay, and westward through Mysore city and Mercāra to the west coast at Mangalore. Bangalore also has a national airport.

Cultural life. Mysore possesses a rich cultural heritage, compounded by the contributions of successive dynasties, which have fostered various religions and philosophies, which, in turn, have influenced literature, architecture, folklore, music, painting, and minor arts. Mauryan architecture is to be seen at Sravana Belgola, 56 miles from Mysore, where there is a giant stone figure of Gommateśvara, the Jain sage, believed to be 2,000 years old. Huge monolithic Jain statues are peculiar to the Kannada culture. The influence of the Cālukya and the Pallava dynasties is still apparent in temple architecture dating from the 7th century AD.

BIBLIOGRAPHY. LEWIS RICE, *Mysore: A Gazetteer*, 2 vol. (1897), is an authoritative work on the historical aspects of Mysore. See also the multivolume *Mysore Gazetteer* (1927–30), ed. by C. HAYAVADANA RAO, and his *History of Mysore, 1399–1799 A.D.*, new ed., 3 vol. (1943–46). Research papers on contemporary social conditions by G.K. GHORI are collected in *Man and Food in Mysore* (1957). L. DUDLEY STAMP, in *Asia*, 12th ed. (1967), provides a discussion of the economic and geographic situation. For current statistical data, see A.T.A. LEARMONTH and L.S. BHAT (eds.), *Mysore State*, vol. 1, *An Atlas of Resources* (1961), and the various censuses of India.

(G.K.G.)

Mystery Religions

Mystery religions were secret cults of the Greco-Roman world that offered to individuals a way to feel religious experiences not provided by the official public religions. They originated in tribal ceremonies that were performed by primitive peoples in many parts of the world. But, whereas in these tribal communities almost every member of the clan or the village was initiated, initiation in Greece became a matter of personal choice. The mystery religions reached their peak of popularity in the first three centuries of the Christian Era; at that time an individual could choose among many religions of this type. Their origin, however, goes back to the earlier centuries of Greek history.

Etymology of the word mystery

Etymologically, the word mystery is derived from the Greek verb *myein* ("to close"), referring to the lips and the eyes. Mysteries were always secret cults into which one had to be "initiated" (taken in). The initiate was called *mystēs*, the introducing person *mystagōgos* (leader of the *mystēs*). The leaders of the cults included the *hierophantēs* ("revealer of holy things") and the *dadouchos* ("torchbearer"). The constitutive features of a mystery society were common meals and common dances and ceremonies. People who had taken the same food and drink, joined in the same dance, and witnessed the same dramatic performances belonged together.

HISTORY

Hellenic roots. *Dionysiac.* In every Greek city the god Dionysus was worshipped by fraternities and sororities and also by mixed communities. Dionysus was a god of fruitfulness and vegetation but especially of wine. The Dionysiac festivals provided an opportunity for stepping outside of the daily routine. The festivals included not only drinking of wine and engaging in sexual activity but also participating in such significant features of Greek civilization as choral singing and mimes. In many cases, only the initiated could participate in the ceremonies. As almost every Greek did join in, initiation into the Dionysiac cult might be compared to tribal initiations (although the high quality of the art and literature produced within the religion of Dionysus lift these ceremonies to so different a level that the comparison with tribal initiations may be misleading). It seems that initiation into the Dionysiac Mysteries was accompanied by initiation into sexual life. The act of producing offspring, however, could never be wholly separated from the thought of death, so that the worshippers of Dionysus were aware of a mystic communion among the ancestors, the living generation, and the future members of the community.

Eleusinian. The most important sanctuary of Demeter (Ceres), the goddess of grain, and her daughter Kore (Persephone) was in the city of Eleusis in Attica, between Athens and Megara. Famous religious agricultural festivals—known as the Greater and the Lesser Eleusinian Mysteries—celebrating the sowing, sprouting, and reaping of the grain, were re-enacted in this city. The cy-

The myth of Kore

cle of the grain, pictured in the myth of Kore (Persephone), was thought to be parallel to the cycle of man. The myth, as told in the Homeric hymn to Demeter, tells how Hades (Pluto, or Pluton), god of the netherworld, wanted a wife and how he carried off Kore into the depths of the earth. Her mother, Demeter, through long days of searching, during which she came to Eleusis, refused to make the grain grow. Finally, Hades was bidden to send Kore back to earth. She came back to light as the grain maiden and gave birth to her son Plutus (Kore, "the maiden"; Pluton, "the rich one"; Plutus, "wealth," especially in grain). But, because Kore had eaten a pomegranate seed, a symbol of death and birth, she could not be completely released, and a compromise was reached by which she spent one-third of the year with her husband, the rest with her mother. Satisfied with this, Demeter caused grain to grow again and taught the Eleusinians her rites. The entire story of Demeter and Kore was elaborately re-enacted in the Eleusinian ceremony. Just as in the myth Kore was carried away to marry Hades and to give birth to Plutus, so was grain thrown into the field and buried in the earth to bring forth new life. Just as grain came up out of the ground and was reaped to yield man's bread and to be used as seed, so was a girl taken from her parents and her virginity "killed" to bring forth new offspring. And when a man died, he was buried in the earth to partake mystically in the cyclic renewal of life. This was the message of Eleusis: out of every grave new life grows—for the initiates there are "good hopes" for a glorious immortality in the afterlife.

Although there were festivals of Demeter throughout Greece, the true Eleusinian Mysteries were celebrated at Eleusis only. At first, the cult of Demeter was local and initiation was tribal rather than personal. By participating in the mysteries, a man became a full member of the civic body. This was changed when Eleusis was annexed to the Athenian territory about 600 BC. Initiation lost its importance as a means of conferring civic status; it became a purely religious ceremony. Every Athenian was admitted to the Eleusinian Mysteries, and soon the mysteries were open to every Greek, so that the ceremonies received an "international" character. Whoever wished to be initiated, however, had to go to Eleusis. It was a day's journey from Athens, a longer distance from most of the other Greek cities. The mystery rite became no longer a tribal ceremony. Each person had to decide for himself whether or not he wanted to be initiated. This development was possible only because Athens had become a large city with a differentiated culture that gave the individual ample choice of a way of life, including religion.

Meaning of the Dionysiac and Eleusinian mysteries

Both Dionysiac and Eleusinian mysteries had a wide range of meaning. Their essence was not contained in any written record but only in the festivals themselves—the holy days of the community. Many participants appreciated only the superficial level of the ceremonies and considered them as an opportunity for having a good time—good company, good food, intoxication, and sometimes (in the Dionysiac cult) sexual pleasures. The ceremonies were open to a deeper understanding, however, that was not made explicit by any theology or by any set of creeds but by the religious action itself, which contained the meaning and conveyed it to the participants without the interposition of words. Therefore, it was not possible to disclose to the noninitiated the mysteries by words, but it was treachery to reveal the secret dances.

Secular mystery communities. A society of initiates could drop its religious connections and become merely a social club. But because secrecy, common meals, and common drinking were implied, the Greeks and Romans regarded such clubs as mystery societies; they did not differentiate between religious associations and private clubs. The role of aristocratic clubs in Athenian politics was very important. In 415 BC the famous mystery scandal occurred. Several aristocratic societies conspired to overthrow the Athenian democracy. In order to pledge all members, a common crime was committed in which each member had to participate. One night the members of the social clubs took hammers and removed the genitals of the many Hermes statues in the city. Whoever would desert the common political cause would be denounced by his former friends for having committed a crime against religion, and many witnesses against him would be at hand. The people of Athens immediately understood that a conspiracy was developing. By a series of severe trials, the conspirators were traced and exiled. The speech of the orator Andocides, one of the conspirators, delivered in his defense in 399 BC, when the old affair was again taken up in a trial, still survives. The title of the oration is "About the Mysteries."

The secular mystery clubs continued throughout Greek and Roman history, and it was often difficult to distinguish them from religious associations. The Romans were especially distrustful of secret societies. This suspicion was justified in the case of Catiline, who led a conspiracy that attempted to overthrow the government in 63 BC. But Trajan, the Roman emperor from AD 98 to 117, did not allow the citizens of Nicomedia (modern İzmit, Turkey) to form a club that planned to provide a fire brigade, and he only reluctantly allowed the citizens of Amisus (modern Samsun, Turkey) to establish an association for charitable purposes.

Orphic. Besides community initiations, there were ceremonies for individual persons of deeper religious longing. Such persons were called Orphics after Orpheus, the Greek hero with superhuman musical skills who was supposedly the author of sacred writings; these writings were called the Orphic rhapsodies and they dealt with such subjects as purification and the afterlife. It is possible to reconstruct a common pattern for these initiations of individuals, although an Orphic "church" never existed, and the doctrines of the many small communities of individualists varied on a broad scale.

The Orphic sense of sin

Many Orphics seem to have had a strong feeling of sin and guilt. They believed that there was a divine part in man—his soul—but it was wrapped up in the body, and man's task was to liberate the soul from the body. This could be achieved by living an Orphic life, which included abstinence from meat, wine, and sexual intercourse. After death the soul would be judged. If a man had lived a righteous life, his soul would be sent to the meadows of the blessed in Elysium; but, if he had committed misdeeds, his soul would be punished in various ways and perhaps sent to hell. Following a period of reward or punishment, the soul would be incarnated in a new body. Only a soul that had lived a pious life three times could be liberated from the cycle.

Pythagoreans. The Orphic creeds were the basis of the Pythagorean brotherhood, which flourished in southern Italy beginning in the 6th century BC. The Pythagoreans were aristocratic fraternities that sometimes had a political scope. Their main achievements, however, lay in the fields of music, geometry, and astronomy. They discovered that these subjects could be explained by numbers and ratios. Combining Orphic eschatology (the study of the last things, especially death and afterlife) with their discoveries, they invested music, geometry, and astronomy with religious values. According to their doctrine, the original home of the soul was in the stars. From there it fell down to earth and associated with the body. Thus, man was a stranger on the earth, and he had to strive to liberate himself from the ties of the flesh and return to the soul's celestial home.

Platonists. The philosophy of Plato (*c.* 428–348 or 347 BC) by no means resulted from connections with a mystery cult. Yet Plato did take up many ideas from earlier Greek religion, especially from the Pythagorean brotherhood and from the Eleusinian communities, and often described his philosophy in terms derived from the mysteries. For example, the notion of searching and finding, so important in Eleusis, became an important notion in Plato's philosophy: the philosopher should never cease or relax in his quest for truth. A value was thus attached to the very act of searching. Later mystery religions, in their turn, borrowed freely from the rich imagery of Plato's dialogues and are thus deeply tinged with Platonism.

In the *Timaeus,* which is an exposition of his theory of

Plato's theory of the soul

the universe, Plato also developed his theory of the soul. The earth is surrounded by the spheres of the seven planets; the eighth sphere is that of the fixed stars. Beyond the eighth sphere is the realm of the divine. The sphere of the fixed stars, moved by the divine, continuously turns to the right at an even speed. This clockwise rotation affects the spheres of the planets, although they have their proper movement, which runs to the left, or counterclockwise. The sphere of mortality begins with the planets. The original home of each soul is in one of the fixed stars. As a result of the movement of the spheres, the soul falls through the planetary spheres to earth, where it is united with the body. The soul must then try to liberate itself from the body and ascend to the fixed star from which it fell. In later generations this picture was vividly worked out. The soul, in the course of its fall through the planetary spheres, was thought to acquire the qualities of the planets: sloth from Saturn, combativeness from Mars, lust for power from Jupiter, voluptuousness from Venus, greed from Mercury. After death, when the soul returned to the fixed star, it discarded these qualities, just as the *mystēs*, in certain initiations, discarded his everyday garment before entering the sacred place.

Many other traditional religious images were taken over by Plato, including the music of the spheres, the migration of the soul, the soul's remembrance of its celestial origin, and the idea of rewards for the righteous and punishment for the wicked. Later mystery associations adopted these concepts, which Plato had expressed so beautifully, and were deeply influenced by Plato's explanations.

The Hellenistic period. When Alexander the Great conquered the Asiatic kingdoms as far east as the Indus River, the Greek world was extended immensely. The religious ideas in Greece itself and the western part of the Alexandrian Empire, however, changed very slowly, because the Greeks, now masters of the world, felt no need for change.

In the Messenian town of Andania mysteries were celebrated in honour of the goddesses Demeter and Kore. A long inscription of 92 BC gives elaborate directions for the conduct of the rites, although, naturally, it gives no details of what went on during initiation. The mysteries in honour of the Cabeiri (gods of fertility) on the island of Samothrace attracted great attention in this period. These gods were thought to be helpers of the seafarers, and initiation into their mysteries was looked upon as a general safeguard against all misfortune but particularly against shipwreck. The Dionysiac Mysteries, with their revels and merriment, continued throughout the whole of Greek history. Together with most of the elements of Greek civilization, this cult was transferred to Italy. In 186 BC a scandal about the Bacchanalia—the Latin name for the Hellenistic Dionysiac Mysteries—so upset the Romans that a decree of the Senate prohibited them throughout Italy, except in certain special cases. These mysteries were celebrated in a lower middle class milieu and involved gross sex parties and violence conducted under the cover of mystery secrecy.

Hellenistic Mysteries in the Greek Orient

The important developments in the mystery rites during the Hellenistic period took place in the Greek Orient, where elements from the Greek and Oriental religions were blended. Contact with Greek civilization completely changed life in the Orient, where the knowledge of writing had been confined to a few priests and scribes. Society first disintegrated after the conquest of Alexander and then developed along new lines. Changes in religion were inevitable, and some influence of Oriental traditions upon the Greeks was bound to follow. But the process was a slow one and became manifest only a few centuries later.

With regard to the institution of kingship, however, syncretism worked quickly. Ancient Near Eastern kingship was originally sacral. The Syrian and Egyptian inhabitants of the newly created Greek kingdoms inevitably regarded the Greco-Macedonian kings as semidivine beings. The Greeks themselves soon submitted to this mixture of politics and religion. Such a mixture was perfectly natural to the Egyptians and Syrians, who did not perceive the structure of society as an abstraction, such as "the state" or "the nation," but saw the unity of the body politic in the person of the king. He was the symbol of the security and help that man derives from an orderly society. Mystery rituals, called royal mysteries, were developed especially in Egypt. According to traditional Egyptian religion, the ruling pharaoh was an incarnation of Horus (the sun-god), his mother or wife an incarnation of Isis (the heavenly queen), and his deceased father an incarnation of Osiris (the god of fertility). In Hellenistic times, Osiris was commonly known by the name Sarapis. These gods became equated with Greek gods: Isis with Demeter and Aphrodite; Horus with Apollo and Helios; Sarapis with Zeus, Dionysus, and Hades (Pluto). Both Greek and Egyptian myths were adopted for these divinities.

One of the suburbs of Alexandria, the newly constructed Greek capital of Egypt, was called Eleusis after the city of Demeter in Greece, and the Eleusinian Mysteries were instituted in a Greco-Egyptian adaptation. Dionysiac Mysteries were introduced on an even greater scale, so that the royal court was temporarily thrown into turmoil by the number of Bacchic ceremonies in which the king was considered to be a reincarnation of Dionysus. The Pythagorean concept of the migration of the soul was also taken over and was blended with the Egyptian belief in the reincarnation of the sun-god Horus in the reigning king.

The cult of rulers thus introduced ideas from the Greek Orient into Greek communities. But the mixture of religion and politics was a great obstacle for the propagation of the Greco-Oriental mysteries in the Mediterranean world. Even the numerous Greeks who lived in Egypt and Syria maintained the traditional Greek concept of the separation of god and man, and it was only after the political aspect of the mysteries was discarded that the religious elements could gain a life of their own. Inscriptions discovered on the Greek island of Delos demonstrate this well. The worship of Sarapis was introduced at Delos during the time the island was temporarily a naval base of the Greco-Egyptian kings. When the Egyptian influence on the island receded, the cult of Sarapis not only remained but reached new heights. The Romans later used Delos as a free port for the eastern part of the Mediterranean, and from there the worship of Sarapis and Isis spread to most of the harbours of the Greek world and to the cities in the Bay of Naples, whence it was brought by Italian merchants to Rome.

The combination of mystery elements with ruler worship is also evident in the kingdom of Commagene (eastern Turkey and northern Syria). Here, the kings assigned large funds to construct throughout the country gigantic sanctuaries, where festivals of the gods and the royal ancestors were celebrated annually on the kings' anniversary days. Long inscriptions discovered in the remains of these sanctuaries bear striking similarities to the language of the mysteries. The ceremonies, however, seem to have contained little true religion.

Roman imperial times. The great period of the mystery religions began when the Romans imposed peace upon the Mediterranean world. The Dionysiac, or Bacchic, societies flourished in the whole empire—in Greece proper, on the Greek islands, in Asia Minor, along the Danube River, and especially in Italy and at Rome. Hundreds of inscriptions attest to Bacchic Mysteries. In some circles, Orphic and Dionysiac ideas were blended, as in the community that met in the underground basilica near the Porta Maggiore ("Major Gate") at Rome. There was also a blend of ideas in the community for which the Orphic hymns were written. The members of this community (probably in Asia Minor) assembled at night in a clubhouse and held their services by the light of torches. Their rite consisted of a bloodless sacrifice and included the use of incense, prayer, and hymns. In addition to the mystery cults that were familiar from earlier times, the national religions of the peoples of the Greek Orient in their Hellenized versions began to spread. A faintly exotic flavour surrounded these religions and made them particularly attractive to the Greeks and Romans. The most popular of the Oriental mysteries was the cult

of Isis. It was already in vogue at Rome in the time of the emperor Augustus, at the beginning of the Christian era. The Emperor, who wanted to restore the genuine Roman religious traditions, disliked the Oriental influences. But men of reputation, such as Messalla, a general and patron of writers, were strongly inclined toward the Isis Mysteries. Isis, the goddess of love, was the patroness of many of the elegant Roman courtesans. The religion of Isis became widespread in Italy during the 1st and 2nd centuries AD. To a certain extent, the expansion of Judaism and Christianity over the Roman world coincided with the expansion of the Egyptian cults.

Mysteries from Asia Minor

Far less important was the influence of cults from Asia Minor. By 200 BC the Great Mother of the Gods (Magna Mater) and her consort Attis were introduced into the Roman pantheon and were considered as Roman gods. Their cult seems to have been encouraged especially under Emperor Claudius about AD 50. The Great Mother was characterized by her universal motherhood, especially over wild nature. The mysteries symbolized, through her relationship to Attis, the relations of Mother Earth to her children and were intended to impress upon the *mystēs* the subjective certainty of having been united in a special way with the goddess. There was a strong element of hope for an afterlife in this cult. The Persian god Mithra (Mithras), the god of light, was introduced much later, probably not before the 2nd century. The cult of Mithra was concerned with the origin of life from a sacred bull that was caught and then sacrificed by Mithra. According to Persian sources, the bull by its death gave birth to the sky, the planets, the earth, the animals, and the plants; thus Mithra became the creator of life. From Syria came the worship of several deities, of which Jupiter Heliopolitanus (the local god of Heliopolis; modern Ba'labakk, Lebanon) and Jupiter Dolichenus (the local god of Doliche in Commagene; modern Dülük, Turkey) were the most important. Adonis (a god of vegetation) of Byblos (in modern Lebanon) had long been familiar to the Greeks and was often considered to be closely related to Osiris; the myths and rituals of the two gods were similar. Adonis' female partner was Atargatis (Astarte), whom the Greeks identified with Aphrodite. At the time of Emperor Marcus Aurelius, in the latter half of the 2nd century AD, a "pseudoprophet" named Alexander of Abonouteichos (on the north coast of Turkey) devised a great mystery spectacle centred around a holy snake called Glycon and had great success during his lifetime.

The height of Syrian influence was in the 3rd century AD when Sol, the Syrian sun god, was on the verge of becoming the chief god of the Roman Empire. He was introduced into Rome by the emperor Elagabalus (Heliogabalus) in about AD 220, and by about AD 240 Pythian Games (*i.e.*, festivals of the sun god Apollo Helios) were instituted in many cities of the empire. The emperor Aurelian (270–275) elevated Sol to the highest rank among the gods. Sanctuaries of Sol and the gods of other planets (*septizonium*) were constructed. Even the emperor Constantine the Great, some 50 years later, wavered between Sol and Christ. For some time his religious policy was devised so as to allow the coexistence of both religions. Finally, Christianity was accepted as the official religion.

The different mystery religions were not exclusive of one another, but they appealed to different sociological groups. The middle class of the Greek and Roman cities preferred the Dionysiac societies, the festivals of which were a cult of beauty and merriment. Isis was worshipped by lower middle class people in the seaports and trading towns. The followers of the Great Mother in Italy were principally craftsmen. Mithra was the god of soldiers and of imperial officials and freedmen. There were no special societies for slaves; but they were usually admitted to the societies, and, during the time of the festival, all men were considered equal.

BELIEFS AND PRACTICES

Common features in Roman imperial times. For the first three centuries of the Christian Era, the different mystery religions existed side-by-side in the Roman Empire. They had all developed out of local and national cults and later became cosmopolitan and international. The mystery religions would never have developed and expanded as they did, however, without the new social conditions brought about by the unification of the Mediterranean world by the Romans. In the large cities and seaports, men from the remotest parts of the empire flocked together. Many people were removed from their accustomed surroundings and suffered from loneliness. They longed for new acquaintances and for assimilation, and they needed the assurance that only the knowledge of belonging to a community can give. Economic and political conditions in the Roman Empire also accelerated the growth of the mysteries. Members of a mystery society helped one another. For a lawyer, a craftsman, or a contractor, membership in a club could be the road to success. Furthermore, there is less opportunity for private initiative in a society ruled by a monarch than in a democratic society. The individual who felt that his initiative was frustrated by the preponderance of the imperial structure might well turn to a community that offered him the hope of a better future. The mystery societies, thus, commonly satisfied both a taste for individualism and a longing for brotherhood. At least in principle, the members of the communities were considered equal: one man was the other man's brother, irrespective of his origin, social rank, or nationality.

Mysteries fostered by new social conditions in the Roman Empire

Because membership in each of the mystery communities was a matter of personal choice, propaganda and missionary work were inevitable. In the religions of Isis and Mithra, missionary zeal was particularly obvious. The followers of Isis and Mithra considered Rome to be the centre of their worship, and the city was called *sacrosancta civitas* ("sacred city") in an Isis romance written in the 2nd century AD by the Latin author Apuleius.

Priesthood. The organization of the mystery religions was rather loose. The priests of Dionysus were wealthy laymen, as the priests in Greece always were. The Roman community of the Great Mother had a large group of priests (the Galli), headed by a chief priest (the Archigallus). They were eunuchs who wore female garb, who kept their hair long and perfumed with ointment, and who celebrated the goddess' rites with wild music and dancing until their frenzied excitement found its culmination in self-scourging, self-laceration, or exhaustion. Besides the priests there were priestesses and many minor officials. The followers were organized according to their function in the ritual procession as bearers of the tree (*dendrophori*) or bearers of the reed (*cannophori*). The men who carried the statue in the rites of Jupiter Dolichenus were called the sedan-chair men (*lecticarii*).

The higher grades of the Isis Mysteries were reserved to persons born of the priest caste of Egypt. To be born into this caste was more important than talent or skill. This limited the quality of the priests and was a serious disadvantage in the community's competition with other religions. But a second way of advancement within the religious group was devised for men of Greek or Roman origin. In Egypt, there was a group of elevated laymen—the porters of the holy shrine (*pastophori*). They were inferior in rank to everyone of the priest caste; but in Greek and Roman countries the rank of the *pastophori* became a surrogate for the native priest caste of Egypt. The *pastophori* were, in fact, the religious leaders of the communities.

Rites and festivals. A period of preparation preceded the initiation in each of the mysteries. In the Isis religion, for example, a period of 11 days of fasting, including abstinence from meat, wine, and sexual activity, was required before the ceremony. The candidates were segregated from the common folk in special apartments in the holy precinct of the community centre; they were called "the chastely living ones" (*hagneuontes*).

In all the mystery religions the candidates swore an oath of secrecy; the oath of the Isis Mysteries is preserved on papyrus. Before initiation, a confession of sins was expected. The candidate sometimes told at length the story of the faults of his life up to the point of his baptism, which was commonly a part of the initiation ceremony,

and the community of devotees listened to the confession. It was believed that the rite of baptism would wash away all the candidate's sins, and, from that point on, his life would be changed for the better, because he had enrolled himself in the service of the saviour god.

In the Mithraic ceremonies, there were seven degrees of initiations: Corax (Raven), Nymphus (Bridegroom), Miles (Soldier), Leo (Lion), Perses (Persian), Heliodromus (Courier of the Sun), and Pater (Father). Those in the lowest ranks, certainly the Corax, were the servants of the community during the sacred meal of bread and water that formed part of the rite.

Initiation ceremonies

The initiation ceremonies usually mimed death and resurrection. This was done in the most extravagant manner. In some ceremonies, candidates were buried or shut up in a sarcophagus; they were even symbolically deprived of their entrails and mummified (an animal's belly with entrails was prepared for the ceremony). Alternatively, the candidates were symbolically drowned or decapitated. In imitation of the Orphic myth of Dionysus Zagreus, a rite was held in which the heart of a victim, supposedly a human child, was roasted and distributed among the participants to be eaten.

The baptism could be either by water or by fire, and the rites often included actions that had an exotic flavour. Sulfur torches were used during the baptism ceremony; they were dipped into water and then—contrary to the expectations of the observers—burned when drawn out of the water. In a dark room a script would suddenly become visible on a wall that had been prepared accordingly. Instructions still exist for producing a nimbus effect —the appearance of light around the head of a priest. The priest's head was shaved and prepared with a protective ointment; then a circular metal receptacle for alcohol was fixed on his head; it was set aflame in a dark room and would shine for some seconds. In the Dionysus and Isis mysteries, the initiation was sometimes accomplished by a "sacred marriage," a sacral copulation. Two cases are known in which a priest speaking from the statue of the god ordered a credulous woman to come to the temple and be the god's concubine, the part of the god being enacted by the priest.

The initiation ceremonies were usually accompanied by music and dance and often included a large cast of actors. In the Dionysiac societies, especially elaborate provisions were made for mimic representations. The names of the sacred roles varied from place to place; among the roles were: Dionysus and Ariadne (a vegetation goddess and wife of Dionysus), Palaemon (a marine deity), Aphrodite (the goddess of love and beauty), Proteurhythmos (the inventor of elegant rhythm), the "foster-father of Dionysus," Kore, Demeter, Asclepius (the god of medicine), Pan (the god of flocks and shepherds), Curetes (long-haired youths), nymphs (minor nature goddesses), shepherds, sileni and satyrs (creatures of the wild, part man and part beast), maenads (female attendants who shared in the nocturnal orgiastic rites of Dionysus), the "guardian of the grotto," and centaurs (a race of beings half man and half horse).

The ceremonies always contained a prayer for the welfare of the emperor and for the good fortune of the whole Roman Empire. In fact, the amalgamation of religion and politics was sometimes so close that the designation "imperial mysteries" is used. The pattern of imperial mystery ceremonies could vary widely. This was especially true of the Dionysiac rites. In the clubs of the upper middle class and wealthy, for example, the festivals were chiefly social events. But the members of these communities were grateful for the security and peace and for the opportunity to make a good living that the emperor guaranteed to them. They felt loyalty toward the Roman Empire and expressed this by ceremonies of the imperial mysteries.

Dionysus was the patron god of the important international society of actors, and their reunions were celebrated in the mode of Dionysiac Mysteries. When an emperor travelled in the empire, responsibility for dignified receptions of him was handed over to the society of actors. Because his route was known beforehand, a voyage of the emperor was turned into a series of pompous festivals that were organized in a manner closely resembling mystery ceremonies.

The meetings of the mystery clubs were often named after the common meal. The Dionysiac meetings were called *stibas* ("straw") because the participants ate their dinner sitting on straw. The meals of the followers of Sarapis and Attis were called *klinē* ("couch"), because the diners lay on couches.

Seasonal festivals. The religions of Dionysus and Demeter and of Isis and the Great Mother had something of an ecclesiastical year. The seasonal festivals were inherited from old tribal ceremonies that had been closely associated with the sowing and reaping of corn and with the production of wine. The dates varied greatly according to the geographical conditions and the emphasis of the seasonal rites in the country in which the mysteries had originated. Dionysiac festivals were held in all four seasons; vintage and tasting of the new wine were the most important occasions. But the religion of Dionysus was closely associated with that of Demeter, and, thus, sowing and reaping were also celebrated in Dionysiac festivals. In the religion of the Great Mother, a hilarious spring festival celebrating the renewal of life was enacted in Rome.

Festivals connected with agriculture

The festivals of the Isis religion were connected with the three Egyptian seasons caused by the cycle of the Nile River (inundation, sowing, and reaping). About July 19, when the whole country was almost desiccated by the heat and the drought, the high waters of the new flood miraculously arrived from Ethiopia. On that day, just before sunrise, Sirius (the Dog Star, or the star of Isis) would make its first appearance of the season on the horizon. This was the sacred New Year's Day for the Egyptians, and the festival of the Nile flood was their greatest festival. There were, in addition, the festivals of sowing and reaping. But because the Egyptian year was a solar year of 365 days without intercalation (leap years), the seasonal festivals that were fixed upon a particular date were retarded by one day every four years and complete confusion resulted. The Romans fixed the calendar of Egypt by introducing an intercalary day every fourth year. In Roman times, important Isis festivals were held on December 25, January 6, and March 5. The March festival, as it was celebrated in Corinth, is described at length in Apuleius' *Metamorphoses*, or *The Golden Ass*. It was a spring festival that celebrated the beginning of the seafaring season. A ship was carried on a cart (*carrus navalis*) through the city. It was followed by a procession of choruses, candidates, *mystai* in bright clothes wearing masks, and priests carrying the insignia of the goddess. The ship was let into the sea, and the participants returned to the temple, where initiation ceremonies, banquets, and dances were held.

In the religion of Sol, the festivals were determined by astronomy. The greatest festival was held on December 24–25, at the time of the winter solstice. Because from this date the length of the day began to increase, it was regarded as the day of the rebirth of the god and of the renovation of life.

Literature. The mystery communities had religious hymns, but almost nothing of them has been preserved. The initial words of some hymns from the Sta. Prisca Mithraeum in Rome are known, and some Isiac poems exist. More important is a text of 40 sentences in which the goddess Isis reveals herself; it was found at four different and geographically distant places and was probably exhibited in every Isis sanctuary. Narratives of the miracles wrought by the gods were preserved in many temple libraries; examples of these narratives, on papyrus and on stone, have been found. According to a recent theory, the literary genre of the romance was developed from these narratives. The last part of the *Metamorphoses* of Apuleius is an Isis text and narrates in detail the initiation into the Egyptian mysteries.

Hermetic writings and *Chaldean Oracles*

Hermes Trismegistos, the Greek name for the Egyptian god Thoth, was the reputed author of treatises that have been preserved. Thoth was the scribe of the gods, the inventor of writing, and the patron of all the arts

dependent upon writing; he was sometimes thought of as an attendant of Isis and sometimes as the repository of all wisdom. These treatises are not exactly mystery texts, but they are works of revelation on occult subjects and on theology. Because the pagan mysteries had no official creed, each congregation of initiates was free to construct a theology of its own and to change it again. The Hermetic writings were attempts to provide a theology for a particular community. Although no authorized interpretation could exist for a doctrine that was in constant fluctuation and although none of the Hermetic treatises could claim to be the correct interpretation of the pagan mysteries, nevertheless, the texts give an instructive picture of spiritual life in mystery communities.

There are some contemporary texts that shed light on the mystery communities. Plutarch, the Greek biographer, wrote the philosophical treatise "About Isis and Osiris," which gives an interpretation of the Isis Mysteries. Arnobius, a 3rd-century Christian apologist, described an interesting semiphilosophical, semireligious mystery community known as the *viri novi* ("the new men"). Arnobius seems to have lived among them in North Africa for a time before his conversion to Christianity. They had a religious doctrine of the soul, with marked affinities to the teachings of the Neoplatonic thinkers Plotinus and Porphyry.

Only fragments are preserved of the *Chaldean Oracles,* a theosophical text in verse that was composed by Julianus the Theurgist and his son late in the 2nd century AD and had great influence on the Neoplatonists. The work combined Platonic elements with Persian or Babylonian creeds and was regarded by the later Neoplatonists as their basic religious book, something of a heathen bible. The doctrine of the *Chaldean Oracles* was associated with esoteric fire rituals. Julianus and his followers were called theurgists—*i.e.,* men who could perform divine operations. Their religion was partly one of meditation about the hidden and wondrous magical processes within the cosmos.

Theology. The creeds of the mystery religions were never worked out to the same extent that the Christian creeds were. Nevertheless, the doctrines of the mysteries may be called a theology. One of the central subjects in mystery writings was cosmogony—the theory of the origin or creation of the world. In the Hermetic treatises, in the *Chaldean Oracles,* and in the little known writings of Mithraism, the cosmogony was modelled after Plato's *Timaeus,* and it always dealt with the creation of the soul and with the soul's subsequent fate.

The myth of the soul

The theological doctrine of the soul and the myth about its celestial home, its fall, and its redemption were inseparable. The sequence is beautifully told in the "Hymn of the Soul," preserved in the *Acts of Thomas,* an apocryphal account of the journeys and death of the apostle in which some episodes were certainly transmitted from pagan mystery texts. The hero of the hymn, who represents the soul of man, is born in the Eastern (the yonder) Kingdom; immediately after his birth, he is sent by his parents on a pilgrimage into the world with instructions to take a pearl from the mouth of a dragon in the sea. Instead of wearing his heavenly garment, he dresses in earthly clothes, eats earthly food, and forgets his task. Then his parents send a letter to rouse him. As soon as he has read the letter, he awakes and remembers his task, takes the pearl, and begins the homeward journey. On the way, his brother (the Redeemer) comes to accompany him and leads him back home to his father's palace in the east. This myth is a figurative representation of the theological doctrine of the soul's fall and its return to heaven.

Many of the questions that were the subject of later Christian theological discussions were already eagerly debated in the mystery religions. In a Hermetic treatise, for example, the existence of God was proved from the evident order of the world. This argument, which had first been formulated by Zoroaster, a 7th-century Iranian prophet, was expressed in the form of questions: Who could have created the heavens and the stars, the sun and the moon, except God? Who could have made wind, water, fire, and earth (the elements), the seasons of the year, the crops, the animals, and man, except God?

Passionate debates were held about the question of whether man was subject to blind fate. The Stoics (a Greek and Roman school of philosophy holding that men should be free from passion and calmly accept all occurrences as unavoidable) had adopted the doctrine that all events are determined by the stars. Thus, for many Greeks and Romans astrology became the only sensible method of studying man's life and fortune. But, for others, the idea that man could achieve nothing by his own will was frightening, and they wanted to be liberated from this fear; the mystery religions promised to liberate them. The theology of the mystery religions admitted that the stars ruled the world and especially that the planets had evil influences. But the highest god of the religion (*e.g.,* Sarapis in the Isiac Mysteries) stood far above the stars and was their master. A man who decided to become a servant of this god stepped out of the circle of determination into the sphere of liberty. The god could suspend determination, because he ruled over the stars; he could unravel the threads of the Moirai (the three spinners of fate); he could save his servant from illness and prolong his life, even against the will of fate. In the Isis Mysteries there was a theology of grace foreshadowing Christian doctrine.

In many of the mystery cults, there was a marked tendency toward henotheism—the worship of one god without denying the existence of other gods. Thus, Isis was the essence of all pagan goddesses; Sarapis was the name uniting the gods Zeus, Pluto, Dionysus, Asclepius, Helios, and even the Jewish god YHWH (Yahweh). In the religion of Sol, an elaborate syncretistic theology was developed to show that all known gods of all nations were nothing but provisional names for the sun god.

RELIGIOUS ART AND ICONOGRAPHY

Much of Greco-Roman art was executed for use in the mystery communities. The Dionysiac monuments are by far superior to all others in artistic quality. This is to be expected, because the worship of Dionysus often took the form of a worship of beauty. Nevertheless, the other communities also produced a great number of art objects.

Superior artistic quality of Dionysiac art

Architecture. The mystery religions developed different types of edifices for their purposes. Every Greek city had temples and precincts of Dionysus. The Isis Mysteries adopted the Greek temples, frequently adding a cupola. Many Isis temples were modest in size, but the temple at Pergamum (modern Bergama, Turkey) was a great basilica with a vaulted roof and strong towers, in the fashion of the best Roman architecture. The Isis temple that the emperor Domitian erected on the Campus Martius (the Field of Mars) in Rome at the end of the 1st century AD was a stately building, and the Temple of Sarapis (the Sarapeum) at Alexandria was a huge construction. The subterranean basilica near Porta Maggiore in Rome (used by an Orphic or Pythagorean society) was a strong and magnificent structure, hidden in a large garden. The Mithraic sanctuaries were artificial caves illuminated from above by light shafts. They were built for communities of from 50 to 100 persons.

The buildings were designed to be functional for the religious ceremonies. The Mithraeum under the Church of S. Clemente at Rome contained a system of underground galleries for initiation ceremonies. Beneath the temple of the Egyptian gods at Pergamum, subterranean passages existed for the use of the priests. One of the paths led into the huge, hollow statue of the god, so that the priest could speak from the mouth of the statue. By another secret way, an officiant could climb the huge corner towers of the temple to make announcements from there. The Sarapeum at Alexandria was directed toward the east; on a certain day of the year, at a certain time, sunbeams directly struck the head of the god's statue. This same temple was so arranged that those waiting to be initiated could hire rooms in an adjacent building during the time of preparation before the ceremony.

Because the use of water was such an important element in most of the mystery rites, the location of the tem-

ples was often determined by the availability of water; Mithraic sanctuaries were always erected on the spot at which a fountain had its source. In the temples of Isis, a cistern for the holy water was required; in Delos and in a house at Pompeii in Italy, a system of water basins could imitate the flood of the Nile. The Dionysus temple at Corinth had an underground system of tubes and barrels that could be operated by buttons from the outside. The priest showed the worshippers of the god a barrel filled with water. They left the temple together, and the door was sealed from without. By pressing the buttons, the water was let out of the barrel, and wine was poured in. The following day, when the seal was removed, the spectator witnessed the Dionysiac miracle of water turned into wine.

On the ground floor of the Mithraic sanctuaries at Ostia, mosaic pavements showed the seven grades of the initiation and their symbols together with the ladder of the seven steps that led to religious salvation. In initiation ceremonies the mosaic was perhaps used to indicate the place where the different participants were to take their places.

Statuary. A great many statues were exhibited in the temples and shrines of the mystery gods. They were usually executed in the traditional Greek style. In the sanctuary of Isis and Sarapis at Thessalonica (modern Thessaloníki), in northern Greece, there were statues of a whole series of Greek goddesses, each of whom was identified with Isis in one way or another to show that the Egyptian goddess was the essence and synthesis of Greek religion. In the 4th century BC, the sculptor Bryaxis created a famous colossal statue of Sarapis in the temple at Alexandria. It represented the god—as a combination of the Greek gods Zeus (the father of the gods), Hades, and Dionysus—seated upon a throne, with Cerberus, the three-headed monster, beside him. An interesting statuette found at Cyrene (modern Shaḥḥāt, Libya) shows a female initiate of Isis. The woman is wrapped from feet to waist like a mummy; but the upper part of her body is free, and she is wearing the crown of Isis on her head. The statue thus showed how an initiate would first die and subsequently resurrect in triumph during the ceremony. Many terra-cotta statues of Isis and her son Horus have survived from Roman Egypt; they are similar to the later statues of the Christian Madonna and Child. Syrian statues of Jupiter Heliopolitanus represent the god in a rigid attitude, like a pillar. In the base of some of these statues are holes, into which sticks could be inserted for the purpose of carrying the statue in procession. In Mithraic sanctuaries a great number of statues, especially of the gods of the planets, were exhibited. Statues of the Mithraic time god were also frequent; they were often hollow and were constructed so that they could spit fire.

Reliefs. The Dionysiac reliefs are numerous. They show symbols of the religion, such as the shepherd's staff, the winnow (an ancient device for separating chaff from grain), and the phallus; they depict the gay life of satyrs and maenads, shepherds and shepherdesses; and they represent the "golden age" of the gods with tame and wild animals enjoying a peace that the god had instituted. A great, silver dish, dating from about the 4th century AD and found at Mildenhall, England, shows the swift and elegant dance of the maenads. Dionysiac sarcophagi represented Bacchic revels and the pastime of the Erotes and Psyches in afterlife. Many reliefs of the Isis Mysteries also survive. They display the mystical cista (a receptacle for carrying sacred objects) with the snake of Horus, the priest carrying the holy water in a procession, female attendants with a ladle, and a man in a dog's mask, who represents Anubis (the guardian god). Other Isiac reliefs show Isis riding on a dog, symbolic of her position as goddess of Sirius (the Dog Star).

In Mithraic caverns there was always a relief depicting the god sacrificing the bull. Representations of the sacramental meal were also frequent; a relief recently discovered in Konjic, Yugoslavia, shows the initiates at a banquet wearing masks, among them a lion, a raven, a soldier, and a Persian. Two reliefs—at Rome (now at Modena, Italy) and at Housesteads, England (the best preserved fort along Hadrian's Wall)—depict the creation of the world out of an initial egg; in this case, Orphic and Mithraic ideas were amalgamated. Other episodes of Mithraic mythology that were commonly displayed include the birth of Mithra from a rock with the shepherds who welcome him and his dealings with the sun god.

The stucco reliefs in the subterranean basilica near Porta Maggiore are of outstanding quality. In the central episode, Sappho—an early Greek poetess who supposedly killed herself in a "lover's leap" from the island of Leucas into the Ionian Sea—is shown leaping toward Apollo, the god of the sun; this symbolized the soul's transcendence into more favourable existence. Many of the reliefs in the basilica allegorize episodes from Greek mythology in the fashion of the Pythagoreans, who found a hidden religious or philosophical meaning behind the mythical tales of the Greek tradition.

Painting. There are few paintings from the temples of the mystery religions that have been preserved; nevertheless, some of these deserve comment. The superb Dionysiac frescoes of the Villa of the Mysteries (Villa dei Misteri) at Pompeii show the initiation of a girl into the Bacchic Mysteries: in one fresco she is lifting the cover of a sacred casket; in a second scene three followers of Dionysus are practicing lecanomancy (divination by the inspection of a bowl filled with water); in a third scene the girl is unveiling an erect phallus and because of this she is being flagellated; finally, she is seen dancing in happy bliss. A number of Isiac frescoes, preserved in the temple of Isis at Pompeii, show the sacred dance of the initiates, the presentation of an urn filled with the ritual holy water to the initiates, the coffin of Osiris and his resurrection, and episodes from the cycle of Io, a Greek heroine equated with Isis. Isiac frescoes dating from the time of the emperor Caligula in the 1st century AD are also found in the ruins on the Palatine at Rome. In the Mithraeum under Sta. Prisca in Rome, two layers of frescoes were found that show the procession of the initiates toward ritual sacrifice of a bull, called Suovetaurilia, and the sacred meal of the sun god and Mithra. Sometimes a fresco replaced the relief of the sacrifice of the bull. The initiation ceremonies are shown in the Mithraic sanctuary at Capua (in western Italy): the candidate, accompanied by the *mystagōgos*, is blindfolded, kneels down, and lies prostrate. At Rome, in the tomb of Vincentius and Vibia, who worshipped the god Sabazius (a Thracian form of Dionysus), frescoes show how Vibia was carried away by Death, as Kore had been carried away by Hades, how she was judged and acquitted, and how she was introduced by a "good angel" to the sacred meal of the blessed.

Mosaics. A mosaic at Antioch represents the Phoenix —the solar bird who died and resurrected from its own ashes and who was its own father and son at the same time—with sunrays encircling its head. A Dionysus mosaic at Cologne, Germany, depicts in several panels the life of satyrs and maenads and also Bacchic symbols such as the winnow (an implement of purification) and the oyster (which has to be liberated from the shell as the soul from the body). The room evidently was used for banquets and Dionysiac merrymaking.

MYSTERY RELIGIONS AND CHRISTIANITY

Christianity originated during the time of the Roman Empire, which was also the time at which the mysteries reached their height of popularity. This was by no means an accident. The Christian theologian Origen wrote in the 3rd century that it was part of the divine plan that Christ was born under the emperor Augustus: the whole Mediterranean world was united by the Romans, and the conditions for missionary work were more favourable than ever before. The simultaneousness of the propagation of the mystery religions and of Christianity and the striking similarities between them, however, demand some explanation of their relationship. The hypothesis of a mutual dependence has been proposed by scholars—especially a dependence of Christianity upon the mysteries—but such theories have been discarded. The similarities must rather be explained by parallel developments from similar origins. In both cases, national religions of a ritualistic type

Similarities between mysteries and Christianity

were transformed, and the transformation followed similar lines: from national to ecumenical religion, from ritualistic ceremonies and taboos to spiritual doctrines set down in books, from the idea of inherited tradition to the idea of revelation. The parallel development was fostered by the new conditions prevailing in the Roman Empire, in which the old political units were dissolved, and the whole civilized world was ruled by one monarch. People were free to move from one country to another and became cosmopolitan. The ideas of Greek philosophy penetrated everywhere in this society. Thus, under identical conditions, new forms of religious communities sprang from similar roots. The mystery religions and Christianity had many similar features—*e.g.*, a time of preparation before initiation and periods of fasting; baptism and banquets; vigils and early-morning ceremonies; pilgrimages and new names for the initiates. The purity demanded in the worship of Sol and in the Chaldean fire rites was similar to Christian standards. The first Christian communities resembled the mystery communities in big cities and seaports by providing social security and the feeling of brotherhood. In the Christian congregations of the first two centuries, the variety of rites and creeds was almost as great as in the mystery communities; few of the early Christian congregations could have been called orthodox according to later standards. The date of Christmas was purposely fixed on December 25 to push into the background the great festival of the sun god, and Epiphany on January 6 to supplant an Egyptian festival of the same day. The Easter ceremonies rivalled the pagan spring festivals. The religious art of the Christians continued the pagan art of the preceding generations. The Christian representations of the Madonna and child are clearly the continuation of the representations of Isis and her son suckling her breast. The statue of the Good Shepherd carrying his lost sheep and the pastoral themes on Christian sarcophagi were also taken over from pagan craftsmanship.

In theology the differences between early Christians, Gnostics (members—often Christian—of dualistic sects of the 2nd century AD), and pagan Hermetists were slight. In the large Gnostic library discovered at Naj'Ḥammādī, in upper Egypt, in 1945, Hermetic writings were found side-by-side with Christian Gnostic texts. The doctrine of the soul, as it was taught in Gnostic communities, was almost identical to that of the mysteries: the soul emanated from the Father, fell into the body, and had to return to its former home. The Greeks interpreted the national religions of the Greek Orient chiefly in terms of the philosophical and religious concepts elaborated by Plato. Interpretation in Platonic concepts was also the means by which the Judeo-Christian set of creeds was thoroughly assimilated to Greek ideas by the early Christian thinkers Clement of Alexandria and Origen. Thus, the religions had a common conceptual framework. The doctrinal similarity is exemplified in the case of the pagan writer and philosopher Synesius. When the people of Cyrene wanted the most able man of the city to be their bishop, they chose him, and he was able to accept the election without sacrificing his intellectual honesty. In his pagan period he wrote hymns that closely follow the fire theology of the *Chaldean Oracles*; later he wrote hymns to Christ. The doctrine is almost identical.

The similarity of the religious vocabulary is also great. Greek life was characterized by such things as democratic institutions, seafaring, gymnasium and athletic games, theatre, and philosophy. The mystery religions adopted many expressions from these domains: they spoke of the assembly (*ekklēsia*) of the *mystai*; the voyage of life; the ship, the anchor, and the port of religion; and the wreath of the initiate; life was a stage and man the actor. The Christians took over the entire terminology; but many pagan words were strangely twisted in order to fit into the Christian world: the service of the state (*leitourgia*) became the ritual, or liturgy, of the church; the decree of the assembly and the opinions of the philosophers (*dogma*) became the fixed doctrine of Christianity; the correct opinion (*orthē doxa*) about things became orthodoxy.

Differences between mysteries and Christianity

There are also great differences between Christianity and the mysteries. Mystery religions, as a rule, can be traced back to tribal origins, Christianity to a historical person. The holy stories of the mysteries were myths; the Gospels of the New Testament, however, relate historical events. The books that the mystery communities used in Roman times cannot possibly be compared to the New Testament. The essential features of Christianity were fixed once and for all in this book; the mystery doctrines, however, always remained in a much greater state of fluidity. The theology of the mysteries was developed to a far lesser degree than the Christian theology. There are no parallels in Christianity to the sexual rites in the Dionysiac and Isiac religion, with the exception of a few aberrant Gnostic communities. The cult of rulers in the manner of the imperial mysteries was impossible in Jewish and Christian worship.

The mysteries declined quickly when the emperor Constantine raised Christianity to the status of the state religion. After a short period of toleration, the pagan religions were prohibited. The property of the pagan gods was confiscated, and the temples were destroyed. The precious metal from which Constantine's gold pieces were coined was taken from heathen temple treasuries. To show the beginning of a new era, the capital of the empire was transferred to the new Christian city of Constantinople. The centres of pagan resistance were Rome and Alexandria: in Rome the old aristocracy clung to the mysteries, and in Alexandria the pagan Neoplatonist philosophers expounded the mystery doctrines. When Julian the Apostate, Roman emperor from AD 361 to 363, tried to re-establish pagan worship, he found allies at Rome and Alexandria. After his death, the pagan opposition to Christianity continued for one more generation. The Roman aristocrats multiplied their efforts to maintain the piety of the mysteries, and the pagan philosophers tried to refine their theology by oversubtle interpretations. In 391, however, the Sarapeum at Alexandria was demolished, and in 394 the opposition of the Roman aristocracy was crushed in the battle at the Frigidus River (modern stream of Vipacco, Italy; stream of Vipava, Yugoslavia).

Only remnants of the mystery doctrines, amalgamated with Platonism, were transmitted by a few philosophers and individualists to the religious thinkers of the Byzantine Empire. The mystery religions exerted some influence on the thinkers of the Middle Ages and the philosophers of the Italian Renaissance.

BIBLIOGRAPHY

General: J.G. FRAZER, *The Golden Bough*, 3rd ed., 12 vol. (1911–15, reprinted 1963), a famous and influential treatise on ancient religion, especially on Adonis, Attis, Osiris, Dionysus, and Demeter; M.P. NILSSON, *Geschichte der griechischen Religion*, 2nd ed. (1955; Eng. trans., *A History of the Greek Religion*, 2nd ed., 1963), the best handbook on the whole subject of Greek religion.

Greek mysteries: E. ROHDE, *Psyche*, 2nd ed. (1897; Eng. trans., 1925), by the first author to understand the religion of Dionysos; L.R. FARNELL, *The Cults of the Greek States*, 5 vol. (1896–1909), a careful and conscientious collection and interpretation of the sources for Demeter and Dionysus; J.E. HARRISON, *Prolegomena to the Study of the Greek Religion*, 2nd ed. (1908), original and readable accounts about Demeter, Dionysus, and Orphics; E.R. DODDS, *The Greeks and the Irrational* (1951), a famous book, relevant especially for Dionysiac and Orphic religion.

Mysteries in the Roman Empire: F. CUMONT, *Les Religions orientales dans le paganisme Romain*, 4th ed. (1928; Eng. trans. of 2nd French ed., *Oriental Religions in Roman Paganism*, 1911, reprinted 1956), the best general account; A.D. NOCK, *Conversion* (1933), a well-written and original study of the relation of Oriental cults and Christianity; *Early Gentile Christianity and Its Hellenistic Background* (1964), a learned collection of articles; M.P. NILSSON, *The Dionysiac Mysteries of the Hellenistic and Roman Age* (1957), a clear and scholarly survey of the religion of Dionysus in later times; R. MERKELBACH, *Roman und Mysterium in der Antike* (1962), on the ancient romances as mystery texts; *Isisfeste in griechischrömischer Zeit* (1963), treats the Isis festivals and their relationship to Christian festivals.

(R.Me.)

Mysticism

Mysticism, a quest for a hidden truth or wisdom ("the treasure hidden in the centres of our souls"), in the 20th century is experiencing a renewed interest and understanding and even a mood of expectancy similar to that which had marked its role in previous eras. Such a mood stems in part from the feeling of alienation that many persons experience in the modern world. Put down as a religion of the elite, mysticism (or the mystical faculty of perceiving transcendental reality) is said by many to belong to all men, though few use it. The British author Aldous Huxley has stated that "a totally unmystical world would be a world totally blind and insane," and the Indian poet Rabindranath Tagore has noted that "Man has a feeling that he is truly represented in something which exceeds himself."

NATURE AND SIGNIFICANCE

The goal of mysticism is union with the divine or sacred. The path to that union is usually developed by following four stages: purgation (of bodily desires), purification (of the will), illumination (of the mind), and unification (of one's will or being with the divine). If "the object of man's existence is to be a Man, that is, to re-establish the harmony which originally belonged between him and the divinized state before the separation took place which disturbed the equilibrium" (*The Life and Doctrine of Paracelsus*), mysticism will always be a part of the way of return to the source of being, a way of counteracting the experience of alienation. Mysticism has always held —and parapsychology also seems to suggest—that the discovery of a nonphysical element in man's personality is of utmost significance in his quest for equilibrium in a world of apparent chaos.

Re-establishment of the harmony between man and the divine

Mysticism's apparent denial, or self-negating, is part of a psychological process or strategy that does not really deny the person. In spite of its lunatic fringe, the maturer forms of mysticism satisfy the claims of rationality, ecstasy, and righteousness.

There is obviously something nonmental, alogical, paradoxical, and unpredictable about the mystical phenomenon, but it is not, therefore, irrational or antirational or "religion without thought." Rather, as Zen (Buddhist intuitive sect) masters say, it is knowledge of the most adequate kind, only it cannot be expressed in words. If there is a mystery about mystical experience, it is something it shares with life and consciousness. Mysticism, a form of living in depth, indicates that man, a meeting ground of various levels of reality, is more than one-dimensional. Despite the interaction and correspondence between levels—"What is below is like what is above; what is above is like what is below" (*Tabula Smaragdina*, "Emerata Tablet," a work on alchemy attributed to Hermes Trismegistus)—they are not to be equated or confused. At once a praxis (technique) and a gnosis (esoteric knowledge), mysticism consists of a way or discipline.

Mysticism and religion

The relationship of the religion of faith to mysticism ("personal religion raised to the highest power") is ambiguous, a mixture of respect and misgivings. Though mysticism may be associated with religion, it need not be. The mystic often represents a type that the religious institution (*e.g.*, church) does not and cannot produce and does not know what to do with if and when one appears. As William Ralph Inge, an English theologian, commented, "institutionalism and mysticism have been uneasy bedfellows." Although mysticism has been the core of Hinduism and Buddhism, it has been little more than a minor strand—and, frequently, a disturbing element—in Judaism, Christianity, and Islām. As the 15th- to 16th-century Italian political philosopher Niccolò Machiavelli had noted of the 13th-century Christian monastic leaders St. Francis and St. Dominic, they had saved religion but destroyed the church.

The founders of religion may have been incipient or advanced mystics, but the inner compulsions of their experience have proved less amenable to dogmas, creeds, and institutional restrictions, which are bound to be outward and majority oriented. There are religions of authority and the religion of the spirit. Thus, there is a paradox: if the mystic minority is distrusted or maltreated, religious life loses its sap; on the other hand, these "peculiar people" do not easily fit into society, with the requirements of a prescriptive community composed of less sensitive seekers of safety and religious routine. Though no deeply religious person can be without a touch of mysticism, and no mystic can be, in the deepest sense, other than religious, the dialogue between mystics and conventional religionists has been far from happy. From both sides there is a constant need for restatement and revaluation, a greater tolerance, a union of free men's worship. Though it validates religion, mysticism also tends to escape the fetters of organized religion.

RELATION OF MYSTICAL EXPERIENCE TO OTHER KINDS OF EXPERIENCE

Mysticism shares a common world with magic, theurgy (power of persuading the supernatural), prayer, worship, religion, metaphysics (transcendent levels of reality), and even science. It may not be always easy to distinguish mysticism from these but its approach and emphasis are different. Though there is an element of magic, psychism, and the occult in much of what passes for mysticism, it is not to be equated with a science of the unseen or with voices and visions. Powers of the occult (or *siddhis*) are viewed as real, but they can also be dangerous and are not of interest to genuine mystics, who have warned against their likely misuse.

Prayer and worship may form part of mysticism, but they are viewed as means and not as essence; also, they are usually continuations of sensory experience, whereas mysticism is a pure unitary consciousness, or a union with God. As for science, it is analytic and discursive and expresses its findings in precise and abstract formulas; mysticism, however, like poetry, depends more on paradoxes and an unusual use of language. Philosophies may lead to or follow from mysticism, but they are not the same. Nature mysticism is another prominent variant, to which poets and artists are particularly prone. This has often been described or dismissed as pantheism (the divine in all), though it is perhaps other than a simple assertion of identity.

Emotionalism and purified emotion are quite different. Emotionalism, a kind of unsuccessful ecstasy, may arise from unpurged elements in the being; it could also be a concession or inability to hold the flow or touch from above. The natural state of man and, even more, that of the true mystic is serene and not agitated, not at the mercy of what the medieval mystical book *The Cloud of Unknowing* calls "monkey tricks of the soul." "Be still, and still, and know." Mysticism, among the many forms of experience, confirms the claims of religion and is viewed as providing a foretaste of the life after death.

DEFINITIONS OF MYSTICISM AND MYSTICAL EXPERIENCE

Differences between mysticism and similar phenomena. To define is to limit, and no single definition will cover every aspect of mysticism. Some have objected to the word itself and believed that "enlightenment" or "illumination" might be better. Though they meet, mysticism has to be distinguished from prophetic religions as well as from shamanism (a belief system built around psychic transformations). Working through chosen individuals—not necessarily saints and chosen for no other reason than God's will—prophetic religions emphasize action to a far greater extent than most forms of mysticism, with its penchant for inwardness and the beyond. Though in ecstasy the barriers seem to disappear, in prophetism God and man are rarely identified. Shamanism, a technique of ecstasy generally found in Siberia and Central Asia but with parallels in primitive society, provides a sort of correspondence with the purgative stage of mysticism (in which physical needs are negated). The closeness to paranormal (or supernatural) phenomena seems more pronounced, however, in shamanism. Both the shaman and the mystic, as communicants with a world beyond normal experience, reveal an identity of goal, if not of practice and content.

Basic patterns. Paradigmatic pronouncements in regard to mysticism pose problems of their own. The classic Indian formula—"that thou art," *tat tvam asi* (*Chāndogya Upaniṣad,* 6.9)—is hedged in with the profoundest ambiguity. The difficulty reappears in the thought of the medieval Christian mystic Meister Eckehart, who had the church raising questions for such unguarded statements as "The knower and the known are one. God and I, we are one in knowledge" and "There is no distinction between us."

Mysticism as a third kind of knowledge

Mysticism may be defined as the belief in a third kind of knowledge, the other two being sense knowledge and knowledge by inference. Adolf Lasson has written:

> The essence of Mysticism is the assertion of an intuition which transcends the temporal categories of the understanding. . . . Rationalism cannot conduct us to the essence of things; we therefore need intellectual vision.

This same view was held by the 3rd-century-AD Greek philosopher Plotinus. But the pattern misses the other dominant quality of mystical experience—love, or union through love. The medieval, theistic view of mysticism (as of religious life) was that it was "a stretching out of the soul into God through the urge of love, an experimental knowledge of God through unifying love." Its other name was joy, and the endeavour of the mystic to grasp the divine essence or ultimate reality helped him to enjoy the blessedness of actual communion with the highest. This was considered both a science and an art. As a science; *i.e.,* intuitive knowledge or the "science of ultimates," mysticism is viewed as being able to help in "the overcoming of creatureliness," and also as being able to maintain "the tendency to stress up to an extreme and exaggerated point the non-rational aspect of religion."

Reality, a kingdom of values, is viewed not as a faceless infinite, an impersonal something or somewhat. If not an ego, it is a being, and most mystics would call it God. Mysticism arises when man tries to bring the urge toward a communion with God—a "Being conceived as the supreme and ultimate reality," according to the British scholar William Ralph Inge—toward a higher consciousness and being in relation with the other contents of his mind and total personality, when he tries to realize the presence of the living God in the soul and in nature or, more generally, in the attempt to realize (in thought and feeling) the immanence of the temporal in the eternal. A 19th-century scholar, Otto Pfleiderer, indicated that religious mysticism is "the immediate feeling of unity of the self with God; it is nothing, therefore, but the fundamental feeling of religion, the religious life at its very heart and centre." Against such exclusive concentration the British writer Richard Nettleship suggests a corrective element, that of wholeness and symbolism. "Mysticism is the consciousness that everything that we experience is an element, and only an element, in fact, *i.e.* that in being what it is, it is symbolic of something else."

Introvertive mysticism. Certain forms of mysticism, however, would seem to strive toward a naked encounter with the Whole or All, without and beyond symbols. Of this kind of direct apprehension of the absolute, introvertive mysticism offers examples from different times and traditions. Instead of looking out, the gaze turns inward, toward the unchanging, the undifferentiated "One without a second." The process by which this state is attained is by a blotting out or suppression of all physical sensations—indeed, of the entire empirical content of consciousness. *Cittavṛttinirodha* ("the holding or stopping of the mind stuff") was how the 2nd-century-BC Indian mystic Patañjali described it. The model of introvertive mysticism comes from the *Māṇḍūkya Upaniṣad:*

Unitary consciousness—One without a second

> The Fourth, [aspect of self] say the wise, . . . is not the knowledge of the senses, nor is it relative knowledge, nor yet inferential knowledge. Beyond the senses, beyond the understanding, beyond all expression is The Fourth. It is pure unitary consciousness wherein [all] awareness of the world and of multiplicity is completely obliterated. It is ineffable peace. It is the supreme good. It is One without a second. It is the Self. (From *The Upanishads, Breath of the Eternal;* trans. by Swami Prabhavananda and Frederick Manchester.)

Other definitions and experiences of mysticism. Such undifferentiated unity or union between the individual and the supreme self is unacceptable to certain traditions and temperaments. The Jewish philosopher Martin Buber emphasized an "I–Thou" relationship: "All real living is meeting," and one Thou cannot become It. But even his own "unforgettable experience" of union he would explain as "illusory." With a wider range, a British scholar, R.C. Zaehner, has tried to establish different kinds, or types, of mysticism: of isolation, the separation of spirit and matter, eternity from time; pantheistic, or "pan-enhenic," in which the soul is the universe—all creaturely existence is one; the theistic, in which the soul feels identified with God; and the beatific, with its hope of deification when "the perishable puts on the imperishable."

Definitions of mysticism include a bewildering variety, ranging from the biological through the psychological to the theological. The origin of the word and certain of its features strongly suggest the possibility that mysticism is the science of a hidden life. But there is also a growing belief among 20th-century scholars that "the people of the hidden" should not remain hidden too long and should come out in the open, befitting an era of "open development" and "open realization." Some 20th-century scientists, among them physicists, biologists, and paleontologists, have shown a marked mystical bias. A biologist, Ludwig von Bertalanffy, has confessed to "peak experiences" of a great unity and liberation from ego boundary: "In moments of scientific discovery I have an intuitive insight into a grand design." He finds no necessary opposition between the rational way of thinking and intuitive experience culminating in what the mystics have tried to express. Both have their place and may coexist. Earlier there had been a sharp dichotomy between scientific and mystical knowledge. The logic of levels may never be amenable to analysis or intellectual understanding, but that is not to deny the role of reason.

Twentieth-century views of mysticism

Attitudes toward mysticism since the middle of the 20th century have been considerably modified by an awareness of subliminal consciousness, extrasensory perceptions, and, above all, an evolutionary perspective. The Roman Catholic paleontologist Teilhard de Chardin asked if in an expanding universe mysticism would not burst the limits of narrow cults and religious rigidity and move toward an ecumenical future. In a larger view, mysticism has not so much to be defined as renewed and redefined.

UNIVERSAL TYPES OF MYSTICAL EXPERIENCE

Intellectual and contemplative forms. Mystical experience, which is centred in a seeking for unity, or union, admits of wide variations but, nevertheless, falls into recognizable types: mild and extreme, extrovertive and introvertive, and theistic and nontheistic. Another well-known typology—corresponding to the faculties of thinking, willing, and feeling—employs the Indian formula, the respective ways of knowledge (*jñāna*), works (*karma*), and devotion (*bhakti*). Exclusive claims have been made on behalf of each, though maturer mystics have tried to accord to each its place and also to arrive at a synthesis, as in the *Bhagavadgītā* (Hindu sacred scripture). Depending on the powers of discrimination, the intellectual or the contemplative type tries to reach the Highest, the One, or the Godhead behind God. In its approach toward the supreme identity it tends to be chary of multiplicity, "to deny the world that it may find reality." Plotinus was "ashamed of being in the body." In the 17th century, Spinoza's nondenominational concept of intellectual love of God revealed a sense of aloofness or isolation reminiscent of the ancient Hindus.

Man, however, does not live by thought alone; to live is to work, and faith without works is dead. The mystic injunction is that works should be done in a spirit of nonattachment, with the ego sense (I, the doer) taken away. In a larger sense, not merely the doing of religious chores but all activity is offered to the Supreme. All life, according to many mystics, turns into a sacrament. "All life is *yoga* (meditation practice)."

Devotional forms. For the emotional type of person there is the mysticism of love and devotion. A theistic

attitude, or devotional mysticism, depends upon mutual attraction. In the words of a Ṣūfī poet, "I sought Him for thirty years, I thought that it was I who desired Him, but no, it was He who desired me." The path of devotion includes the rituals of prayer, worship, and adoration, which—if done with sincerity, inwardness, and understanding—can bring some of the most rewarding treasures of the religious life, including ecstasy (or *samādhi*). There is a paradox and a danger here: the paradox of avoiding the loss of personality, the danger of self-indulgence.

Ecstatic and erotic forms. Also, in an unpurified medium, the experiences may and do give rise to erotic feelings, a fact observed and duly warned against by the wiser spirits and the Fathers of the Church. (Zen Buddhism avoids both the overly personal and erotic suggestions.) Sometimes the distinction between *eros* (Greek: "erotic love") or *kāma* (Sanskrit: "sexual love") and *agapē* (Greek: "a higher love") or *prema* (Sanskrit: "higher love") can be thin. In the Indian tradition the Vaiṣnava (devotional) and Tantric (sexual) experiments were, in their apparently different ways, bold and honest attempts at sublimation, though the majority of these experiments turned out to be failures and disasters.

Use of drugs as aids to visionary experience

The same fate is likely to overtake the craze for psychedelic drugs and pharmacological aids to visionary experience—practices that are by no means new. A yogic writer, Patañjali, speaks of the use of *auṣadhi* (a medicinal herb) as a means to yogic experience, and the Vedas (Hindu scriptures) and *Tantras* (Hindu occultic writings) refer to wine as part of worship and the initiatory rites. The Greek Mysteries (religions of salvation) sometimes used sedatives and stimulants. Primarily meant to remove ethical, social, and mental inhibitions and to open up the subconscious no less than the subliminal, these techniques, as a rule, were frowned upon, even though those who took the help of such artificial aids had undergone prior training and discipline.

A whole new life-style and vocabulary have developed around medicinal mysticism in the 20th century. Peyote, mescaline, hashish, marijuana, *Cannabis indica*, LSD, and other similar products have become familiar to much of the world's population. The visions induced by such aids at best resemble the extrovertive type and cannot be easily equated with genuine mystical experience. According to taste, temperament, and tradition, the experience—a parody of creative spontaneity—may come from unexpected sources. In any case, utilizing such medicinal aids rarely achieves union with Self or God, and no permanent change of personality (in the mystical sense) has been known to occur.

GOAL OF MYSTICAL EXPERIENCE AND MYSTICISM

Experience of the divine or sacred. The goal of mysticism is "ghostly," a state or condition in which the soul is "one'd with God," according to the Western medieval work *The Cloud of Unknowing*. This "one-ing" is because all men, according to mystics, are called to their origin. Self-realization is basically one in intent with the injunctions of the Greek Mysteries: "Know thyself." This knowing, union, or communion with the divine and the sacred is of the essence of the ascent of man. As the only answer to the problem of identity, mystics look upon it as the final end, the *summum bonum*. At the journey's end waits the knowledge by identity. The direct, intuitive perception is more akin to revealed religion than to science and philosophy, though it is of itself a science, and philosophies spring from as well as lead to it.

Union with the divine or sacred. In the movement toward the goal there are, naturally, stages and processes, marked differently in different traditions. The discipline of prayer, purification, and contemplation culminates in the highest wordless union with the divine and the ultimate. As the process unfolds itself along the mystic way, an alteration of personality—a conversion, if not reintegration—occurs. The unregenerate "old man" (in Christianity) is replaced by the new being. The "twice born" (in Hinduism) becomes more than a metaphor or sacrosanct social arrangement. There is a change of level and mind. One of the aims or methods of mysticism is to make possible this change and conversion, a shift from the profane to the sacred, from "here" to "there": "Lead me from the unreal to the real, from darkness to light, from death to immortality" (*Bṛhadāraṇyaka Upaniṣad*). Before the transition, or the "great passage," is completed, however, the individual or pilgrim feels successively or simultaneously his oneness with nature, with people, and with things—an extension of awareness or expanded selfhood to which no limits can be assigned. Cosmic consciousness is thus a stage in a progressive self-discovery.

Conversion from the profane to the sacred

Experience of the universal. The nature of the goal, however, introduces a paradox. Like every other aim and activity, mysticism operates in a historical context. Yet, sooner or later, it also tends to reveal a timeless stance. The mystic is both in and out of time. The eternal now is a kind of release from the temporal order. Such a release may lead to a shift from the local to the universal, to a growing sense of unity of all experience. Though not a declared or conscious aim, this result could be looked upon as a not unworthy goal as well as a pragmatic standard.

To cure man of a provincialism of the spirit, from which more people suffer than either know or admit it, is one of the goals of a mysticism that has come of age. The true mystic is a cosmopolitan. In man's many-sided growth toward the real, a sane and mature mysticism leads to an ecumenical insight and obligation. Local colour, particulars, and uniqueness will not cease, but, in the perspective of the future and of wholeness, the universal alone will have survival value.

Experience of oneness with people. The apotheosized (divinized) field of consciousness is mysticism's ultimate goal and gift to the life of an evolving humanity. It alone is fitted to mediate between the anguish of existence and the serenity of essence, between *saṃsāra* ("cycle of birth and rebirth") and Nirvāṇa (the State of Bliss). According to an American Roman Catholic mystic, Thomas Merton, "The spiritual anguish of man has no cure but mysticism."

Though the mystic goal may seem to be tied to a transcendent reality, this does not mean a sundering of all relations and responsibilities. On the contrary, it is the guarantee of a set of altered relationships and a rehabilitation of what may be called the higher reason. Intuitions that sink into private fancy and morbidity have a short life to live. As for the mystic's "yonder," it is not spatially or posthumously remote but rather refers to a different order of reality and consciousness. The healthier forms of mysticism do not abjure action or the claims of love. It is an ancient maxim that one becomes what one loves. This is how the psychic birth repeats itself in the mystic soul, as stated, for example, by Meister Eckehart, a medieval German mystic: "It is more worthy of God that he should be born spiritually of every virgin, or of every good soul, than that he should have been born physically of Mary."

The claims of love versus mysticism

The mystic is not always amorous of the beyond, leaving an unredeemed world to its own ways. Not escape but, rather, victory is mysticism's inner urge and promise. The more sober among the mystics do not merely withdraw; they also return to the base and attempt the ancient alchemy, the transformation of men. A solitary salvation does not satisfy either head or heart.

MYSTICAL RELATIONSHIP BETWEEN MAN AND THE SACRED

Nature of the relationship. Within man is the soul of the holy, said Ralph Waldo Emerson in the 19th century. This is true of society, too. As the French sociologist Émile Durkheim saw it, the sacred is but a personified society. Mysticism, one might say, is the art and science of the holy. Theologically, it is but "the experience of the Holy Ghost, . . . the realization of the Spirit of Holiness." As the opposite of the profane and as a distinct and irreducible quality of the religious and mystical life, the sacred has always existed. It is indeed a mark of the real, and, when the German theologian Rudolf Otto isolated the sacred as a "quite distinctive category" of mystical apprehension, he had no lack of evidence. The emphasis, however, was not unanimously accepted. Some, like Inge,

thought the sacred might as well be elicited from such ultimate values as "truth, goodness, and beauty."

The "otherness" or unbridgeable gulf of the sacred

According to the respective world view, the interpretation or emphasis varies, but the universal core remains unaffected. The sacred is in its own way a coherent system, though not rational. The dualists no less than the theists insist on the unqualified and irreducible "otherness," the unbridgeable gulf, even when one speaks of union or communion. It is the distance that preserves the sacred.

Christian mystics, who often speak of "union with God," generally do not imply identity with the divine, since this might lead to heresy. The 16th-century Spanish mystic St. Teresa of Avila could write with impunity: "It is plain enough what unity is—two distinct things becoming one." But most others could not be so plain and had to use special strategy to cover up traces of possible deviation from what was permissible. Even if there had been a semblance of interpenetration between man and the divine, there could be no substantial identity. "Each of these," wrote the medieval Dutch mystic Jan van Ruysbroeck, "keeps its own nature. There is here a great distinction, for the creature never becomes God, nor does God ever become the creature." The same doctrine is preached in the Middle Ages by the mystic Heinrich Suso:

> In this merging of itself in God the spirit passes away and yet not wholly; for it receives indeed some attribute of God, but it does not become God by nature. It is still something that has been created out of nothing, and continues to be this everlastingly.

Identification of man with the divine, according to many the heart of mysticism, raises problems from other points of view as well. Pantheism, which asserts that all is God (or Nature), and God (or Nature) is all, is looked upon as a false doctrine in many religions. To John Calvin's leading question—"The Devil also must be God, substantially?"—the unsuspecting Spanish theologian and physician Michael Servetus had answered smilingly: "Do you doubt it?" The opinion cost him his life. The Hindus' *Upaniṣads*, however, insist on this identity in passage after passage. Closely looked at, this may not be simple pantheism but an identity in difference, a paradox present in even Vedānta (a Hindu monistic system). Islām has been fiercely critical of these claims of oneness and the medieval mystic Manṣūr ol-Ḥallāj had to pay with his life (922) for making the unorthodox announcement of his identity with the divine: *anā al-ḥaqq* ("I am the Creative Truth"). He was not the only one to speak in this manner. The more moderate Maḥmūd Shabestarī had reported an experience (*c.* 1320):

> In God there is no duality. In that presence "I" and "we" and "you" do not exist. "I" and "you" and "we" and "He" become one. Since in the unity there is no distinction, the Quest and the Way and the Seeker become one.

But Muslim theologians as a rule tended to dismiss those who "boasted of union with the Deity" as merely "babblers." In the Jewish tradition, it is generally considered improper and indecorous for any man to give a personal account of his own mystical experience.

The dimension of attraction and repulsion

Behind these and other interpretations, the reality of the sacred—and its persistent ambiguity—appears to be too true to be denied or ignored. Awe may or may not be the best part of man, but without it a necessary dimension is left out of the image of man, the dimension of what Otto called the *mysterium tremendum et fascinans* ("the mystery that repels and attracts"). The mystics are loath to leave this dimension out and, directly or indirectly, insist on its inclusion. The reason was suggested in the 5th–4th centuries BC by the Greek philosopher Plato, who maintained that the divine was the head and root of man. The mystic's is the eye, the third eye, with which the world beholds itself and knows itself divine. Though the vision is partial and passes away, there could be an ideal state of unbroken awareness of the Real Presence, an epiphany (manifestation).

According to the mystical point of view, the rational content of religion is not enough; it is not of the essence of religion. The sense of the holy, the mark of man's encounter with the "other," is usually invested with an ethical aura or undertone. This is how most people understand it. But this lowers its potency considerably. There is clearly an overplus, below good and evil and beyond good and evil. The numinous (spiritual) is not altogether free of the ominous. Thus, though the holy may be discussed, it cannot be well defined. It can, however, be experienced and evoked, as part of that wordless mystery that man must face—even if he is not able to explain satisfactorily—in his journey toward the real. This may happen early or late in his mystical journey, and the notion of evolution may not be applied to it uniformly.

The holy is not always and altogether a pleasant experience. Often shrouded in a fear that is more than fear, it is an inward dread and shuddering. The holy as awe-inspiring can be found in the Indian pantheon in such figures as Rudra and Kālī, the dark and wrathful faces of the divine, in which—in a collapse of finitude—majesty and unapproachability are inexplicably blended. The feeling of being consumed in the presence of the divine is a profound expression of man's relation to the holy. As for the ultimate mystical identity with the Supreme, Self, God, or the Unknowable, that also confirms the nonrational and suprarational nature of the experience in which ego, logic, and grammar are shattered alike. A frightful and traumatic adventure, not unlike the Greek Mystery rites, it can erupt at every crisis, break through an insulated universe. A clergyman cited by William Starbuck, an American psychologist of religion, spoke of having experienced

> a [silent] presence [in the night], all the more felt because it was not seen. I could not any more have doubted that *He* was there than that I was. Indeed, I felt myself to be, if possible, the less real of the two.

Diminution apart, the holy generally gives rise to a sense of energy and urgency, which may take different forms. At a higher level, the consuming fire of love reported by mystics could be an extension or refinement of the same energy, for "Love is nothing else than quenched Wrath." The "nothing else" may be an exaggeration, but such paradoxes of the religious life—*e.g.*, the unity of opposites—meet man at every turn. The void in Buddhism, like the nothing in Western mysticism, may be a numinous ideogram of the "wholly other."

Means and modes of the relationship. As means to meet the divine, some mystics have taken recourse to fasting, breath control, meditation, ecstasy, simplification, autosuggestion, and monoideism (absorption in a single idea). Rituals, in some cases, provide contact. An old method is the *via negativa* ("negative way"): "the emptier your mind, the more susceptible are you to the working of the presence." In other words, the impediments have to be removed. Among other indirect—but no less effective—means would be the shock therapy of the blood-curdling images that one notes in Tibetan iconography and symbology, which have their links with the archaic and the chthonian (infernal). On more negotiable levels, works of art—as far apart as Sung (Chinese) paintings, Gothic cathedrals, medieval temple architecture in India, the Egyptian Sphinx, music like *Missa Solemnis* (Roman Catholic solemn mass), or Sanskrit (Hindu) hymns—are accredited conductors of the numinous. Darkness, solitude, silence, and emptiness are sometimes enough for the sensitive soul, and the doors of perception open to a wider world beyond. A wide stretch of land or cranes flying against a cloudy sky were enough to throw the 19th-century Indian saint Ramakrishna into transport. But, always, it is less the object than something seen through the object, a bodiless presence, that forms the essence. Without symbols in which the holy is embodied, the experience of the holy vanishes.

The *via negativa*

Though it creates a sense of awe and exaltation, the idea of the holy also produces a mood of dependence, leading to action aimed at appeasing the deity or the powers behind the universe. At first, the policy of appeasement may have been inspired by fear and hope of reward. But, since the deity is not ultimately malevolent, it could also evolve into an idea of grace. Mystical theology, both in the East and in the West, has sometimes been divided

over the issue whether the union with the divine is the result of one's unaided effort or supernatural grace.

The various approaches to the divine

The approaches to the divine or sacred are various rather than uniform. Moving through physical, intellectual, devotional, and symbolic rituals and disciplines, it moves toward the ultimate goal: the annihilation of the self, *unio mystica* ("mystical union") in Western Christianity, *mokṣa* ("salvation") in Hinduism, Nirvāṇa (the State of Bliss) in Buddhism, and *fanā'* ("the snuffing out of self") in Islām. Though the words differ, the experiences are perhaps allied, if not the same. In a Ṣūfī (Islāmic mystical) poem the divine voice speaks exultantly:

> Annihilate yourself gloriously and joyously in Me, and in Me you shall find yourself; so long as you do not realize your nothingness, you will never reach the heights of immortality.

The description could as well be applied to the Buddhist *śūnyatā* ("void") and the self-negating of the Christian mystics.

The ultimate has been, as a rule, thought of as something "other" and apart, even if in mysticism what is sought is union or unity. Hierophany (manifestation of the holy) implies a choice and a distinction: between that which manifests the sacred and that which does not. Also, though a hierophany may represent a historic event that does not minimize its larger validity (and in any culture there may be local as well as general hierophanies), a hierarchy is not unlikely. On occasions, the sacred may manifest itself in something profane. Ideally, to a mystic, "the integrated quality of the cosmos is itself a hierophany." From this follows the possibility of consecrating the whole of life, so that by sacramental transformation, at any moment, "the flash of a trembling glance" may be inserted into the great time and project the man amphibian (having dual life) into eternity. Deification, without doubt one of the goals of the mystical life and a fundamental concept of orthodox Christendom, is part of the dialectics of the sacred. The alchemic undertone, in the man–God idea, has never wholly been extinguished. But, as part of the continuing paradox, one should also mention a resistance to the sacred. Depending on the ambivalence of the response to the sacred, which at once repels and attracts, the resistance is ultimately a flight from reality.

SEMANTICS AND SYMBOLISM IN MYSTICAL EXPERIENCE

Union of opposites. Mystical experience is flanked with a communication hazard, a "polar identity." The linguistic liberties and extravagances are part of the logical impossibility of having to describe one order of experience in terms of another. Hence, the rhetoric of mysticism is largely one of symbols and paradoxes. The most striking of the strategies, as the medieval Christian scholar Nicholas of Cusa put it, is *coincidentia oppositorum* ("union of opposites"). Since the opposites coincide without ceasing to be themselves, this also becomes an acceptable definition of God, or the nature of the Ground. God, said Heracleitus, is day and night, summer and winter, war and peace, and satiety and hunger—all opposites. A 5th- to 6th-century-AD Christian mystical writer called Dionysius the Areopagite advised people to

Role of symbols and paradoxes

> strip off all questions in order that we may attain a naked knowledge of that Unknowing and that we may begin to see the superessential Darkness which is hidden by the light that is in existent things.

This use of language or view of things is obviously not normal.

Old myths and archetypes are full of examples of such dichotomy. The Zoroastrian tradition has Ormazd (the Good Lord) and Ahriman (the Lie); the Gnostic myth speaks of Christ and Satan as brothers; and the same idea is found in the Vedas, where the *suras* ("good spirits") and *asuras* ("bad spirits") are shown to be cousins. In a different context there is the androgyne ("man-woman"), the *ardhanārīśvara* in Indian myth. As for the Hindu *jīvanmukta*, the liberated individual, he is liberated from duality. This is also part of what the Lord Kṛṣṇa (Krishna) said, when he asked the hero Arjuna to rise above the three *guṇas* ("modes"). The Tantras refer to the union of Śiva (a Hindu god) and Śakti (Śiva's consort) in one's own body and consciousness and provide appropriate practices to this end. The Chinese had their Yang and Yin (opposites), the Tibetans their Yab and Yum (opposites), and Buddhism its *saṃsāra* and Nirvāṇa as aspects of the Same. In *Prajñāpāramitā*, a Mahāyāna (northern Buddhist) text, the Illumined Ones are supposed to engage in a laughter in which all distinctions cease to exist.

Emptiness and fullness. Mystical experience permits complementary and apparently contradictory methods of expression: *via affirmativa* ("affirmative way," or fullness) as well as *via negativa* ("negative way," or emptiness). For fullness and freedom both are needed. This is because the reality affirmed contains its own opposite. In fact, the apparent negations—*neti-neti*, ("not this, not that") of the *Upaniṣads*, the *śūnyatā* ("void") of the Buddhists, or the Darkness beyond Light of Dionysius—perform a double function. They state a condition of being as well as its utter freedom from every determination. As Dionysius explains it, "While God possesses all the attributes of the universe, being the universal Cause, yet in a stricter sense He does not possess them, since He transcends them all." The "negative way," a way of turning the back upon the finite, is part of an old, positive, verified insight, at once the last freedom and, as far as many men are concerned, perhaps a lost freedom.

Symbolism of divine messengers. Experiences relating to these realities could not at any time have been common or widespread and must have come mainly through consecrated channels: yogis (Hindu meditation practitioners), gurus (Hindu teachers), prophets, mystics, saints, and spiritual masters of the inner life. This channeling through human agents has given rise to a host of divine messengers: a hierarchy of angels, intermediaries, and incarnations, singly or in succession. This manner of approaching or receiving the divine or holy is the justification of *avatāras* ("incarnations of God") and the man-God in various religions. "God was made man in order that man might be made God."

The mystical experience is a renovation of life at its root; that is, of the forgotten language of symbols and symbolism. The mystic participates in two worlds at once, the profane and the sacred. Rituals and ceremonies become the means of integration with a higher reality and consciousness. But symbols cannot be deliberately manufactured, nor do they make an arbitrary system. "Being for ever communicating its essence" is the source of their abundance, potency, and unity. Even a nontheistic mysticism, such as Buddhism, has deployed symbols freely, of which perhaps the most well-known is the formula *oṃ maṇi padme hūṃ* ("the jewel in the lotus").

Participation in two worlds at once

Symbols point beyond themselves, participate in that to which they point, open up levels of reality that are otherwise closed to man, unlock dimensions and elements of the soul that correspond to reality and cannot be produced intentionally or invented. Symbols may be inner or outer. To some, nature symbolism comes easily.

Symbolism of love and marriage. A far more risky but inescapable mode of symbolism than pantheism has been the use of the analogy of human love and marriage. Not all the mystics have been deniers or champions of repression. The soul, it may be added, is always feminine. The Christian mystics St. Bernard and St. John of the Cross, the Islāmic Ṣūfī poets, and the Hindu Dravidian and Vaiṣṇava saints could teach lovers. Not only the church but the faithful are viewed to be among Christ's brides and speak the language of love. "O that you would kiss me with the kisses of your mouth!" The speaker is the bride, thirsting for God. St. Bernard has shown that through carnal, mercenary, filial, and nuptial love the life of man moves toward the mystery of grace and union.

The hermeneutics (critical interpretation) of "the Bridegroom-Word" is that "the soul's return is her conversion to the Word, to be reformed through Him and to be conformed to Him." In the West, the roots of the tradition go back to the Song of Solomon in the Bible, not, perhaps, the best of models. The Hindu *līlās* ("love plays") of Rādhā and Kṛṣṇa have been freely misunderstood in spite of the repeated disclaimer that the events

described are not facts but symbols. The charge of immorality has been loudest against the Tantras, which had made a subtle, bold, and strict experiment in sublimation, whose inner sense may fail to be intelligible even to those who are attracted by it. That the marriage symbol should find a readier response among the brides of Christ is only to be expected. In *The Interior Castle*, St. Teresa has been fairly outspoken: "He has bound Himself to her as firmly as two human beings are joined in wedlock and will never separate Himself from her." But this was not a monopoly of nuns. The medieval theologian Richard of Saint-Victor has described as well as explained the "steep stairway of love" made up of betrothal, marriage, wedlock, and fruitfulness. In a slightly different set of symbols, St. John of the Cross states that after the soul has driven away from itself all that is contrary to the divine will, it is "transformed in God by love."

The notion of the "stairway of love"

Symbolism of the journey. Another prominent mystical symbol is the way, quest, or pilgrimage. Having lost the paradise of his soul, man, as the 16th-century physician and alchemist Paracelsus says, is a wanderer ever. A Christian monk, St. Bonaventure, has written about the mind's journey to God, and an English mystic, Walter Hilton, has described the Christian journey thus:

> Right as a true pilgrim going to Jerusalem, leaveth behind him his house and land, wife and child, and maketh himself poor and bare from all that he hath, that he may go lightly without letting: right so, if thou wilt be a ghostly pilgrim, thou shalt make thyself naked from all that thou wouldst be at Jerusalem, and at none other place but there." (From *The Ladder of Perfection.*)

According to the Ṣūfīs, the pilgrim is the perceptive or intuitional sense of man. Aided by attraction, devotion, and elevation, the journey leads, by way of many a wine shop (divine love), to the tavern (illumination), "the journey to God in God." In his *Conference of the Birds*, the 12th-century Persian Ṣūfī ʿAṭṭār refers to the seven valleys en route to the king's hidden palace: the valleys of quest, love, knowledge, detachment, unity, amazement, and, finally, annihilation. Others have gone further and spoken of "annihilation of annihilation." In the symbolic universe, denudation may be viewed as a way of fullness.

Men are called to the journey inward or upward because of a homing instinct. Eckehart put the matter simply: earth cannot escape the sky. All men are called to their origin, which implies God's need of man. A mutual attraction, the tendency toward the Divine cannot be stifled indefinitely, since it returns after every banishment. For some, paradise is not enough; it is too localized and perhaps perishable. They strain toward eternity, a leap beyond history into the incommunicable forever. A white radiance to some, to others it is "a ring of pure and bright light." The Veda speaks of the *kālahaṃsa* ("the swan of time") winging back to the sky and nest, eternity.

Essentially a way of return, *ricorso*, the final aim of mysticism is transfiguration. But

> by what alchemy shall this lead of mortality be turned into that gold of divine Being? But if they are not in their essence contraries? If they are manifestations of one Reality, identical in substance? Then indeed a divine transmutation becomes conceivable. (From Sri Aurobindo, *The Life Divine.*)

This is a clue to the Vedas, those hymns to the mystic fire and the inner sense of sacrifice, burning forever on "the altar Mind." Hence the abundance of solar and fire images: birds of fire, the fire of the sun, and the isles of fire. The symbol systems of the world religions and mysticisms are profound illuminations of the human–divine mystery. Be it the cave of the heart or the lotus of the heart, "the dwelling place of that which is the Essence of the universe," the third eye, or the eye of wisdom—the symbols all refer back to the wisdom entering the aspiring soul on its way toward progressive self-understanding. "I saw my Lord with the Eye of the Heart. I said, 'Who art Thou?' and he answered, 'thou'." Throughout the ages man, *homo symbolicus*, has been but exploring the endless miracle of being. Mystical experience is a living encyclopaedia of equations and correspondences, pointer readings that partly reveal and partly conceal.

Solar and fire symbolism

PSYCHOLOGICAL ASPECTS OF MYSTICISM

Awareness. Mysticism has been accused of passing off psychological states for metaphysical statements. But the psychological base has never been questioned seriously. It would, however, be proper to call it autology (the science of self). If the word psychology is to be retained, it must be in the original sense of the word now discarded. The contrast between the old and the new has been well expressed by the Russian philosopher Peter Ouspensky:

> Never in history has psychology stood at so low a level, lost all touch with its origin and meaning, perhaps the oldest science and, unfortunately, in its most essential features, a forgotten science, the science of his possible evolution.

Mysticism is that science in which the psychology of man mingles with the psychology of God. The major change or orientation is from the level of the profane to the sacred, an awareness of the divine in man and outside. The source and goal of such a psychology was revealed in the 18th-century Methodist leader John Wesley's dying words: "The best of all is this, that God is with us."

A mark of the mystic life is the great access of energy and enlarged awareness, so much so that the man who obtains the vision becomes, as it were, another being. Mansions of the mind, *maqām* (Arabic: "place"), and *bhūmi* (Sanskrit: "land"), open up to the gaze of the initiate, a wayfarer of the worlds. This means a renewal or conversion until one knows that the earth alone is not man's teacher. The mystic begins to draw his sustenance from supersensuous sources. He has "drunk the Infinite like a giant's wine," and a hidden bliss, knowledge, and power begin to sweep through the gates of his senses.

Role of identification. The state of energizing is facilitated by controlled attention. It is customary to fix the mind on some object or idea, some focus of contemplation. According to the Indian formula, to worship God one must become like him (*devam bhutva devam yajet*). Exercises, physical no less than mental, including methods of worship and prayer, have been developed to this end until one becomes what one contemplates. The ranges and creative aspects of the mind are part of the psychology of the mystics and one of the oldest traditions of mankind. The old Indian psychology divided consciousness into three provinces: waking state (*jāgrat*), dream state (*svapna*), and sleep state (*suṣupti*), and added a fourth (*turīya*), which is the consciousness of man's pure self-existence or being. The fourfold scale represents the degrees of the ladder of being by which man climbs back to the source, the absolute divine. The change, from "here" to "there," is not an uneventful process. There come dry periods, deviations, violent alterations, and temptations. If there are raptures and blue heavens, there are python agonies and absolute abandonments, howling deserts and "dark nights of the soul" to go through. Tears of joy, horripilation (bristling of the hair), stigmata (bodily marks or pains), and parapsychological phenomena have been known to develop.

Indian states of consciousness

The earlier phases of a naturalistic psychology had no qualms in relegating most of these experiences to the scrap heap of obsolete and archaic vanities, disorders, and morbidities—in a word, hallucination. One reason for such overall denigration was that complacent aliens to the mystical life did not care to distinguish between abnormal and supernormal phenomena. To them all were the same, at best some kind of religious sport. An American Quaker philosopher, Rufus Jones, has noted that psychology, as 20th-century man knows it, is empirical and possesses no ladder by which it can transcend the empirical order.

According to mystics, most men live in a prison, the walls thick with ego, the senses, and restricting interests. But some prisoners develop a passion to scale the walls and move toward an unwalled horizon, an adventure of ideas, if nothing more. Thus, the hypothesis that there might be cherubs and seraphs (angels of knowledge and of love) who call and guide men in the upward way is difficult to ignore. But if the distinction between love and knowledge is at all valid, the achievements of men would

seem to be the products of love, since, as Aristotle maintained, the intellect by itself moves nothing. Without "the driving and drawing that we feel in the heart," mysticism would lack power and might sink into quiescence, as has sometimes happened. To will what the Supreme wills is the supreme secret, the *primum mobile*. "Nothing burns in hell but self-will" (*Theologia Germanica,* ch. 34). The mystic approaches this knowledge and mobility even when he is compelled to withdraw from society for long or short periods. But withdrawal without return is not complete. As scientists of the psyche, the mystics insisted on the primacy of the inner factors. Modern psychoanalysis claims to have made available to man's knowledge areas of darkness beneath the conscious levels. However revealing these evidences of the ape and the tiger, psychoanalysis is debarred from understanding the superconscious, and it is viewed by mystics as being less than correct in its reading of the irrational in man. The inescapable pessimism of the psychoanalytic conclusion stands in contrast to the possibilities of self-development and sublimation to which mystics have always pointed.

Among other discoveries on the mystical way is that of ambivalence, or the alternate ways of looking at the world: temporal as against eternal. The double vision characterizes the saint whose life forms a point of intersection between time and timelessness. Mystical psychology assumes a transcendental faculty, in the hiddenness, beyond the threshold. It is committed to a breakthrough and could never have sustained itself without constant verification. In many ways a guarded secret, meant for the competent few, the experiment has hazards and could upset any but the most disciplined. The rousing of energy, the infusion of grace, and confrontation of the levels of reality create tensions and difficulties. Hence, the insistence on moderation and balance on all hands. "The higher the love, the greater the pain," a voice had consoled a 13th-century German mystic, Mechthild von Magdeburg. "Believe me, children," wrote a 14th-century German mystic, Johann Tauler, "one who would know much about these matters would often have to keep to his bed, for his bodily frame could not support this."

These upheavals of "mystic ill health" are part of a developing consciousness that has to move through and adjust to habits of inertia and resistance in the system and to an inability to support the emerging powers and their demands. A little imbalance now and then should take no one by surprise. The possibility of ranges of consciousness without thinking is one of the basic premises of yogic or mystical psychology. It constitutes a confutation of the formula of the 17th-century French philosopher René Descartes: *cogito ergo sum* ("I think; therefore, I am"). Being can exist without *cogito* (or *ratio*, "reason") in a direct awareness of things that is the function of intuition, *prajñā*.

Ranges of consciousness without thinking

SYSTEMATIC EXPOSITION OF MYSTICAL EXPERIENCE

Attempts of mystics to record the nature of their experiences. The theory or interpretation of mysticism is not mysticism. Generally, there are two sides to the theory: philosophical and practical. There may be another: confessional and justificatory. Though some mystics have been content to record what happened, others have worked out manuals of praxis (techniques), or *sādhāna*. As a rule, mystical method, experience, and exegesis cannot be sharply set apart from one another. However ineffable, raids on the inarticulate and expositions of the same have not ceased. The expositions have formed part of a particular framework of culture, tradition, and temperament. The 8th- to 9th-century-AD Indian philosopher Śaṅkara and the 16th-century Spanish mystic St. John of the Cross are not likely to talk in the same tone or accent. However universal in intent, all expositions tend to be localized.

Under present conditions, the study of comparative mysticism as well as the spirit of the age make it possible and perhaps mandatory for modern man to move toward an open and untethered mysticism, the "ocean of tomorrow." Indications of this change in attitude and emphasis are not wanting, especially in the 20th-century writings of the Indian mystic philosopher Sri Aurobindo and Teilhard de Chardin, who represent something totally new but allied. R.C. Zaehner has explained that both, though unknown to each other,

> not only accepted the theory of evolution, but enthusiastically acclaimed it, indeed were almost obsessed with it. Both were profoundly influenced by Bergson, both were deeply dissatisfied with organized religion, and both were vitally concerned not only with individual salvation or "liberation," but also with the collective salvation of mankind.

The value and meaning of mystical experience. Among the attempts to explain the value and meaning of mystical experience, a few features may be indicated. Claimed to be a guarantee for order and reconciliation, mysticism does not take away mystery from the world, nor is it essentially irrational. Though in their penchant for the beyond or God-intoxication some mystics have inclined to reject the world, the maturer variety has not divided the world of spirit and matter but has tried to mediate between spirit and matter with the help of emanations, correspondences, and a hierarchy of the real. As a giver of life, mysticism is meant to fulfill and not to destroy. Thus, it need not be world negating.

Pointing to a scale of senses and levels of mind, mysticism provides an escape from a life of uninspired existence. It magnifies man and gives him a hope and destiny to fulfill. With its abiding sense of the "more," mysticism may be called the religion of man or the religion of maturity. It offers not irrational developments or inducements but the working out of inherent potentials. Evolution, according to mystics, is not yet ended.

The mystical life is not for those who are well adjusted and other oriented. In Ramakrishna's homely phrase, at some point or other one has to "take the plunge." A change so radical calls for a kind of attention other than what most people seem prepared to give. To make it his supreme business one must have a call to holy living. He who seeks the divine must consecrate himself to God and to God only.

Problems of communication and understanding. The problem of communication, of tidings from another country, is obvious. Transvaluation of values is not easy to accept, adjust to, or express. The dialogue between mystical and other pursuits is an unsolved problem. After he had undergone a spiritual experience, the 13th-century Christian philosopher St. Thomas Aquinas is reported to have said, "I have seen that which makes all that I have written and taught look small to me. My writing days are over." This, from the author of the voluminous *Summa theologica* ("Summary of Theology"), is not without its importance.

Even if it is difficult to describe visions and dangerous to systematize, the direction in which mysticism points is clear: relational transcendence. The 20th-century crises and the mass media suggest the possibility of a mysticism brought up to date that will serve "the Creative Intention that past ages have called God." Whether it comes through symbols, systems, paradigmatic examples, or extreme situations, there will probably always be some response to the call of the real.

MYSTICISM AS A SOCIAL FACTOR

Mystical experience is no doubt solo, the experience of a singular person. But more than "a flight of the alone to the Alone," it could also be a redemption of solitude no less than of society. In the mystic experience, as Jakob Böhme said, the world is not destroyed but remade. At times a protest against heteronomy (*i.e.*, external authority and ecclesiastical machinery), mysticism has expressed itself in diverse backgrounds and flourished during dark periods of history.

Because of its other-worldly bias, the belief still persists that the solitary mystic, absorbed in a vertical relation with God or reality, owes no social responsibility. Altogether an outsider, he has deliberately undergone a civil death. This is not an ideal or wholly accurate picture. "A Mystic who is not of supreme service to the Society is not a Mystic at all" (from preface to R.D. Ranade, *Mysticism in Maharashtra*). According to Zen Buddhism, the great

The mystic as a man of action "doing nothing"

contemplative—even when "sitting quietly, doing nothing"—has been a man of action, perhaps the only kind of action that leaves no bitter residue behind. The less extravagant forms of mysticism represent attitudes and principles of charity, detachment, and dedication, which should guide the relation of the individual to the group. The mystics have fought the inner battle and won, creating themselves and their world.

Mysticism proves the individual's capacity to rise above the conditioning factors of nature, nurture, and society and to transform collective life, though this has not been generally recognized. With a hidden and potent force, mystics have tried, as best as circumstances permitted, to mend the universal ill. As in the classic resolve of the *bodhisattva* ("buddha-to-be"), they have looked forward to universal enlightenment. If the attempt by mystics to create a new order or a better society has failed, the incapacity or defection of the majority may be the reason for the failure.

"Revolution" is a word too often profaned. The change suggested is mainly, if not wholly, from without. In such contrived salvation by compulsion, the inner core is hardly touched. "But it is an eternal law that there can be no compulsion in the realm of the spirit. It is essentially a world of free creative choices" (Rufus Jones, *Some Exponents of Mystical Religion*). Mystics insist on a change of consciousness, a slower and more difficult process, and also on a scrupulous equation between ends and means. Impatience, deviations, and subterfuges in this respect can be costly, ironic, and instructive. According to mystics, the individuals who will most help the future of humanity will be those who recognize the unfinished and ultimate revolution—the evolution of consciousness—as the destiny and therefore the great need of all men as of society.

Holiness does not mean a retreat from or a rejection of the world. To be a mystic or a seer is not the same thing as being a spectator on the fence. As the Swedish secretary-general of the United Nations, Dag Hammarskjöld, proved with his life, in the modern era the road to holiness necessarily passes through the world of action. Many with a mystical frame of mind look beyond what mystics call quasi-revolutions to a great life—an entire civilization, the civilization of consciousness. The need of synthesis places its stake on the future and the All.

> The outcome of the world, the gates of the future, the entry into the super-human—these are not thrown open to a few of the privileged nor to one chosen people to the exclusion of all others. They will open only to an advance of *all together*. (From Teilhard de Chardin, *The Phenomenon of Man*.)

According to mystics, here may be the outline of a revolution whose message has reached but a few. The hope of a Kingdom of Heaven within man and a City of God without remains one of mysticism's final gifts to the life of what many mystics view as an evolving humanity.

BIBLIOGRAPHY. WILLIAM JAMES, *The Varieties of Religious Experience* (1902, many later editions), marks the beginning of an intelligent and general interest in the subject, in which the reality of the experience is established and its four characteristics (freely adapted by later writers) clearly brought out. RICHARD M. BUCKE, *Cosmic Consciousness: A Study in the Evolution of the Human Mind* (1905, many later editions), introduced two important ideas as one, ideas that would recur, with modifications, in later writings. RUFUS M. JONES, *Studies in Mystical Religion* (1908, reprinted 1970), provided a balanced and liberal attitude that emphasized its experiential quality, its value as a practical guide, and the presence of a mystical brotherhood through the centuries. EVELYN UNDERHILL, *Mysticism*, 12th ed. rev. (1957), has been a pioneer work, though its insistence on the Mystic Way has been questioned. REYNOLD A. NICHOLSON, *The Mystics of Islam* (1913, reprinted 1966), is one of the earliest studies in Ṣūfism that still holds interest. SRI AUROBINDO, *The Synthesis of Yoga* (first published, serially, 1914-21, later in book form), with much collateral illumination, explains the idea of Integral Yoga. JAMES BISSETT PRATT, *The Religious Consciousness: A Psychological Study* (1920), is critical but sympathetic. HENRI BERGSON, *Les Deux sources de la morale et de la religion*, 3rd ed. (1932; Eng. trans., *Two Sources of Morality and Religion*, 1935), forms part of a general thesis on creative evolution and, paradoxically, on the need for mysticism in an age of mechanization. GERSHOM G. SCHOLEM, *Major Trends in Jewish Mysticism*, 3rd rev. ed. (1954), clearly brings out the distinction that the concept of union is not an essential of mystical experience as understood in the Jewish tradition. ALDOUS HUXLEY, *The Perennial Philosophy* (1946), is an anthology with sophisticated, sometimes cynical, commentary and with an ascetic bias. PIERRE LECOMTE DU NOUY, *Human Destiny* (1947), attempts to mediate between science and faith and appeals to modern man to act as a responsible actor in evolution. JACQUES DE MARQUETTE, *Introduction to Comparative Mysticism* (1949), is a fair and straightforward survey in which its relevance to modern life and thought is brought out and an awareness of possibilities hinted at. R.C. ZAEHNER, *Mysticism, Sacred and Profane* (1957), beginning as a caveat against the use of drugs for transcendental experience, goes on to make critical distinctions between four types of mysticism. D.T. SUZUKI, *Mysticism: Christian and Buddhist* (1957), offers a sympathetic study of contrasts as well as some resemblances between two traditions. RUDOLF OTTO, *Das Heilige*, 9th ed. (1922; Eng. trans., *The Idea of the Holy*, 1958), is indispensable, but the "otherness" and ambiguity of the sacred, according to some mystics, seems to have been overemphasized. RADHAKAMAL MUKERJEE, *The Theory and Art of Mysticism* (1960), is an overall study, particularly good with regard to the Eastern material. WALTER T. STACE, *Mysticism and Philosophy* (1960), is balanced and analytic but singles out introvertive mysticism as more genuine and superior. SIDNEY SPENCER, *Mysticism in World Religion* (1963), is a helpful anthology with a reliable introduction to the field of comparative mysticism. PIERRE TEILHARD DE CHARDIN, *Le Phénomène humain* (1955; Eng. trans., *The Phenomenon of Man*, 1959), though its scientific accuracy has been questioned, its poetic and impassioned attempt to mediate between religious insights and a hope for man and the future has made it the object of much attention.

(S.K.G.)

Myth and Mythology

Myth is a collective term used for one kind of symbolic communication and specifically indicates one basic form of religious symbolism, as distinguished from symbolic behaviour (cult, ritual) and symbolic places or objects (such as temples and icons). Myths (in the plural) are specific accounts concerning gods or superhuman beings and extraordinary events or circumstances in a time that is altogether different from that of ordinary human experience. As with all religious symbolization, there is no attempt to prove that these unusual, transcendent, or divine events are "possible," or otherwise to justify them. For this reason, every myth presents itself as authoritative and always as an account of facts, no matter how completely different they may be from the ordinary world. The original Greek term for myth (*mythos*) denotes "word" in the sense of a decisive, final pronouncement. It differs from *logos*, the word whose validity or truth can be argued and demonstrated. Because myths present extraordinary events without trying to justify them, people have sometimes assumed that myths are simply unprovable and false stories and thus have made the word myth a synonym for fable. In the study of religion, however, the difference between myth and fable must be borne in mind.

The term mythology is used in two distinct ways. It is the study of myth, but it is also the body of myths given in a particular religious tradition.

NATURE AND SIGNIFICANCE

Myth occurs in the history of all human traditions and communities and is a basic constituent of human culture. People in every culture or tradition have developed a stock of myths in addition to other early forms of literature (such as legends and fairy tales). Because the variety of mythical themes, characters, and styles of narration is so great, it is difficult to make general statements about the nature of myths, which in their details indicate what the self-image of people in a given civilization is. The study of myths thus claims a place next to the study of language, art, philosophy, and science.

The primacy of myth

The primacy of myth stands out most clearly in certain practical situations. In the middle of the 19th century, for instance, a newly appointed British governor of New Zealand, Sir George Grey, found himself face to face with

the problem of how to come to terms with the warlike Maori. He learned their language, but that was not enough to understand what caused their military, social, and political upheaval. To understand the intention of Maori chiefs in negotiations, he found it above all necessary to study their frequent references and allusions to Maori mythology. Similarly, other government officials and Christian missionaries of the 19th and 20th centuries made extensive efforts to understand the mythologies of nations or tribes, because only such understanding made communication possible. The extensive studies of various mythologies were not only a means to an end, whether the efficient administration or the conversion of people, but they also amounted to a discovery that was then new to Europeans: myths present the model for man's behaviour. The world of myth is a continuous source of the knowledge needed for crucial problems in man's existence: war and peace, life and death, truth and falsehood, good and evil.

DEFINITIONS OF MYTH AND MYTHOLOGY

Unity of meaning and variety of description. Distinguished philosophers and folklorists represent opposite extremes in the study of mythology: the former aim at the most general statements possible and focus on myth as one general factor in human thought; the latter emphasize the variety of myths. Between the two, there lies a wide spectrum of interpretations, yet the two tendencies remain discernible in most attempts at definition, in harmony with the distinct attitudes prevailing in the study of man. On one hand, efforts are made to fathom the inner meaning of myths because of the authoritative, indeed revelatory, function they have for human existence. Many scholars, such as Mircea Eliade, a historian of religion, emphasize that myth narrates sacred history; and whether or not scholars agree with Rudolf Otto, a German theologian, that experience of the holy is a unique experience, they believe that myth is to be understood as a religious phenomenon and cannot be fully, or cannot at all be, explained in terms of nonreligious (*e.g.*, literary, psychological, sociological, or economic) categories. This tendency leads to sharp distinctions between myths and other, nonsacred forms of literature. On the other hand, there is a tendency to deal with myth in terms of a general theory of man that may be inspired biologically, psychologically, or any other way.

It is possible, however, to stress the sacred character of myth and therefore its uniqueness. It is also possible to describe and list the themes of various accounts and thus place myths and other narratives side by side. Whichever way one leans, in the delineation of myth, it is useful to venture comparisons and contrasts with other manifestations of oral tradition.

Characteristic contents of myths

Myths are accounts with an absolute authority that is implied rather than stated; they relate events and states of affairs surpassing the ordinary human world, yet basic to that world; the time in which the related events take place is altogether different from the ordinary historical time of human experience (and in most cases is unimaginably long ago); the actors in the narrative are usually gods or other extraordinary beings (such as animals, plants, the very first people, or specific great men who changed the human condition). Many other forms of literature share in one or more of the features of this definition of myth without becoming mythical.

Relation of myths to other forms. *Etiological tales.* The etiological tale seems closest to myth. In modern usage the term etiology is used to refer to the description or assignment of causes (Greek *aitia*). Accordingly, an etiological tale explains the origin of a custom, a peculiar condition or feature in nature, or the human or divine world. Many tales explain the origin of rocks or mountains. Others, such as the famous Hindu narrative that states that the blue neck of the god Śiva is the result of a poison he drank in primordial times, explain iconographic features. According to Karl Kerényi, a classicist, however, the original meaning of *aitia* (the word from which etiology derives) was not merely that of "explanatory causes" but an indication of a sort of "primeval condition" and thus, by implication, first principles or primary states that produce everything always.

Modern usage of these terms is vague, and the vagueness unnecessarily obscures the distinction between etiological tale and myth. Generalizing, it can be said that the typical etiological tale answers the question *why* such and such is the case, and the typical myth answers the question *whence* and thus indicates the *basis* of something or the validity of a custom. As a rule, the etiological tale has an amusing quality, and the etiological theme often seems to be added to a mythical narrative as an afterthought. In other words, the etiology is not the distinctive characteristic of myth. The story of the god Śiva's neck turning blue as a result of the poison he drank illustrates the transition from myth to etiological tale as well as the distinction between the two. In some presumably earlier versions, the etiological motif did not occur. The earlier versions narrated a series of creational events, among which was the episode in which Śiva drank a deadly poison (which had mysteriously turned up and threatened the whole creation) in order to save the world from the danger. The story of his neck turning blue as a result seems to have been added as an afterthought in later versions. When etiological passages occur in mythical narratives, it is these passages that are prone to the greatest change in the course of time because of the entertaining quality of etiological tales. Even if etiologies occur as features in myths, etiological tales are a distinct category and do not share in the claim to truth that typifies myth.

Fairy tales and folktales. The fairy tale tells of extraordinary beings and events and, in that respect, resembles myth, though it differs markedly in other respects. The time suggested by the fairy tale is the time of man's ordinary experience. Whereas the typical fairy tale opens with: "Once upon a time . . .," the typical myth begins with: "In the beginning. . . ." The fairy tale carries no authority, even if sometimes a moral is presented. Its outstanding quality is entertainment. Like the fairy tale, folktales can be collectively distinguished from myth. In all cases the miraculous plays an important part, but the intent remains entertainment. Thus, this whole group of narratives is quite different from myth, even if historically many characters in the fairy tale or folktale (*e.g.*, ogres, witches, wizards, werewolves, and giants) may be transformations of figures in a past mythological tradition (see also FOLKLORE; FOLK LITERATURE).

Fables. The Latin word *fabula* originally carried about the same meaning as Greek *mythos*; but, like *mythos*, it also came to be used with the meaning of fictitious or untrue story. The same emphasis exists in the English "fable." Something untrue can be referred to as a "mere fable," and fictitious and instructive stories like those of Aesop, an ancient Greek storyteller, are called fables. Genuine myths, on the contrary, are not presented as fictitious or untrue.

Resemblance of sagas to myth

Sagas and epics. Sagas are tales with a claim to truth, and in that important respect they resemble myths. The time of action, however, is a specific time in the past (not unspecified as in the fairy tale, nor an altogether different time as in the case of myth). In many primitive traditions that have no pronounced interest in history, saga-like tales are virtually indistinguishable from myths. The word *saga* (literally "things said") is Old Norse, and there are a great many examples of the form in medieval Scandinavian literature. In addition to defined periods in the past, the saga usually stresses specific locations. In this respect too, sagas resemble the myths of certain traditions (as in Tuamotuan religion in the Pacific). The epic can be seen as an extension of the saga. The time of action is presented as a specific time in the past. As in the saga, the protagonists are usually great ancestors of the race or nations, heroes, and heroines. Sometimes again as in the saga, superhuman bringers of good fortune play a role (*e.g.*, Gesar in the Tibetan epic; Kṛṣṇa in the *Mahābhārata*, a Hindu scripture). The epic as such, however, is not of a mythical character. Instead, it presupposes the reality of myths and a knowledge of myth among the listeners.

Legends. The notion of time in legends is like that in sagas. The protagonists, unlike those in the sagas, are

as a rule great religious figures. The legend, however, carries no authority comparable to that of myth. Rather, the legend illustrates a truth that is known from another (usually mythological) source.

It is to be noted that this list of literary types is a list that conforms by and large to Western scholarly views. Other traditions, whether of "primitive" or "advanced" civilizations, have their own classifications. In most of these other traditions, however, a basic distinction is somehow made between "true" and "fictitious," whereby the "true" stories correspond to what Westerners would understand as genuine myths.

FUNCTIONS OF MYTH AND MYTHOLOGY

Explanatory and narrative. The first function of myths that strikes the outside observer of any tradition is that of explanation. Natural, social, cultural, and biological facts are explained by myth. This is not to say, however, that myths become identical with etiological tales. A North American Indian (Abnaki [Wabanaki]) myth, for example, explains the origin of corn: a lonesome man meets a beautiful woman with long, light hair; she promises to remain with him if he does what she tells him; she tells him in detail how to make a fire and drag her over the burned ground; as a result of these acts, he will see her silken hair (viz., of the cornstalk) appear and have corn seeds for his use. Henceforth, whenever the Indians see her hair, they know that she has not forgotten them. Obviously, a myth such as this one functions as an explanation, but, just as obviously, the narrative form distinguishes it from a straightforward answer to an intellectual question about causes. The function of explanation and the narrative form go together, and myths have thus been significant in many traditional systems of education.

Justifying or validating. A great variety of myths answer questions about the nature and foundation of ritual and cultic customs. According to myths from the island of Ceram (Indonesia), in the beginning human life was not complete, or not yet "human": vegetation and animals did not exist, and there was neither death nor sexuality. In a mysterious manner Hainuwele, a girl with extraordinary gift-bestowing powers, appears. The people kill her at the end of their great annual celebration, and, as a result of this first killing, life as man knows it comes into being. With a certain circularity that often occurs in mythology, the myth validates the cultic celebration that itself occurs in the myth. The mythical account has an optimistic tone, since its events make life "livable." The cult can be understood as a commemoration of those first events. Hence, the myth can be said to validate life itself together with the cultic celebration. Comparable myths are told in a number of root-cropper societies.

Cultic or ritual validation

Bloody sacrifices are typical of traditional peasant cultures. In most cases such customs are related to mythical events. Among important themes validating the cultic proceedings are the necessity of death (*e.g.*, the grain "dies" to yield a harvest), the renewal of time, and the significance of woman and sexuality. New Year's celebrations, often accompanied by a temporary abandonment of all rules, often are related to or justified by mythical themes concerning a return to chaos and a return of the dead.

Dynasties and ruling families in several ancient civilizations found justification of their positions in myths, which state that they originated in the world of the gods or in heaven or from the sun or the moon (as in China, Egypt, Babylon, the Hittite Empire, Polynesia, the Inca Empire, and India). Social classes also can be validated by myths; the best known are the mythologically established classes of ancient Indo-European traditions: the sacerdotal, warrior, and producing classes. No cultural tradition exists without some mythological foundation for marriage and for funerary customs. Also, many myths account for specific groups of a social or socioreligious nature, such as medicine men, shamans (men having curative and psychic transformation powers), and priests.

Other functions. *Descriptive.* The descriptive function of myth is linked with the authoritative presentation of facts that transcend ordinary reason and observation. Myths can describe the origin of the world, the end of the world, or a paradisiacal state. Thus, a myth is capable of describing what persons, using reason and observation, can never "see for themselves." Perhaps even more than the explanations that are provided, this descriptive function accounts for the educational value of myths in traditional societies. Myths set the pattern for theoretical as well as practical instruction. In primitive traditions they form the only available model of instruction, in the absence of any separate philosophical enquiry. The descriptive function of myth expresses a perennial human need and is a reminder that modern men do not have to regard myth as something that comes to an end with the termination of primitive or classical civilizations.

Controlling the universe. In every mythological tradition one myth or cluster of myths tends to be central. In comparison to these, other myths—whether about the first shaman, the establishment of a secret society, or anything else—relate more to a specific sector of society or more specialized functions. The central mythology as a rule is primarily occupied with the creation that is presupposed in all other creations; *i.e.*, cosmogony (origin of the cosmos). In many events, on which the well-being of man's world depends, references are made to the beginnings of the world. Among such events are enthronements of kings, which in some traditions (as in Fiji and in ancient India) were associated with a creation or re-creation of the world. Such associations are not to be viewed merely as products of power-hungry monarchies. Creation myths also play a significant role in healing the sick; they are recited (*e.g.*, among the Navajo Indians of North America) when an individual man's world—his individual life—is in jeopardy.

Magical and poetic-artistic functions. Healings through the recitation of the cosmogony, referred to above, can be examples of the magical function. Equally "magical" is the example of an old Norse poet, who was to renew his poetic inspiration by recalling the myth of the god Odin's acquisition of poetic power.

The poetic–artistic function of myths in archaic and primitive traditions cannot be overestimated. In civilizations in which great specialization and separation of art forms have not yet occurred, mythical themes and images are a "given" source of all expression. The creative force of well-developed mythologies also influenced later cultural and artistic developments.

THE STUDY OF MYTH AND MYTHOLOGY

A determined study of myth developed in the West, especially since the 18th century. Much of its material came from the study of the classics, from which it also derived some of its methods of interpretation.

Brief history of the study. The growth of a philosophical interest in ancient Greece furthered allegorical interpretations of myth; *i.e.*, finding other or supposedly deeper meanings hidden below the surface of mythical texts. Such meanings usually involved phenomena of nature or human morals. Related to this was a tendency toward rationalism or pragmatism, especially when interpreters, aided by faulty etymologies, took the opportunity to stress the value of the ideas or the civic sense of their own time. Of special and long-lasting influence was Euhemerism (named after Euhemerus, a Greek writer who flourished about 300 BC), according to which certain gods were originally great people venerated because of their benefacts to mankind.

Influence of Euhemerism

The early Church Fathers continued with concepts that were general in late antiquity, particularly a modified Euhemerism, according to which mere men were raised to superhuman, demonic status because of their great evil deeds. The Middle Ages did not add greatly to the study of myth, and, despite some elaborate works of historical and etymological erudition (which did not stand up to modern criticism), neither did the Renaissance.

Early in 18th-century Italy, Giambattista Vico, a philosopher, building on traditional scholarship—especially in law and philosophy—was the first to make a clear case for the role of man's creative imagination in the forma-

tion of distinct myths at successive cultural stages, but his work had no influence in his century. The notion that pagan myths were distortions of the biblical revelation in paradise was first expressed in the Renaissance and continued to find favour in the 18th century, alongside other earlier theories. At the same time, Enlightenment philosophy, reports from voyages of discovery, and missionary reports (especially the reports of Jesuits on North American Indians) contributed to 18th-century scholarship and fostered a desire for objectivity. Bernhard Le Bovier de Fontenelle, a French scholar, compared Greek and American myths and theorized on a universal human predisposition toward mythology. Around 1800, the growing fascination of the Romantics for human speech, the postulation of an Indo-European language family, the study of Sanskrit, and the growth of comparative studies, especially in history and philology, all gave a firm footing to the study of myth. The interest in myth was greatly stimulated in Germany by Friedrich Schelling's philosophy of mythology. Scholarly interest in myth declined during the heyday of evolutionism but re-emerged around 1900 among ethnologists as well as some philologists. Many depth-psychological studies followed in the wake of the discoveries of Sigmund Freud. A renewed interest in myth was fostered by new approaches in sociology and anthropology and by various philosophical studies. In the history of religions, Eliade, for example, has stressed that myth is best understood as narrating sacred history and cannot be explained in terms of nonreligious (literary, psychological, sociological, or economic) categories. Among some classicists (*e.g.*, Karl Kerényi) and others, a conscious resumption of Romantic ideas is noticeable (see also RELIGION, STUDY OF).

Basic modes of approach. *Rationalistic.* Many types of rationalistic explanation of myth have found adherents since antiquity. They have in common the fact that they posit a definable system of thought as the cause of mythological texts. Allegorical interpretations are examples of this, for it is typical of allegory that it can be rendered adequately in terms of a clear idea, though myths as a rule do not really present such ideas. Also since antiquity, many (*e.g.*, David Hume, an English philosopher) have attributed the creation of myths to fear. This too can usually be regarded as a rationalistic interpretation, if it posits a known system of psychological functions.

Romantic. Georg Johann Hamann, a German philosopher, and William Blake, the English mystic and writer, both rejected as a rationalistic construct the notion of "natural religion," a type of religion supposedly common to all mankind. Schelling was convinced that the Absolute is expressed in myth but that this expression also occurs in terms of certain contradictions and that it is the task of philosophy to regain the transcendent unity of the Absolute. Schelling's metaphysical concern has been a stimulus for subsequent discussions. Many Romantic studies were less visionary and have been Romantic only in the sense of stressing emotion as a factor in the creation of myths, sometimes bordering on sentimentality. Also Romantic is the long-lingering explanation of myth as an expression of man's experience of nature.

Rejection of "natural religion"

Comparative. Since the Romantic movement, all study of myth has been comparative. This is not to say that a comparative study as such is Romantic, for comparative attempts were made earlier. But the fact remains that in the 19th century and later, even the most specialized studies have made generalizations about more than one tradition and at the very least have had to take comparative works by others into account. Philosophical enquiry aiming at an understanding of the function of myth assumes the possibility of a comparative approach.

Folkloristic. The classical folkloristic approach is that of Wilhelm Mannhardt, a German scholar, who attempted to collect data on the "lower mythology," which he considered more or less homogeneous in ancient and popular peasant traditions and basic to all formation of myth. Sir James Frazer, an English anthropologist, and the brothers Grimm (Jacob and Wilhelm), German philologists, are well known for their collections of folklore, and the American folklorist Stith Thompson for his classification of folk literature. Collecting and classifying mythological themes have remained the principal contribution of folklore.

Functional. Marcel Mauss, a French sociologist, spoke of "total social facts" in reference to religious symbols and myths and their irreducibility in terms of other functions. Bronisław Malinowski, an anthropologist, changed the emphasis and attributed the necessity of myth to the functions it fulfilled in primitive cultures: justifying and preserving social, ethical, pragmatic, aesthetic, and religious customs and institutions. Functionalism, especially as an interpretation of myth as a mirror of social relationships, has had a wide appeal to anthropologists in Britain and America.

Structural. Structural approaches begin with the observation that myths throughout the world resemble each other and proceed to define the resemblance. Some structuralists (*e.g.*, anthropologist Claude Lévi-Strauss, in France) look for the resemblance in a certain logical coherency in analogy to patterns in linguistics. Others emphasize the total view of the world manifest in all aspects of a community's life.

MYTH IN HUMAN CULTURE

Myth and psychology. The paramount importance of Freud's investigations for the study of myth is his view that the formation of concepts does not depend on cultural history. Until Freud, all study of myth had assumed that historical analyses and an understanding of stages of development of human consciousness were absolutely necessary. Freud's analysis of the psyche created an independent, transhistorical mechanism, based not on cultural history but on a biological conception of man. The fact that he made use of anthropological views that have since been refuted (*e.g.*, totemic [symbolic animal] sacrifice as the earliest ritual custom, which he related to the first parricide) does not change this. Symbolizations of guilt feeling or symbolic expressions of wish fulfillment are not limited to specific times in history and occur in myths as in dreams. Criticism of Freudian methods has been levelled against the explanation of myths in terms of only one theme (the Oedipus complex; a psychosexual concept concerned with guilt) and in terms of the "repression" of conscious ideas. In Switzerland, the psychoanalyst Carl Jung, like Freud, was stimulated by a theory that no longer has much support; *i.e.*, the theory of Lucien Lévy-Bruhl, a French philosopher, associating myth with prelogical mentality. This, according to Lévy-Bruhl, was a type of thought that supposedly would have been common to archaic mankind, that was still common to primitives, and in which people supposedly experienced a "mystical participation" with the objects of thought, rather than a separation of subject and object. Jung's theory of the "collective unconscious," which bears resemblance to Lévy-Bruhl's theory, enabled him to regard the foundation of mythical images as positive and creative, in contrast with Freud's negative evaluation. For Jung, archetypes are the specific images that arise from the collective unconscious and form the basis of the human psyche. In principle, however, these images that make up myths are, like the Freudian mechanism, transhistorical. Critics of Jung have been hesitant to accept his neo-romantic reasonings concerning the collective unconscious as an eternal source of meaning. The criticism has come not only from Freudians but also from critical students of Jung himself (see also CONCEPT FORMATION).

Freud's creation of a transhistorical method

Myth and society. Myth is obviously related to questions about language, because it is a matter of human communication, and to questions of man's social life, because the narration of myths is proper to a community and to communal tradition. There is no general agreement, however, on the precise nature of these relationships. The reason for the uncertainty is not difficult to detect. The methods used by modern science allow for empirical investigation and inference, but the classification and segmentations of language and society that can thus be established leave an essential part of the subject untouched. Some linguists, reflecting on their own discipline, have spoken explicitly of the need for a more en-

compassing science; *e.g.*, a new general science of semeiology, whose task it should be to study all signs central to social life; and a new science of psychology, which would begin with "the problem of characterizing various systems of human knowledge and belief." The study of society and of language can only begin with what is given in human speech and social relations, but in each case it is confronted by a coherence of traditions that is not directly open to research. This is the area in which mythology functions.

Intricacy and variety of relationships

The following mythological conceptions exemplify the intricacy and variety of the relationships between myth and society. The Lugbara (a tribe in northwestern Uganda and the adjoining Congo region) make use of a certain conceptual system to relate their sociopolitical order to two hero ancestors, who in turn are related to the creation of the universe. The narrations about their descent from their hero ancestors are saga-like, though the earlier "history" is presented in myths. It is remarkable, however, that the one conceptual scheme of the Lugbara social system relates the mythical and the genealogical (nonmythical) past and in its entirety is expressed in spatial rather than temporal–historical categories. Ancient China does not present an extensive system of myths and in this respect differs markedly from postclassical China, with its vast numbers of myths and gods. Instead, in ancient China, literary references are made to figures and developments that seem historical but are in fact historicizations of earlier myths whose mythical form was abandoned. Many of these references, especially in words attributed to Confucius, have to do with the nature of the state and the quality of rulers. Similarly, the Roman historian Livy's history of Rome presents the origin, the governmental forms, and the social forms of the city in terms of a historicization of earlier (Indo-European) mythological materials. The Indian institutions of class and caste are embedded in myth, and precisely for that reason they present a strong force in the empirical reality of Indian society. The variety of mythological influences on societies does not facilitate an overall system of interpretation, but an understanding of myth in each instance has a bearing on an understanding of social relations. Also, a particular conception of history is often taken for granted in modern studies of societies and social change, perhaps too uncritically. Indeed, some scholars have argued in various ways that the appeal made to history by nations or social groups is itself mythological by nature.

Myth and science. Some students have drawn attention to the changes occurring in the "reality" reflected and assumed in myths. Such "reality" is never final but changes continually in history. Questions about the changes of "reality" have especially occupied philosophers and historians of science, for an image of reality is basic to any scientific pursuit, beginning with the earliest philosophical enquiries into the nature of the world. Though it is not necessary to identify the images of reality in science with myth, the parallels of science and myth and the mythological dimension of science as such should be noted.

Scientific models as paradigms

Parallels. The function of models in physics, biology, medicine, and other sciences resembles that of myth as paradigms (patterns) for the human world. In medicine, looking upon the human body as a machine and viewing the human brain as a computer are models that are easy to understand as examples. Once a model has gained acceptance, it is difficult to replace, and in this respect it resembles myth, while at the same time, just as in myth, there may be a great variety of interpretations. In the period after René Descartes, 17th century, it was simply assumed that the universe could be explained in terms of minute corpuscles, their motion and interaction, and that entities of another sort did not exist. The absolutistic bent of many models in the history of science most resembles myth. There are, however, important differences. Despite the difficulty of replacing models in science, replacement does occur, and an awareness of the limitations of models has grown strong in modern science. This awareness of scientists makes the difference evident, for, as a rule, a myth cannot be regarded as replaceable by the community in which it functions, although an outside observer might record changes and even replacements. Moreover, models in science have their value only for specific scientists. Hence, they function only for a small segment of society, even though, for instance, a medical theory held in academic circles in one century can filter down into folk medicine in the next. As a rule, myth has a much wider impact.

Mythological dimensions of science. In spite of the distance between scientific models and myths, there are reasons to speak of mythological dimensions and components of science. The history of science has made clear that modern science did not come about in its entirety as a rebellion against myth, nor at its birth did it suddenly shake off the shackles of myth. In ancient Greece, the naturalists of Miletus, long regarded as the originators of science, developed views of the universe that were in fact very close to the creation myths of their time. Precursors of modern science, such as Nicholas of Cusa, Johannes Kepler, Sir Isaac Newton, and Gottfried Leibniz, were involved in metaphysical problems of which the traditional, indeed mythological, character is evident. Among these problems were the nature of infinity and the omnipotence of God. Clear remnants of mythological views are seen in the English physician William Harvey's association of the circulation of the blood with the planetary movements and Darwin's explanation of woman's sexual cycles by the tides of the ocean.

Several thinkers (*e.g.*, Paul Tillich, a theologian, and the philosophers Karl Jaspers and G. Gusdorf) have argued convincingly that a mythological dimension is constitutive of all science. Myth is that which is taken for granted when thought begins. It is at the same time the limit science reaches in its own course: certain elementary ideas or principles are found to be of such lucidity that no further definition can be undertaken. In recent scientific researches, especially in astronomy and biology, questions of teleology (final ends) have gained in importance, as distinct from earlier concerns with questions of origin. These recent concerns stimulate discussion about the limits of what can be scientifically explained, and they reveal anew a mythological dimension to human knowledge. It should be underlined that there are such mythological dimensions to science in the same way that there are to society. They do not make the study of science (as of society) itself a mythological affair.

Myth and religion. The place of myth in various traditions differs; and within some religious traditions there is opposition.

Relation to ritual and other practices. The idea that the principal function of a myth is to explain a ritual had adherents before anyone tried to make a case for it. At the beginning of the 20th century, many scholars thought of myths in their earliest forms as statements of what is being done and what is desirable, and they found support in what Frazer called "sympathetic magic." According to this view, the narration of myth developed to explain and accompany magical procedures that were first meant to fulfill human desires and later grew into true religious cult.

The Myth-Ritual school

In a much more articulate way, biblical scholars stressed the necessity to look for the situation in life and custom (the "Sitz im Leben") that mythical texts originally possessed. A number of scholars, mainly in Britain and the Scandinavian countries and usually referred to as the Myth-Ritual school, have concentrated on the ritual purposes of myths. Their work has not extended far beyond the philological study of the ancient Near East and Islām and has focussed almost exclusively on rituals connected with sacred kingship and New Year's celebrations.

The existence of connections between myths and cultic behaviour is certain, but there is no solid ground for the suggestion that, in general, ritual is chronologically first and myth an added explanation. In the first place, it is not correct to regard genuine myths as etiologies. Secondly, if it were only the myth (later) that made the sense of the ritual (earlier) explicit, the meaning of the ritual remains a riddle. Endeavours to solve this problem have not met

with general approval. Evolutionists since Frazer have largely thought of early man's behaviour as something primitive and irrational and have posited slowly growing intellectual powers that demanded explanations. Other attempts, betraying influences of Existentialist philosophy, have seen earliest myth as something not in need of words because it was "alive and lived."

The view that has seemed to many to be closest to the truth begins with the recognition that there is no universal law governing the relationship between myth and ritual. An absolute chronology, establishing which of the two came first, is outside the reach of scholarship. It is quite certain that there is no ritual without myth, but there are many myths without ritual. Most often the content of important myths concerning the origin of the world reflects the dominant cultural form of a tradition. Hunters' societies have myths that tell of the origin of game animals and hunting customs, agricultural civilizations tend to give weight to agricultural practices in their myths, pastoral cultures to pastoral practices, and so on. Thus, many myths present models of necessary acts and organizations and relate these to primordial times. Myths in specific archaic traditions deal with harvest customs, initiation ceremonies, circumcision and subincision (cutting open the urethra on the underside of the penis), the customs of secret societies, the use of ceremonial masks, cannibalistic and headhunting customs, or special skills and crafts.

Relation to religious symbolism and iconography. Sacred objects are found in all religious traditions, and sacred images (icons) in most. They are analogous to the myths because they represent sacred realities or figures, as myths do in narrative form. Representing does not mean copying of natural or human forms. In this respect religious symbolism is immediately related to myths because they also narrate extraordinary things. Many symbolic representations have their sources in myths. Representations in human form or perfectly "natural" human forms are rare. The famous sculptures of divine figures in classical Greece (as by Phidias and Praxiteles) are the exception, and they were exceptional even in the religion of Greece. Everywhere, less naturalistic forms have always been preferred in cultic practices and in the representation of mythical themes. Geometric and animal figures—*e.g.*, on cultic vessels as well as on images of deities—abound in the history of religions. Of special significance is the use of sacred masks: in the mysteries of Dionysus, a Hellenistic salvation cult, and the indigenous traditions of Australia, America, prehistoric Europe, and elsewhere (see also SYMBOLISM AND ICONOGRAPHY, RELIGIOUS).

Relation to sacred texts. The relation of myth to the sacred texts of the great religions naturally depends on the character of the religion in question. The Old Testament contains much mythological material, yet the religion of Israel was in many respects critical of myths in the same manner in which it rejected any representation of God in natural forms. Similar antimythological tendencies exist in the religions that have their roots in Israel. The New Testament of Christianity in some instances derogates myths by describing them as "godless" and "silly." Islām's emphasis on the transcendence of God, attested in the Qur'ān, is strongly antimythological. Many scholars have pointed out, however, that all these antimythological tendencies are correlated to mythological motifs, especially motifs concerning the end of time (eschatology). In other traditions with sacred scriptures, such as Hinduism and Buddhism, opposition to myths is not dominant.

Biblical and Qur'ānic criticism of myth

Myth and the arts. *Oral traditions and written literature.* Myths in ancient civilizations are known only by virtue of the fact that they became part of a written tradition. In the case of Greece, virtually all myths are "literature" in the form in which they survived, the oldest source being the epics of Homer, a Greek poet of the 8th century BC. Though it is difficult to speak in detail about the original function of many myths when they were alive in oral tradition, scholars are in a better position to understand the importance that myths had for literary artists and the mythological function that literature inherited. The heroes and heroines of epic literature are in many instances narrative rejuvenations of gods and goddesses in myths, and, although scholarly work in this area is far from complete, it is safe to suggest that among the mythical themes that have appealed to epic poets are themes of the culture hero or bringer of salvation and of shamanism. The epic itself can take on certain mythological functions. The Homeric epic became the basis of education in classical Greece. The great epics of India (*Mahābhārata* and *Rāmāyaṇa*) came to function as encyclopaedias of knowledge and provided models for all human existence. Generally, whether literature makes use of mythical motifs or not, the power of the literary word, the sheer felicitous formulation of political, aesthetic, or other certainties, has a spell that is akin to that of myth.

Visual arts. In principle, the sort of relationship that exists between myth and literature exists also with respect to the other arts. In the case of architecture and sculpture, documentation begins much earlier, and it is difficult to imagine their beginnings without myth. Moreover, data owed to archaeological discoveries confirm the primacy of mythical representations. Among the earliest known architectural occupations of man are prehistoric megalithic and sepulchral structures. Mythological details cannot be recovered, but mythical images most likely dictated the shape. Grave monuments of rulers are among the most famous remains of several ancient historical civilizations as well (*e.g.*, the Egyptian pyramids; and the sepulchral structures of Chinese rulers since the Chou period, *c.* 1122–221 BC). There is worldwide evidence in archaic cultures of a mythological orientation in which man established his place in the cosmos with respect to north, south, east, and west, as well as to the vault of heaven and the world below; and such mythological views formed the model of his building activity. As late as the 1st century AD, the Roman architect Vitruvius, summarizing the building lore of the classical world, spoke of the different orders of building with explicit reference to temples and the specific gods and goddesses for whom they were appropriate. Not by coincidence did Vitruvius consider the pillar the principal element, and not by coincidence is the history of architecture linked with temples. The pillar is the archaic mythological image of what connects man's world with heaven. This imagery already existed in hunters' cultures, but it continued in the Neolithic period with a growing interest in firm dwelling places for gods and people. The strict separation between religious and civil architecture that modern man is inclined to take for granted is not applicable for most cultures and periods, and perhaps it is not altogether applicable even in modern times. A certain general mythology, equally normative for civil and religious buildings, is visible in ruling architectural fashions. Classicism, for example, can be seen as a mythology that set a model for building city halls and courthouses, as well as churches. Classical and medieval India has produced a vast literature on architecture and art, setting side by side instructions for temples and houses and complete indications for town planning.

Mythological aspects of architecture

When art ceases to represent mythological matters or no longer produces vessels, instruments, or images used for cultic purposes, it still represents something different from what is naturally given. This fact can be considered in two distinct perspectives. First, it is possible to insist on the universal mythological dimension of man and to look upon what seems to be abandonment of mythological forms as something that in fact is only a change in mythology. Secondly, one can insist on the actual abandonment of myth. These two possibilities are analogous to the positive and negative attitudes that religions have developed toward myth. Factual study of myths and of artifacts by themselves does not suffice to decide the matter, because the choice depends heavily on one's theory of art.

Performing arts. Myth is one of the principal roots of drama. Even in the very "literary," refined tragedies of classical Greece this is clear, not only because of the many mythological subjects treated and the setting of the

Myth as a source of the drama and the dance

performances (organized by the city-state at the festival of Dionysus) but especially because of playwrights', such as Aeschylus', mythlike presentation of events and facts, which may be interpreted in various ways like myth, "at once moral, religious, and purely human," and relating to the community as a whole. The citizens made up the chorus, the crucial part of early Greek tragedy. Theodor Gaster, an American historian of religion, has shown in detail that in the ancient Near East the interrelationship of myth and ritual created drama. Elsewhere, dramatic presentations (as in Japanese Nō plays and the Javanese *wayang*) are similarly rooted in myth. Western distinctions between "religious" and "entertaining" have a very limited value here.

Dance has even earlier rights than the other performing arts in experiencing and expressing mythological characters, and it is a universal form. Dance as an expression of myth, though attested universally, is also a form that is most difficult to comprehend for many people today, because the decay of ritual behaviour typifies modern secularization. In spite of a continuing interest in ballet and some revival of folk dancing, dance as a form of art has perhaps become further removed from myth than any other form of art in the Western world.

Music. Music in its primary rhythmic function is related to dance and thus to myth. Music nevertheless is least dependent on myth but is regarded in many traditions as creative, and hence mythical, in itself. It sometimes is used (as myth is) in healing the sick (*e.g.*, tarantism, a form of hysteria appearing in Italy in the 15th–17th centuries and reportedly cured by frenzied dancing). Although most of such religious purposes of music in modern Western history have declined, mythical themes (*e.g.*, in opera and oratorio) have not ceased.

Mass media. The modern means of radio, film, and television have had an effect on religious symbolism, but equally striking is the fact that—like the printing press, newspapers, magazines, and pocket-books—they have provided channels for old mythological themes. Communication as such has become much more important than it was before, because the reader, listener, or viewer is much more passive than the participant in earlier ceremonies. Inevitably, the mass media try to increase their principal power, viz., their power of suggestion. This development has led to a general increase in receptivity for myth, to the exclusion of ritual. Among the mythological types that are most appealing are those that most resemble the heroes of ancient and classical myths (sports heroes, movie stars, astronauts, and other celebrities). Whatever the details or the order of popularity may be, the modern audience, like all earlier audiences, looks for models by which all things can be measured.

Modern mythological models

MAJOR TYPES OF MYTH

Myths of origin. "Cosmogony" and "creation" are used as synonyms, yet properly speaking, cosmogony is the best general term because it refers to the origin of the world, whereas "creation myth" implies a creator and something created, a suggestion that does not fit a number of myths that, for example, speak of the origin of the world as a growth or emanation, rather than an act. Furthermore, the term origin should be used with some caution for cosmogonic events (as well as for other myths of beginnings), because the origin of the world hardly ever seems the focal point of a mythological narrative—which is not a matter of enquiry into the first cause of things. Instead, cosmogonic myths are concerned with origins in the sense of the foundation (or validity or "ground") of the world as it is. Creation stories in both primitive and advanced cultures frequently speak of the act of creation as a fashioning of the earth that was already present before the act. In African cosmogonies, especially, the earth is pre-existent. A creation out of nothing (*creatio ex nihilo*) occurs as a theme only in certain more philosophical traditions. Water, though important everywhere as source of life and image of endless "potentiality," has a special role in Asia and North America, where the creator (often an animal) is assisted by another figure, who dives for earth in the primordial ocean. The earth-diver helper sometimes develops into an opponent, or Satan-like character, in culturally advanced areas; *e.g.*, those touched by Zoroastrianism, an Iranian religion that was dualistic. Though hardly an "explanation" in the ordinary sense of the word, the theme accounts for the fact that evil is constitutive of the cosmos without holding the creator responsible for it. Other widely diffused motifs are: the cosmogonic egg, found in the Pacific world, parts of Europe and southern Asia, and in very advanced cultures (*e.g.*, Hinduism); the world parents (usually in the image of sky and earth); and creation through sacrifice or through a primordial battle. Creation through the word of the creator also occurs outside the Old Testament account (in Polynesia).

Cosmogony sets the pattern for everything else in most traditions; other myths relate to it or derive from it. Because man's inhabitable world, the cosmos, is the crucial issue, no matter how variegated the contents may be and how different from one period to another, cosmogony probably is the clearest expression of man's basic mythological propensity. All cosmogonic accounts have certain formal features in common. They speak of irreconcilable opposites (*e.g.*, heaven and earth, darkness and light) and, at the same time, of events or things totally outside the common range of perception and reason (*e.g.*, a "time" in which heaven and earth were not yet separated, and darkness and light intermingled). In other words, the basic ingredients of man's world and orientation are presupposed and yet nevertheless are realized, constituted, or brought about anew in the narration. The narrative can arrive at such a reconstitution only by transcending the limits of ordinary perception and reason.

Common features in cosmogonic and creation myths

Because the cosmos is man's world, the origin of man is immediately linked to the cosmogony. Man, for instance, is placed on the earth by God, or in some other way his origin is from heaven. Nevertheless, it is only in mythologies that have been influenced by philosophical reflections that the place of man becomes the conspicuous centre of the cosmogony (*e.g.*, Pythagoreanism, a Greek mystical philosophical system; Orphism, a Greek mystical religious movement; Gnosticism, a heretical Christian dualistic and esoteric movement; and Tantrism, a Hindu and Buddhist esoteric meditation system). Man is sometimes said to have ascended from the depths of the earth (as in Zuni, an American Indian people, mythology) or from a certain rock or tree of cultic significance. These images are often related to the idea of a realm of the ancestors as the origin of newborn children. Man is also said to be fashioned from the dust of the ground (as in Genesis) or from a mixture of clay and blood (as in the Babylonian creation myth). In all cases, however, man has a particular place (because of his duties to the gods or because of his limitations or gifts), even though—especially in many hunters' civilizations (*e.g.*, the African Bushmen and many North American Indian tribes)—the harmony of man and other forms of nature is emphasized.

There are innumerable tales of origin, very many of which belong to etiological tales rather than myth proper. The most undeniable mythical group of tales of origin are those that relate the origins of institutions. Certain initiation ceremonies or ritual acts are said to be founded in the beginning, in mythical times, and therefore they have their validity. Thus according to Raffaele Pettazzoni, an Italian historian of religion, myth forms "the Magna Charta of every institution" (see also CREATION, MYTHS AND DOCTRINES OF).

Myths of eschatology and destruction. Myths of eschatology by definition deal with "the end." The end, being the opposite of the cosmogony, means first and foremost the origin of death but also, in a wider sense, the end of the world. Special forms of eschatology are given in messianism (belief in a coming salvation figure) and millennialism (belief in a 1,000-year reign of the elect).

Myths about the origin of death, which is not considered "natural," are probably as widely diffused as creation stories. Three themes can be distinguished, even though they usually overlap. First, there are many myths that speak of a primordial time in which death did not exist

The origin of death

and relate that it came about as a result of an error or as a punishment or simply because the creator decided the earth would get too crowded otherwise. Secondly, many myths, usually in more "advanced" cultures, describe a paradise and a creator who first intended man to be immortal, so that the coming of death, in whatever manner it arrives, leads to a fall from paradise and is viewed as a tragedy. The third theme links death with sexuality and birth. Each of the three themes and their combinations occur with infinite variations. Possibly the third theme, linking death and generation, should be seen as a necessary existential coherency that was strongly experienced in the early planters' (root croppers') cultures.

Expectations of a cataclysmic end of the world are also expressed by myths. A universal conflagration with a final battle and defeat of the gods is part of Germanic mythology and has parallels in other Indo-European eschatological imageries. In many primitive religions specific expectations about the end of the world do occur but until recently have not received much scholarly attention. Many eschatological views of the universe are linked with religious movements that as a rule are spurred by historical events (see also ESCHATOLOGY).

Myths of culture heroes and soteriological myths. A great many nonliterate traditions have myths about a culture hero, who generally is not responsible for the creation but who completed the world and made it fit for human life; in short, he created culture. Outstanding examples of such culture heroes are Maui in Polynesia and Prometheus in Greek mythology. The bringer of culture is often also the bringer of health. Thus, the culture hero of the Woodlands and Plains Indians in North America is at the same time related to the foundation of the medicine society. A comparable figure occurs in many classical traditions as the "good son"; *e.g.*, Horus, the son of the god Osiris in Egypt, or the figure of the king in the Psalms. Health and (spiritual) salvation are synonymous, and this is implied in the Greek words *sōtēr* ("saviour") and *sōtēria* ("salvation"). Related to soteriological myths in many cases is the hope for a final and total salvation in which the "good" powers will triumph (such as through the Saoshyant, the saviour in Zoroastrianism; see also SALVATION).

Messianic and millenarian myths. Eschatological certainty and hope for a new world surge up from time to time in many civilizations, both primitive and advanced. In Melanesia, Africa, South America, and Siberia, innumerable religious movements have occurred in this century. Christian elements are usually visible, but the basic thrust in virtually all cases is indigenous. These cults and movements centre around prophetic leaders, often emphasize the return of the dead at the renewal to come, and are convinced of a catastrophic end, both historical and cosmic, of the present world. In many cases, the culture hero is expected to return and lead in battle against the evil (colonialist) forces. In the history of Judaism and Christianity, as in many primitive millenarian and messianic movements, there is an expectation of a new heaven and a new earth (see also MESSIAH AND MESSIANIC MOVEMENTS; MILLENNIALISM; TRIBAL RELIGIOUS MOVEMENTS, NEW).

Myths of time and eternity. The regularity of the heavenly bodies has made a great impression on people everywhere. The sky is the very image of transcendence, and the immutable order in the course of sun, moon, and stars suggests a time that transcends man's. That "transcendent time" is eternity. Many myths and mythological images concern themselves with the relationship between eternity and time on earth. The number four for the number of world ages figures most prominently. Ancient Persia knew of a complete world age of 12,000 years, divided in four periods of 3,000 each, at the end of which Ormazd (the Wise Lord) would conquer Ahriman (the Lie). Similarly, the Book of Daniel (in the Old Testament) mentions four kingdoms—of gold, silver, bronze, and a mixture of iron and clay, respectively—after which God will establish an everlasting kingdom. Four world ages, sometimes associated with metals, occur also in classical writers and in later speculations on history. Judaism developed the view of a 1,000-year period between the four world ages and the everlasting kingdom (hence the words millennium and millenarian). Although other numbers occur (three, six, seven, 12, and 72), four is dominant. In ancient Mexico this world was held to be preceded by four other worlds. India, in both Hindu and Buddhist texts, has developed the most complex system of world ages and worlds that arise and come to an end. Here too the number four is important; *e.g.*, the four ages (*yugas*) of decreasing length and increasing evil. Many writings, often with large numbers that have been considered fantastic by humanistic scholars, reflect exact astronomical observations and calculations. Some mythologies—*e.g.*, those of the Maya in Central America—have developed sophisticated views interrelating time and space. Mythological accounts of repetitions of worlds after their destruction occur not only in India but elsewhere, such as in Orphism and Stoic philosophy in ancient Greece (see also CHRISTIAN MYTH AND LEGEND; CALENDAR).

The ages of the world

Myths of providence and destiny. There is a broad range of mythical motifs between astrological calculations and devotional self-surrender, and there are many occasions at which a man may be filled with doubt about his own fate or the fate of his community. In some myths divine supremacy is marked by a god's mastery over fate. Marduk, the patron god of Babylon, acquires the "tablets of fate" in his primordial battle preceding the creation. There is no doubt about Zeus's supremacy in the Greek poet Hesiod's genealogical account of the gods, yet in later Greek texts even Zeus can be spoken of as powerless against fate, as in *Alcestis* by Euripides, a Greek dramatist. The same Euripides nevertheless can use the terms natural law or world reason as synonyms for Zeus. Mythological views of providence, destiny, or fate are given precise shades of meaning vis-à-vis dominant views in a tradition concerning justice and divine law, the philosophical problem of determinism, the theological problems of theodicy (justification of a good god with observable facts of evil), and of predestination. An important difference in mythological accounts of providence exists between those traditions that speak of the creation of the world as a result of God's will (as in Judaism, Christianity, and Islām) and those that attribute a more limited significance to worldly phenomena (as Buddhism does; see also PROVIDENCE, RELIGIOUS DOCTRINES AND MYTHS OF).

Cyclical time

Myths of rebirth and renewal. Myths of archaic traditions generally imply a conception of the world, nature, and man in terms of cyclical time. According to mythical lore, the Australian Aborigine at birth is really reincarnated through his mother into profane life. At his initiation he re-enters sacred time, and through his burial ceremony he returns to his original "spirit" state. Similar statements could be made about many tribal peoples, since the reality of their myths is first and foremost the reality of cosmic cycles. Special myths are narrated in many places in preparation for initiation procedures. In agricultural societies, in addition to the themes of cosmic renewal, renewal through birth, and rebirth through initiation ceremonies at the attainment of manhood and womanhood, the theme of seasonal renewal is of great importance. The cyclical concept of time in all these traditions is present in many of the great religious and philosophical systems, such as Brahmanism (a Hindu system), Buddhism, and Platonism, and to some extent is at variance with the idea of linear time that is typical of the historic religions (Judaism, Christianity, Islām). Jews, Christians, or Muslims, or indeed any culture cannot completely disregard the cyclical patterns of the seasons, work, festivities, or existence, which seem to be engraved on man's perception of the world.

Myths of memory and forgetting. Some of the North American medicine men claim to remember their prenatal existence. Such memory, according to their mythology, is lost in most common people. Similar myths of memory and forgetting are related to the hierarchy that exists in all archaic societies. The fundamental knowledge of the world, transcending ordinary consciousness, is not equal-

ly clear and available to everyone. Myths of memory can take the form of collective nostalgias. The Yaruros in South America, a very primitive tribe, ignorant of all agricultural and pastoral life, are one of many tribes in the course of modern Western cultural expansion that gave up the struggle for their own social and cultural existence. They expressed a yearning to return to the Great Mother ruling the land of the dead and awaiting them in her paradise. Less tragically, however, mythologies of memory and forgetting have a role in many traditions. They are of great significance in traditions that hold the idea of rebirth or reincarnation. Some people are said to remember previous existences, and a few (among them the Buddha), the very first. The veil of *māyā* ("illusion") in many Indian stories prevents a man from remembering his true origin and goal. All Gnosticism knows about a similar forgetfulness that must be resisted. In ancient Greece, Mnemosyne (Memory), the mother of the muses, is said to know everything, past, present, and future. She is the Memory that is the basis of all life and creativity. Forgetting the true order and origin of things is often tantamount to death (as in the case of Lethe, the river of death, which destroys memory, in Greek mythology). Anamnesis, "commemoration" or "recollection," is one of the crucial parts of the Christian celebration of the Communion. Through the anamnesis, the passion and death of the Lord is "applied" to the congregation. In philosophy, the imagery of forgetting and remembering occurs in the thought of Śaṅkara, a medieval Indian philosopher, and Plato in connection with the paramount calling of the thinker and the difficulty of living up to that calling.

Myths of high beings and celestial gods. Supreme celestial deities occur in many mythologies, with various qualities and attributes, in many shapes, and with great diversity in cultic significance. A cardinal distinction exists between the supreme being in many archaic or nonliterate traditions and the God of the great monotheistic systems (Judaism, Christianity, and Islām). Even though certain qualities seem alike in many cases (*e.g.*, transcendence, omniscience), the God of the latter arose historically in a reaction to polytheistic views and practices and demonstrates his supremacy accordingly, whereas the more archaic types of supreme beings nowhere show that militant feature in their mythologies. Another related general consideration about high beings and celestial gods is that their supremacy does not necessarily involve exclusion of other supreme beings. Outstanding examples are Viṣṇu, Śiva, and the great goddess in Hindu literature, who are each described as supreme yet do not reduce the "reality" of the others. "Supremacy" is not as unambiguous and general a term as it seems, and in Hinduism it refers first and foremost to the *perfection* of a deity in himself.

Distinction between high gods and one, unique God

The sky is considered sacred everywhere and is often related to or identical with the highest divinity. Nevertheless, supreme beings are always more than what can be explained from celestial phenomena alone, for they are often called creators of the world, founders of the order of the world, and protectors of law; and they are praised for their eternity and goodness. Often, the supreme being that created the world does not—or has ceased to—receive attention in the cult, although he may still be invoked in moments of great crisis. In a good many ancient agricultural societies, a male creator-god's place seems to have been occupied by a great goddess. The great goddess (as in the ancient Near East and India) is venerated principally because of her omnipotence, especially her power over life. The sky god-creator sometimes cedes to a divinity that is also related to the sky but apparently is experienced more concretely because of his activity. Such a divinity (especially in pastoral cultures) can be a god of atmospheric phenomena (storm, rain, thunder, or lightning), whose power for the good of the people is extolled. In spite of his power, however, he is one of several gods, and in some cases (Yahweh in ancient Israel and Allāh in Islām) one such god regains the full creative function of early creator gods, and in him all "true" divinity becomes concentrated. Also, a divinity related to the sun can take on the place of eminence; this has happened in some ancient imperial traditions (*e.g.*, Egypt; Inca Empire). Among sky gods who remained important in the mythologies of great ancient civilizations are Zeus in Greece, Jupiter in Rome, and T'ien in China.

Myths, tales, and legends concerning founders of religions and other religious figures. Founders of religions are by definition historical figures. Consequently, they cannot be part of myth, which describes events in other than historical time. Information concerning the founders of great religions (Confucius, Zoroaster, the Buddha, Moses, Jesus, Mani, Muḥammad), however, is couched in legendary materials that have many mythological features. The same thing can be said about many other religious figures (prophets, saints, or gurus, *i.e.*, Hindu spiritual teachers). The traditions that have preserved the memory of their founders have, as a rule, carefully guarded those elements that functioned most mythologically, in the proper sense that they stated most categorically realities that could not be known in any ordinary fashion or that raised the founder above ordinary historical conditions. Examples are the account of Jesus' prayer in Gethsemane, which no one heard according to the text itself, his statement that he was before Abraham, and his prophecies. Buddhist texts state that the Buddha not merely surpassed all *yogins* in knowledge of previous existences but, in fact, had conquered time. Well-known too are his predictions concerning the course and decline of Buddhism and (in Mahāyāna texts) his promises as to the future spiritual attainments of the *bodhisattvas*. Other examples are Muḥammad's eschatological teachings in the Qur'ān and those of Zoroaster.

Mythological features in records of historical figures

Myths of kings and ascetics. Genuine myths concerning kings are found only in traditions that know a form of sacred kingship. From ancient Babylon come preserved temple records concerning offerings to kings who were considered divine. Hymns addressed to them make references to the king's union with a goddess; *i.e.*, the mythological motif of the "sacred marriage." One of the epithets for the king in ancient Egypt was "endowed with life" or "imparting life." The twofold meaning of the epithet is significant and can serve to make the mythology of sacred kingship understandable in other places as well, because the function of the king is in fact double. He mediates between the divine world and the world of man, representing each to the other. Hence, in Egypt a sacrifice by an individual was understood as offered to the king and at the same time by the king. The king's role of mediator and protector brings royal mythologies close to myths of cultural heroes. Solemn procedures on which kings become divinized occur relatively late in history. An early and most conspicuous case of such an apotheosis (becoming divine) is that of Alexander the Great. Similarly, in later times, apotheosis took place for Roman emperors. A great many legends have accumulated around the figures of kings, resembling fairy tales or sagas rather than myths (*e.g.*, around King Aśoka of India and King Arthur of England). Stories about Frederick I Barbarossa, a medieval Holy Roman emperor, and Emperor Charles have a somewhat eschatological mythical flavour, because they are said to dwell each in his mountain (in the Kyffhäuser and the Unterberg, respectively) until a certain future moment.

Most narratives about great ascetics, as well as other saints, are much more like legends than myths. There are, however, instances of saints or ascetics who are presented as a more than worldly model, so that a case can almost be made for the mythological function of their legends (*e.g.*, al-Hallāj in Islām and St. Francis in Christianity). In the case of traditions that have asceticism as an integral part, certain figures and the legends around them do indeed function as paradigms (see also HERO WORSHIP; SACRED KINGSHIP; ASCETICISM).

Myths of transformation. Countless stories exist concerning the origin of peculiar rocks, properties of animals, plants, stars, or other features in the world. By nature, the majority of these accounts are etiological tales rather than myths. A number of myths, however, speak of cosmic changes brought about at the end of primordial

times. Also, an altogether different and extensive mythology exists concerning initiation rites and other "rites of passage" that involve transformations of man's being.

Cosmic transformation may concern an original world, without proper human means of existence and also without death, that was transformed through a new event (*e.g.*, the death of Hainuwele, a type of primal being known as a dema-deity) into the world known to mankind, a truly inhabitable world with vegetation, animals, and other necessities that had not existed before.

On a wider scale are myths that could be appendages to cosmogonic myths but that have not turned into mere etiologies. Many myths akin to the type of the dema-deity (like Hainuwele) and to the culture-hero type (like Prometheus, the giver of fire to the Greeks) account for events —such as the invention of agriculture, domestication of animals, and the use of fire—that have transformed the world for the benefit of man. Many others are just as closely related to cosmogonic accounts but tell of "setbacks" in primordial times. In agricultural societies, for example, myths have been collected that ascribe the unevenness of land or the formation of mountains to an ancient mishap or evil force.

Realization of the contents of myths

In "rites of passage" (*e.g.*, rites accompanying birth, attainment of maturity, marriage, death) the contents of myths are realized perhaps more fully than anywhere else. In each case the intention of the rites is that man's mode of being be affected, indeed transformed. Through the birth ceremony the child "becomes" a person, and through initiation an adolescent "becomes" an adult, a member of the men's society, or a warrior. There is a great variety of customs in different communities and traditions, but everywhere these rites actualize the cosmic processes and realities known from the myths. In many traditions the myths of the community are conveyed to the novice at the time of his initiation. In the major world religions, rites of passage have not come to an end, as evidenced in such ceremonies as circumcision, Baptism, weddings, and mortuary rites. In all instances, the rites derive their meaning from the core of the tradition, and for that reason man's being is transformed. In some cases the transformation derived from the dominant myth is far-reaching. The initiated shaman is able to transcend the ordinary human conditions and overcome conditions that would cause the death of a noninitiate. Through his initiation he is believed to have gone through death and thus conquered it. In certain Hermetic (a Hellenistic occultic, magical tradition) and Gnostic texts the certainty of attaining divine being is clearly expressed.

Myths and tales illustrating religious doctrines or reaffirming standards of conduct. The term myth is not properly applicable to most narratives that have as their explicit purpose the illustration of a doctrine or standard of conduct. Illustrations of this sort are allegorical rather than mythical. Familiar examples are the "myths" in Plato's dialogues and the parables of the New Testament. Illustrative tales have a considerable role also in Ṣūfism (Islāmic mysticism), rabbinic (Jewish biblical interpretive) literature, Ḥasidism (Jewish pietism), and Zen (intuitive) Buddhism. The nonmythological character is clear because the point made by the narrative is known or is supposed to be known from another source. Illustrative tales have a more subservient function than myths. They may clarify something to an individual or a group but do not take the place of the revelatory character of myth. The authority of myth sets a standard for conduct, whereas the illustrative tale by contrast makes a case for a certain superior authority or conduct.

Other types of myths and typologies. The types of myths presented thus far form a descriptive list of frequently occurring themes. This list is neither exhaustive nor the only one possible. A great many other themes (the great flood; myths related to totemism; souls, external souls, and guardian angels; weddings; thefts; and so on) have been left aside or have merely been touched on. Different and much longer lists could be drawn up. The greatest problem, however, remains that of defining myth and the effect one's definition has on typology. Crucial is the dichotomy, which is usually implied rather than stated, between the tendency of some to view myth as a lasting dimension of all human life and the tendency of others to regard myth as a stage in the development of human consciousness, which is important but past. Although the latter tendency has been prevalent in scholarship, the former now deserves attention. A typology of myths that begins with the recognition of a lasting mythological dimension of human existence is inclined toward a certain hierarchy of mythical themes. Examples are Mircea Eliade's insistence on the primacy of cosmogonic accounts or, earlier, Friedrich Schelling's interest in theogony as the basic myth.

The two tendencies, however, are a matter of philosophical commitment. This means that the decisive criterion of mythological classification cannot be found only by considering all available mythologies. The point of view is a crucial factor, and the issue is far from closed at the present state of scholarship.

MYTH IN MODERN SOCIETY

Secularization of myth and mythology. The problem of the secularization of myth is almost entirely a problem of definition. If one has decided, however, to regard myth as product of a past that has come to an end, insuperable difficulties arise in establishing the historical moment when that past period was terminated. It is virtually impossible to state with accuracy when a certain mythical theme becomes a "mere" literary theme or to determine in general when myths are no longer being created. It is more fruitful, as well as more defensible in view of the study of religious facts, to recognize that symbols, myths, and rituals are all subject to change in the course of time. Secularization is not an irreversible process that takes place once and for all, eventually leaving all myths devoid of their religious function. It is rather a process that takes place time and again in history. Secularization movements and mythification (or symbolization) movements are distinct aspects of the same historical processes. The formation of epic literature is an early example of secularization of myth. Nevertheless, epic literature itself can resume certain mythical functions.

Although the form of secularization brought about in modern history is one of many such forms, it is a new and complex type, to which many factors have contributed. Scientific, particularly astronomic, discoveries of the late medieval and Renaissance periods were accompanied by a new trust in cosmic laws and an increasingly abstract notion of God. More or less Euhemeristic historical accounts that were common in the Middle Ages and were a symptom of a certain secularization process themselves gave way to history writing, focussing on psychological, social, and economic facts. In philosophy, naturalism of various sorts opposed notions of transcendence that earlier systems had taken for granted. These are a few of the factors in the modern secularization of mythological views. The most common tendency in modern society has been to regard the characters and events in mythical accounts as not real or as by-products of realities that are not transcendent, but rather immanent.

Secularization with mythification

This secularization in modern society, like earlier secularization processes, is accompanied by forms of mythification. Though there is no general agreement on these matters, some of the most prominent features of modern society are mythological by nature. The aversion to transcendent realities has prototypes in ancient religious systems (especially Buddhism).

To recognize the myths that are dominant in one's own society is always difficult. The outline of myths from a past period, however, can be seen quite clearly; *e.g.*, the Marxist eschatological myth of the withering away of the state. This is not surprising, because a myth has its authority not by proving itself but by presenting itself. In this sense the authority of a myth indeed "goes without saying," and the myth can be outlined in detail only when its authority is no longer unquestioned but has been rejected or overcome in some manner by another, more comprehensive myth.

Demythologization of major religious traditions. Demythologization should be distinguished from seculariza-

tion. Every living mythology has to come to terms with the world in which it is transmitted and, therefore, inevitably goes through processes of secularization. Demythologization, however, refers to the conscious efforts people make to purify a religious tradition of its mythological elements. The term demythologization (*Entmythologisierung*) has been coined by Rudolf Bultmann, a German theologian and New Testament scholar. In the strict sense of the word, demythologizing efforts have been limited to theological discussions in 20th-century Christianity (see also CHRISTIAN MYTH AND LEGEND).

Survivals in modern consciousness and behaviour: superstitions and tales. E.B. Tylor, an English anthropologist of the 19th century, used the word survival for customs and beliefs that continued to be adhered to long after the context in which they had their meaning had ceased to exist. They connote the impression of superstitions, and hence the word survival usually has a slightly derogatory overtone. There are many survivals of myth in this sense. The myth of "the noble savage," well-known from Jean-Jacques Rousseau, an 18th-century French-Swiss moralist, can be understood as a survival of a paradisical mythology: Western man expected to find evidence of paradise on earth.

Once more, however, the suggestion that at one time myths ruled supreme and that all that now remain are mere remnants is to be avoided. The secularization process in modern society has affected two of the three forms of symbolization most adversely, viz., symbolic behaviour (cult, ritual, liturgy) and symbolic objects (sacred places and the sacredness of the earth itself) but not symbolic communication (myth). Indeed, the most striking feature of modern Western society is its insistence on communication (*e.g.*, freedom of the press). Because as a rule no reasons are adduced or believed to be necessary to justify this insistence, this modern sense of communication is mythological in structure.

Survival of symbolic communication

Commonly accepted forms of mythology in modern society, however, do not permeate all parts of that society or fulfill all needs. In all likelihood, no society has ever been perfectly homogeneous in its myths. In modern society a number of profound mythological needs are filled by myths borrowed from submerged or alien traditions. Modern society's neglect of cosmic symbolism (which typified all archaic tradition) has elicited a certain reaction and has led various "subcultures" to an interest in cosmic coherencies (such as astrology). The great scientific possibilities opened up by the 20th century have everything to do with the rise of a new literature, resembling myths, and often with an eschatological tenet (*e.g.*, science fiction).

Political and social uses of myth in modern times. In western European countries that have experienced anticlerical movements, it has been observed that politics has taken over the role that was played by religion or myth in previous times. This should not, however, lead to the conclusion that myth was ever apolitical or asocial. At all times reasons used to solidify a political unit, whether a village or a nation, have had a mythological component. In recent times, however, people have begun to think of social and political "ideologies" in terms of mythical functions, and conscious endeavours have been made to utilize myth for social and political purposes. The most conspicuous example has been the Nazi endeavour to revive Germanic mythology. The vast majority of sociopolitical uses of myth, however, are less conscious efforts. Marxism can be regarded as "a new *tao*" or way of nature, according to Robert Charles Zaehner, an English historian of religion, and the word ideology in much contemporary parlance could be replaced by the term mythology without harm. Finally, crucial terms in modern sociopolitical discussion, such as "freedom" and "equality," although they have a long and complex philosophical history, are usually posited in a manner quite analogous to the function of myth presenting its own authority.

BIBLIOGRAPHY. H. BAUMANN, "Mythos in ethnologischer Sicht," 2 pt., *Studium Generale*, 12:1–17, 583–597 (1959), the best general anthropological study of myth, of special importance to an understanding of the function of myth and the difference between myth and other forms of oral tradition; KEES W. BOLLE, *The Freedom of Man in Myth* (1968), on the relation between myth and literature, presenting a view of myth as constitutive of human existence; "Secularization As a Problem for the History of Religions," *Comparative Studies in Society and History*, 12:242–259 (1970), a comparative study of secularization processes in various cultures and periods; JOSEPH CAMPBELL, *The Masks of God*, vol. 1, *Primitive Mythology* (1959), neo-romantic and Jungian in interpretation; HENRY CORBIN *et al.*, *Man and Time* (1957), excellent contributions on mythologies concerning time in early Christianity, Islām, Mazdaism, and the Book of Changes; F. CREUZER, *Symbolik und Mythologie der alten Völker*, 4 vol. (1810–23), a classic (romantic) work; G. DUMEZIL, *Heur et malheur du guerrier* (1969; Eng. trans., *The Destiny of the Warrior*, 1970), on the problem of myth and epic in Indo-European studies; MIRCEA ELIADE, *Birth and Rebirth: The Religious Meanings of Initiation in Human Culture* (1958), indispensable for an understanding of myth in its relation to initiation ceremonies in all times and places; *From Primitives to Zen: A Thematic Sourcebook of the History of Religions* (1967), contains a wide selection of mythological materials; *Mythes, rêves et mystères* (1957; Eng. trans., *Myths, Dreams and Mysteries: The Encounter Between Contemporary Faiths and Archaic Realities*, 1961), of special interest for the study of myth in modern society, especially nostalgia for paradise and the function of psychoanalysis; *Myth and Reality* (1963), a collection of several essays on myth, including an appendix on myths and fairy tales, an essay on the structure of myths, and a number of important observations on the continuation of myths in modern times; *Le Mythe de l'éternel retour* (1949; Eng. trans., *The Myth of the Eternal Return*, new ed. 1965; also published as *Cosmos and History: The Myth of the Eternal Return*, 1959), a discussion of cosmically and historically oriented myths; H. and H.A. FRANKFORT *et al.*, *Before Philosophy: The Intellectual Adventure of Ancient Man* (1951), a discussion of the "mytho-poetic age"—neo-romantic but still stimulating; THEODOR H. GASTER, *Thespis: Ritual, Myth and Drama in the Ancient Near East*, rev. ed. (1961), a most persuasive study on the historical relation between cult and drama (folkloristic in character); L.H. GRAY (ed.), *The Mythology of All Races*, 13 vol. (1916–33), perhaps the best collection and discussion of myths ever published in English; S.H. HOOKE (ed.), *Myth and Ritual* (1933), the "manifesto" of the myth-ritual school; *Myth, Ritual, and Kingship* (1958), contains much information on the myth of divine kingship; WERNER W. JAEGER, *The Theology of the Early Greek Philosophers* (1947, reprinted 1968), a discussion of the historical problem of the relationship between myth and philosophy; KARL JASPERS and RUDOLF BULTMANN, *Die Frage der Entmythologisierung* (1954; Eng. trans., *Myth and Christianity: An Inquiry into the Possibility of Religion Without Myth*, 1958), a discussion between the two authors on demythologization; ADOLF E. JENSEN, *Mythos und Kult bei naturvölkern*, 2nd ed. (1960; Eng. trans., *Myth and Cult Among Primitive Peoples*, 1963), important for the mythology of dema-deities and theoretical questions concerning myth and sacrifice; CARL G. JUNG and K. KERENYI, *Einführung in das Wesen der Mythologie* (1941; Eng. trans., *Essays on a Science of Mythology*, 1950), the basic introduction to the psychological approach to myth in a Jungian manner; ALEXANDRE KOYRE, *From the Closed World to the Infinite Universe* (1958), a discussion of the rise of modern science, not in isolation, but within the context of philosophical and religious problems of the time; S.N. KRAMER (ed.), *Mythologies of the Ancient World* (1961); WALTER KRICKEBERG *et al.*, *Die Religionen des alten Amerika* (1961; Eng. trans., *Pre-Columbian American Religions*, 1968); THOMAS S. KUHN, *The Structure of Scientific Revolutions* (1962), contains a helpful discussion of scientific paradigms and models; G. VAN DER LEEUW, *Wegen en Grenzen*, 3rd ed. (1955; Eng. trans., *Sacred and Profane Beauty: The Holy in Art*, 1963), an elaborate and imaginative work on the relations between religion and art in all forms: dance, drama, literature, plastic art, architecture, and music; CLAUDE LEVI-STRAUSS, *La Pensée sauvage* (1962; Eng. trans., *The Savage Mind*, 1966), a demonstration of structuralism; CHARLES H. LONG, *Alpha: The Myths of Creation* (1963), the best and most available anthology of creation myths in English; BRONISLAW MALINOWSKI, *Magic, Science and Religion, and Other Essays* (1948), contains the influential essay "Myth in Primitive Psychology"; JOHN MIDDLETON (comp.), *Myth and Cosmos: Readings in Mythology and Symbolism* (1967), 18 well-known anthropologists present their views on myth, most of them on the basis of their own findings in a variety of nonliterate societies; WALTER F. OTTO, *Die Götter Griechenlands*, 3rd ed. (1947; Eng. trans., *The Homeric Gods: The Spiritual Significance of Greek Religion*, 1954), mainly a treatise on the significance

of Homeric mythology; R. PETTAZZONI (ed.), *Miti e leggende,* 4 vol. (1948–63), a classic by Pettazzoni (and other Italian scholars), containing an unsurpassed collection of myths and exhaustive bibliographies; MACLINSCOTT RICKETTS, "The North American Indian Trickster," *History of Religions,* 5:327–350 (1966), an excellent essay on a most puzzling character of North American Indian mythology; H.H. ROWLEY (ed.), *The Old Testament and Modern Study* (1951), important for the problem of myth, directly, and by implication; H. SCHAERER, *Die Gottesidee der Ngadju Dajak in Süd-Borneo* (1946; Eng. trans., *Ngaju Religion: The Conception of God Among a South Borneo People,* 1963); F.W.J. SCHELLING, *Philosophie der Mythologie,* 2 vol. (published in *Sämtliche Werke,* 1856–58, reprinted 1966); THOMAS A. SEBEOK (ed.), *Myth: A Symposium* (1958), nine specialists present their basic views of myths (fascinating because of the diversity); STITH THOMPSON (ed.), *Tales of the North American Indians* (1929, reprinted 1966), a readily available collection of North American Indian and other stories; RUTH M. UNDERHILL, *Red Man's Religion: Beliefs and Practices of the Indians North of Mexico* (1965), an introduction to North American Indian mythology, with superb bibliographies; JAN DE VRIES, *Forschungsgeschichte der Mythologie* (1961), the standard work on the history of scholarship devoted to myth; *Heldenlied en heldensage* (1959; Eng. trans., *Heroic Song and Heroic Legend,* 1963), especially important for the relation between myth and epic (or saga); *The Study of Religion: A Historical Approach* (1967; orig. pub. in Dutch, 1961), a historical survey of scholarship devoted to religious phenomena, in which mythology is prominent; F.A. YATES, *Giordano Bruno and the Hermetic Tradition* (1964), a monograph on the persistence of Hermetic views during the rise of science.

(K.W.Bo.)

Nabokov, Vladimir

One of the contemporary novelists most admired by critics, Vladimir Vladimirovich Nabokov, of Russian birth but from his 20th year a resident in western Europe and the United States, used a highly chiselled and lucid prose to express by allegory, parody, and sardonic jest a very personal vision of the world. With the exception of *Lolita,* Nabokov's greatest popular success, his writing was generally too involuted to permit precise description, and more than perhaps any other celebrated 20th-century author, he left it to the reader to interpret his work.

© Philippe Halsman

Nabokov, 1968.

Nabokov was born in St. Petersburg (now Leningrad), Russia, on April 23, 1899, into an old aristocratic family. His father, V.D. Nabokov, was a leader of the pre-Revolutionary liberal Constitutional Democratic Party (Kadets) in Russia and the author of numerous books and articles on criminal law and politics, among them *The Provisional Government* (1922; Eng. trans. 1970), one of the primary sources on the downfall of the Kerensky regime. In 1922, after the family had settled in Berlin, V.D. Nabokov was assassinated by a reactionary rightist while shielding another man at a public meeting; and although his novelist son disclaimed any influence of this event upon his art, the theme of assassination by mistake has figured prominently in Nabokov's novels. Nabokov's enormous affection for his father and for the milieu in which he was raised is shown in his autobiography *Speak, Memory* (revised version, 1967).

Early verse collections

Nabokov published two collections of verse, *Poems* (1916) and *Two Paths* (1918), before leaving Russia in 1919. He and his family made their way to England, and he attended Trinity College, Cambridge, on a scholarship provided for the sons of prominent Russians in exile. While at Cambridge he first studied zoology but soon switched to French and Russian literature; he graduated with first-class honours in 1922 and subsequently wrote that his almost effortless attainment of this degree was "one of the very few 'utilitarian' sins on my conscience." While still in England he continued to write poetry, mainly in Russian but also in English, and two collections of his Russian poetry, *The Cluster* and *The Empyrean Path,* appeared in 1923. In Nabokov's mature opinion these poems were "polished and sterile."

Between 1922 and 1940 Nabokov lived in Germany and France, and while continuing to write poetry, he experimented with drama and even collaborated on several unproduced motion-picture scenarios. By 1925 he settled upon prose as his main genre. His first short story had already been published in Berlin in 1924. His first novel, *Mashenka* (*Mary,* 1970), appeared in 1926; it was avowedly autobiographical and contains descriptions of the young Nabokov's first serious romance as well as of the Nabokov family estate, both of which are also described in *Speak, Memory.* Nabokov did not again draw so heavily upon his personal experience as he had in *Mashenka* until his episodic novel about an émigré professor of entomology in the United States, *Pnin* (1957), which is to some extent based on his experiences while teaching (1948–58) Russian and European literature at Cornell University, Ithaca, New York.

His second novel, *King, Queen, Knave,* which appeared in 1928, marked his turn to a highly stylized form that characterized his art thereafter. His chess novel, *The Defense,* followed two years later and won him recognition as the best of the younger Russian émigré writers. In the next five years he produced four novels and a novella. Of these, *Despair* and *Invitation to a Beheading* were his first works of importance and foreshadowed his later fame.

Work in Germany and France

During his years of European emigration Nabokov lived in a state of happy and continual semipenury. All of his Russian novels were published in very small editions in Berlin and Paris. His first two novels had German translations, and the money he obtained for them he used for butterfly hunting expeditions (he eventually published 18 scientific papers on entomology). But until his best-seller *Lolita* (1955), no book he wrote in Russian or English produced more than a few hundred dollars. During the period in which he wrote his first eight novels, he made his living in Berlin and later in Paris by giving lessons in tennis, Russian, and English and from occasional walk-on parts in films (now forgotten). His wife, the former Véra Evseyevna Slonim, whom he married in 1925, worked as a translator. From the time of the loss of his home in Russia, Nabokov's only attachment was to what he termed the "unreal estate" of memory and art. He never purchased a house, preferring instead to live in a series of houses rented from other professors on sabbatical leave. Even after great wealth came to him with the success of *Lolita* and the subsequent great interest in his previous work, Nabokov and his family (he and his wife had one son, Dmitri) chose to live (from 1959) in genteelly shabby quarters in a Swiss hotel.

The subject matter of Nabokov's novels is principally the problem of art itself presented in various figurative disguises. Thus, *The Defense* seemingly is about chess, *Despair* about murder, and *Invitation to a Beheading* a political story, but all three works make statements about art that are central to understanding the book as a whole. The same may be said of his plays, *Sobytiye* ("The Event"), published in 1938, and *The Waltz Invention.* The problem of art again appears in Nabokov's best novel

in Russian, *The Gift,* the story of a young artist's development in the spectral world of post-World War I Berlin. This novel, with its reliance on literary parody, was a turning point: serious use of parody thereafter became a key device in Nabokov's art. His first novels in English, *The Real Life of Sebastian Knight* (1941) and *Bend Sinister* (1947), do not rank with his best Russian work. *Pale Fire* (1962), however, a novel consisting of a long poem and a commentary on it by a mad literary pedant, extends and completes Nabokov's mastery of unorthodox structure, first shown in *The Gift* and present also in *Solus Rex,* a Russian novel that began to appear serially in 1940 but was never completed. *Lolita* (1955), with its antihero, Humbert Humbert, who is possessed by an overpowering desire for very young girls, is yet another of Nabokov's subtle allegories: love examined in the light of its seeming opposite, lechery. *Ada* (1969), Nabokov's 17th and longest novel, is a parody of the family chronicle form. All of his earlier themes come into play in the novel, and because the work is a medley of Russian, French, and English it is his most difficult work.

First novels in English

Nabokov's major critical works are an irreverent book about Nikolay Gogol (1944) and a monumental four-volume translation of, and commentary on, Pushkin's *Eugene Onegin* (1964). What he called the "present, final version" of the autobiographical *Speak, Memory,* concerning his European years, was published in 1967, after which he began work on a sequel, *Speak On, Memory,* concerning the American years.

As Nabokov's reputation grew in the 1930s so did the ferocity of the attacks made upon him. His idiosyncratic, somewhat aloof style and unusual novelistic concerns were interpreted as snobbery by his detractors—although his best Russian critic, Vladislav Khodasevich, insisted that Nabokov's aristocratic view was appropriate to his subject matters: problems of art masked by allegory.

Nabokov's reputation varies greatly from country to country. In his native land he was not only not published because he was a "White Russian émigré" (he became a U.S. citizen in 1945) but was condemned for other reasons. Critics of strong social convictions in the West also generally hold him in low esteem. But within the intellectual émigré community in Paris and Berlin between 1919 and 1939, V. Sirin (the literary pseudonym used by Nabokov in those years) was credited with being "on a level with the most significant artists in contemporary European literature and occupying a place held by no one else in Russian literature." His reputation after 1940, when he changed from Russian to English after emigrating to the United States, mounted steadily until the 1970s, when he was acclaimed by a leading literary critic as "king over that battered mass society called contemporary fiction." He died in Montreux on July 2, 1977, of a virus infection.

MAJOR WORKS

NOVELS: In Russian (English translations after 1940 mainly by the author and his son Dmitri Nabokov): *Mashenka* (1926; *Mary,* trans. by Michael Glenny with Nabokov's assistance, 1970); *Korol-dama-valet* (1928; *King, Queen, Knave,* 1968); *Zashchita luzhina* (1930; *The Defense,* 1964); *Podvig* (1932; *Glory,* 1971); *Camera obscura* (1933; first Eng. trans. 1936, altered by Nabokov for U.S. version and retitled *Laughter in the Dark,* 1938); *Otchayaniye* (1934; *Despair,* 1937, altered and enlarged by Nabokov for U.S. version, 1966); *Priglasheniye na kazn* (1935; *Invitation to a Beheading,* 1959); *Dar* (incomplete 1935–36, complete 1952; *The Gift,* 1963). In English: *The Real Life of Sebastian Knight* (1941); *Bend Sinister* (1947); *Lolita* (1955); *Pnin* (1957); *Pale Fire,* including a long poem (1962); *Ada; or, Ardor: A Family Chronicle* (1969); *Transparent Things* (1972).

SHORT STORIES AND NOVELLAS: In Russian: *Soglyadatay* (1938), collects a novella and 12 short stories written 1930–36 (the title novella, first published 1930, was trans. as *The Eye,* 1965, and some of the stories have been trans. in later collections); *Vesna v Fialte i drugiye rasskazy* (1952), collects 14 short stories written 1931–40 (some trans. in later collections). In English: *Nabokov's Dozen* (1958), short stories and novella, some trans. from the Russian by the author and others; *Nabokov's Quartet* (1966), collects three novellas written in Russian in the 1930s and trans. by Dmitri Nabokov, with one published in English in 1959; *Nabokov's Congeries* (1969), collects stories written in Russian in the 1930s and trans. by Dmitri Nabokov and others, with some written in English; *A Russian Beauty, and Other Stories* (1973), 13 early stories trans. by Dmitri Nabokov and Simon Karlinsky in collaboration with the author; *Tyrants Destroyed, and Other Stories* (1975), 12 stories trans. from Russian by Nabokov and Dmitri Nabokov and one English story; *Details of a Sunset, and Other Stories* (1976), 13 early stories trans. by Nabokov and Dmitri Nabokov.

OTHER WORKS (MEMOIRS): *Conclusive Evidence: A Memoir* (1951); *Speak, Memory: An Autobiography Revisited* (1967), revised version of *Conclusive Evidence,* with new material and additions from *Drugiye berega* (1954); *Look at the Harlequins!* (1974), a pseudo-memoir. (POEMS): In Russian: *Poemy 1929–1951* (1952). In English: *Poems* (1959); *Poems and Problems* (1971), 39 Russian poems with English trans. by Nabokov and 14 English poems (published earlier in *Poems*). (PLAYS): *Sobytiye* (1938), a three-act comedy; *Izobretaniye valsa* (1938; *The Waltz Invention,* 1966). (CRITICAL BIOGRAPHY): *Nikolai Gogol* (1944). (TRANSLATIONS): *Three Russian Poets: Translations of Pushkin, Lermontov, and Tiutchev* (1944, with Dmitri Nabokov); Lermontov's *A Hero of Our Time* (1958, with introduction by Nabokov); Pushkin's *Eugene Onegin,* 4 vol. (1964, with introduction and commentary).

BIBLIOGRAPHY. VLADIMIR NABOKOV, *Speak, Memory* (1967), the author's autobiography up to 1940; three books by ANDREW FIELD, *Nabokov: His Life in Art* (1967), *Nabokov: A Bibliography* (1972), and *Nabokov: His Life in Part* (1977); C. NEWMAN and A. APPEL, JR. (eds.), *For Vladimir Nabokov on His Seventieth Birthday* (1971), a festschrift of reminiscences of Nabokov and critical articles and tributes first published in *Tri-Quarterly* magazine.

(A.Fi.)

Nāgāland

Nāgāland, a state of India, occupies a region of forested hills and mountains between Burma and the Brahmaputra Valley of upper Assam. It is bordered on the east by Burma, on the south by the Indian state of Manipur, and on the west and northwest by the Indian state of Assam; a small section extends north from the Burmese border into the Tirāp division of Arunachal Pradesh, an Indian union territory. Its total area is 6,366 square miles (16,488 square kilometres) and its population (1971) is 516,449. The capital is Kohīma.

How the word Nāga came into use to designate the many tribes and subtribes that occupy the area is still a matter of conjecture. Some derive the name from Sanskrit *nāga,* "snake"; some from Hindu *naṅgā,* "naked"; and some from *naga,* "belonging to the hills," "highlander." The most widely accepted view is that it is derived from *nok,* meaning "people," or "folk." The people once called themselves by their tribal names, such as Angami, Sema, or Ao, but the upsurge of nationalism that led to statehood made Nāga so popular that it became a matter of pride to be so designated. (For an associated physical feature, see BRAHMAPUTRA RIVER.)

History. Nāgāland has no early written history. The first reference to the Nāgas is found in chronicles of the neighbouring Ahom kingdom of Assam. These chronicles tell of an invasion of the region in 1228 (although scholars consider a later date more probable) and contain information regarding the Nāga tribes, their economy and customs, and indicate that the Ahom did not rule the country but maintained friendly relations with the Nāga chieftains. There were occasional clashes between the Nāgas and the Ahoms, and a Burmese invasion of Assam in 1816 led to an oppressive Burmese rule from 1819 until the establishment of British rule over Assam in 1826, after which the Nāga hill areas were gradually annexed to the British Empire. The advent of British administration, which by 1892 encompassed the whole of Nāga territory (except the wild and rugged Tuensang area in the northeast), brought an end to the practices of headhunting and intervillage raids, and the Nāgas settled down to a peaceful life of cultivation and trade.

Nāga separatism

When India became independent in 1947, a strong nationalist movement sought a union of the Nāga tribes, claiming that Nāgāland was a sovereign state separate from India, and a small armed force was organized among them. At the same time, however, there were

many political differences among the Nāgas. In 1957, after negotiations with a Nāga People's Convention at Kohīma, the Indian government placed the Nāga Hills district and Tuensang area under central administration.

This status, however, proved unsatisfactory to a large segment of the Nāgas, and unrest continued, taking the form of noncooperation with the Indian government, nonpayment of taxes, and armed conflicts. A rebel force engaged in action against the Indian army and committed acts of sabotage. The Indian government and representatives of the Nāgas, however, continued to negotiate, and an agreement reached at a Nāga People's Convention of July 1960 provided that Nāgāland should become a full state in the Indian Union. Nāga representatives found this solution acceptable because of the wide autonomy that would be granted to the new state. Nāgāland accordingly entered the Indian Union in 1962, and in 1964 a democratically elected government took office and proceeded to administer the state. Dissident factions within Nāgāland refused to accept this solution, however, and rebel activity continued, increasingly assuming the form of banditry and often motivated more by tribal rivalry and personal vendetta than by political aspiration. The state received substantial financial help from the central government, but relations were strained somewhat by the increasing presence of the Indian army in the territory. From June 1966 to March 1969 there was a ceasefire, and talks between the dissidents and the government of India continued throughout the period. In clashes that occurred after the breakdown of talks, evidence was produced suggesting that large numbers of Nāga dissidents had received training in China; it can be said with certainty that the Nāga underground had some contacts with the Chinese.

The dissidents became increasingly fragmented in the late 1960s, and several groups surrendered to the Indian army. In the early 1970s, however, a hard core of extremists were continuing their activities. A faction of the dissidents set up a self-styled government in the Tuensang area, but its activities were of minor significance.

The landscape. *Physiography.* Except for a few square miles of plain, the entire state is covered with ranges of hills that geologically are a part of the Himalayan system. They consist of sedimentary formations of the Tertiary Period (formed up to 65,000,000 years ago). The Nāga Hills rise abruptly from the Brahmaputra Plain in the north, initially to low hills of about 2,000 feet, which increase in height toward the east to more than 6,000 feet, with isolated peaks reaching 10,000 feet or more. The hills merge with the Pātkai Range, which separates Nāgāland from Burma and which reaches a height of 12,552 feet (3,826 metres) at Mt. Saramati, on the border. The northwest is drained by tributaries of the Brahmaputra. The southeast, comprising all of Tuensang and the southeastern part of Kohīma district, has an elevation of more than 4,000 feet and is drained by tributaries of the Chindwin River of Burma. The headstreams of the Barak flow through the southwest. The hills, clothed for the most part with dense forest, scrub, and grassland, are deeply dissected, though some slopes provide suitable sites for shifting cultivation. The higher slopes are covered with shallow, acidic, brown forest soils whereas there are comparatively deeper soils on the lower slopes.

Relief and drainage

Climate. Nāgāland has a wet monsoon climate, with the average annual rainfall varying between 70 inches in the southeast and more than 100 inches in the northeast. July and August are the rainiest months. The temperature is relatively low. Varying according to altitude, the summer temperature ranges between 70° F (21° C) and 104° F (40° C), while in winter it rarely goes below 39° F (4° C), although frost is common over higher elevations. Because of the high humidity, the summer is sultry, especially over the lower hills and in the valleys.

Vegetation and animal life. The hills are covered with luxuriant forest approximating in type that of the eastern Himalayas, except that there is no true alpine zone. The hill forests may be separated generally into evergreen, broad-leaved, and pine forests. But where the forests have been cleared for *jhum* (shifting cultivation), high grass, reeds, and scrub jungles have sprung up in profusion. There are many species of palm and several of calamus (climbing palms, or rattan). Below 4,000 feet occur wet monsoon and evergreen forests, with mahogany and other valuable timber species and varieties of bamboo. On the higher hills in the east and southeast, oak, pine, and other coniferous vegetation appear. The surveyed forest area of the state is about 777,000 acres.

In the lower hills, especially in the north and northeast, are found rhinoceroses, elephants, tigers, leopards, bears, sambar deer, buffalo, boars, and wild oxen. Also seen are several kinds of monkey, including the long tailed langur, flying lemur, and pink faced monkey. Porcupines, pangolins (scaly anteaters), goats, wild dogs, foxes, civets, and mongooses are also found. Game birds include wild fowl, partridges, woodcocks, and pheasant. The great Indian hornbill is the most treasured bird—for use in decoration and magic. The gayal, a domesticated breed of wild ox, is a prized animal. The colourful smaller animals and the wide variety of birds make the jungle attractive for tourists.

Animal life

Population. *Ethnic composition.* The Nāgas belong to the Indo-Mongoloid ethnic group. They have long been isolated and they lack cohesion as a single people—there are 17 major tribes and several subtribes. Though all have much in common, they differ in physique, dialect, and customs. The tribal groups have formed new affiliations and are using names hitherto unknown. The Chakhesangs, for example, are a combination of Chakru, Khezha, Sangtams, and a few Rengmas. Similarly, the Zeliangs are a mixed group of Zemis, Liangmais, and others. The big Konyak tribe has two sharply distinguished divisions, one ruled by powerful autocratic chiefs and the other, which now calls itself Shaymnyuyungmang, more democratic.

The Angamis, Chakhesang, Zeliangs, Rengmas, and Kukis are southern tribes living in the Kohīma district. The Tuensang district in the east is the home of the Konyaks and the Upper Konyaks, occupying the north and northeast, and the Phoms, Changs, Sangtam, Yimchungrs, Khienmungans, Sangtams, and a large number of subtribes, occupying the west and south. The major tribes in the western district of Mokokchūng are the Aos, Lhotas, and Semas.

Languages. The Nāga tribes do not have a common language, and there are 60 spoken dialects. In some areas dialects vary even from village to village. The languages all belong to the Sino-Tibetan family. Intertribal conversation generally is carried on through broken Assamese, which is also now commonly used and is called Nagamese. Many Nāgas have also become acquainted with Hindi and English.

Religion. The traditional religion of the Nāga tribes is not based on any deep-rooted philosophy and does not demand any spiritual or mystic participation of its followers. Generally, all tribes give special status to the sun and the moon. Nevertheless, there is a vague conception of a supreme creator, and there are many minor deities. Nature is seen to be alive with invisible forces, with which priests and medicine men mediate. They also lead rites and festivals, bless the marriage bed, and protect the craftsman at his work. Nāga ideas of afterlife vary from tribe to tribe, but there is wide agreement that the soul does not perish at death.

With the advent of British rule, Christianity was introduced. The American Baptist Mission and various church unions were active, especially in the Kohīma and Mokokchūng districts. Out of the total population, about 67 percent are now Christian, about 11 percent Hindu, and less than 1 percent Muslim, the rest belonging to other faiths.

Rural–urban distribution. Nāgāland is extremely rural. About 90 percent of its population live in isolated villages. There are altogether some 860 villages and three urban centres. The villages are normally large, many having more than 5,000 inhabitants. Their size and singular character is striking even in the wilder Tuensang border ranges. Built on the most prominent points along the ridges of the hills, they were formerly stockaded; some had village gates—massive wooden doors approached by

Nāga villages

narrow, sunken, winding paths. The Aos and Lhotas set out their houses in streets, frequently along the top of a ridge. The villages are usually divided into *khels,* or quarters, each with its own headmen and administration. The houses are fairly large, sometimes very large, reflecting the status of the owner. Many are built on stilts. Different areas are dominated by different tribal groups, which have well-defined portions of lands and sources of water. Tuensang district is comparatively densely populated (83 persons per square mile). Kohīma, Mokokchūng and Dimāpur are the only urban centres, with 21,398, 17,381, and 12,292 inhabitants (1971), respectively. Urban growth has been very rapid since 1961.

Administration. In common with the other states of India, the administration of Nāgāland is conducted by a council of ministers, which is headed by a chief minister and is responsible to the Nāga legislative assembly. The legislative assembly has 52 members. The state is divided into three administrative districts—Kohīma, Mokokchūng, and Tuensang—each under a deputy commissioner and magistrate. The deputy commissioner is responsible for the maintenance of law and order, revenue, and criminal administration. Each district is divided into three subdivisions headed by subdivisional officers responsible to the deputy commissioner.

Provisions for tribal autonomy

The administration of Nāgāland differs from other states in that wider autonomy has been provided in the constitution to the tribal communities. Each tribe has three local councils—the village council, the range council, and the tribal council. These deal with disputes involving breaches of customary laws and usages. Appeals are made to the Nāga Tribunal in respect of cases decided according to customary law. Special provisions were made for Tuensang, which was not placed under the direct jurisdiction of the assembly but was put under a regional council consisting of elected representatives of all the tribes in the district, to which the governor might also nominate members. The deputy commissioner is its ex officio chairman. It elects representatives from among its members to the legislative assembly.

Social conditions. This hill and forest land, hitherto an underdeveloped region, has been witnessing a sociocultural renaissance. From its formation in 1964, the popular democratic government rapidly established a good measure of stability and directed its efforts toward improving social conditions through systematic planning. Under a special plan, inaugurated in 1958, the highest item of expenditure was on roads, and greater emphasis was given to public health and water supply. Simultaneously, elaborate educational and community development programs were undertaken.

In 1957 there were 12 hospitals, 18 dispensaries, and five mobile medical units in Nāgāland. By the 1970s there were 28 hospitals, 77 dispensaries, and a large mobile unit to serve the interior villages. Because of the high incidence of tuberculosis, a hospital was established to treat victims of the disease. Malaria is another major health problem, and an antimalaria unit with trained personnel has been deployed. Under extensive community development programs, drinking water has been supplied to over 800 villages.

In 1971 about 27 percent of the population was literate. Since then there has been much educational development, and the literacy percentage has been increasing at a rapid pace. There are four colleges for general education, all affiliated with Gauhāti University in Assam. Also, there are 33 high schools, about 150 junior high schools, and nearly 2,000 primary schools, as well as a limited number of technical schools, basic training schools, and schools for special education.

The social and economic welfare activities of the state include consumer trading societies; farming and dairy, marketing, and other cooperatives; and adult literacy centres, libraries, youth clubs, and women's societies.

The economy. *Agriculture and forestry.* Agriculture is the mainstay of the economy. About 90 percent of all workers are cultivators. Autumn and winter rice, small millets, pulses (edible seeds), oil seeds, fibres, sugarcane, potato, and tobacco are the principal crops. Production, however, is inadequate, and Nāgāland has to depend largely on imports from neighbouring states. This deficiency may be attributed to inadequacy of fertile, cultivable land and the common practice of shifting, or *jhum,* cultivation.

Cultivation practices

Under shifting cultivation the jungle on the hill slopes is cleared and the undergrowth is burned down. The ground is then lightly hoed and sown for two successive years. After two harvests, the land is allowed to remain fallow for eight to ten years. In higher or otherwise unsuitable regions, *jhum* is replaced by terrace cultivation. The Angamis of Kohīma district excel in the art of terracing and irrigation. The implements of the cultivator include the light hoe, *dao* (a heavy knife), and sickle. Except in the plains, the plow is not used. Agriculture is undergoing slow but steady transformation, and wasteful practices are being progressively replaced by improved methods.

Forestry is a primary source of income and employment. Quantities of fine timber, such as mahogany, are produced, as well as timber for fuel and forest products.

Minerals. It is believed that there are rich deposits of minerals in Nāgāland, but limestone and low-grade coal are the only two known at present. The limestone, ranging in colour from light gray to a deep blue, makes an excellent building stone. Low-grade coal occurs in the Wokha, Mokokchūng, and Naginimārā areas. Naginimārā is the only producing area at present but proposals have been made to mine coal from other areas. Many of the coal measures contain iron ore in the form of clay, ironstone, and impure limonite. Salt is locally produced from brine wells. Petroleum seepages occur in the Dikhu Valley, and the geological homogeneity of northern Nāgāland with the oil and coalbearing area of Upper Assam suggests the presence of exploitable oil reserves.

Industry. Lack of raw materials, financial resources, and power, and, above all, poorly developed transport and communication have hindered the development of industry. In 1970 there were in the state some 25 small bakeries, 18 motor vehicle service and repair shops, six rice mills, and 12 cabinet and furniture factories. A sugar mill was under construction at Dimāpur and was expected to manufacture 2,800,000 gallons of alcohol for wine production; it will be the first manufacturing industry of any importance in the state. A small fruit-canning centre is located at Chanki.

Power. Electric power generation started only after World War II. Two small hydroelectric schemes begun by Nāga entrepreneurs were subsequently taken over by the state government. By the early 1970s all the district headquarters and even some of the subdivisional administrative centres had electricity.

Cottage industry. Even though the state is devoid of manufacturing industries, cottage industry is common. Nāgas are noted weavers and expert in basketry, woodwork, and pottery. They weave beautiful shawls, handsome ties, and colourful skirts. Cottage industry, however, suffers from lack of raw materials and improved tools, competition from outside, and the absence of a sufficiently wide market.

Communications. Nāgāland depends mostly on roads for communication. The North-East Frontier Railway runs along the south bank of the Brahmaputra River and provides the single rail link to Nāgāland. It extends a little way across the Assam border. The total length of railway line within Nāgāland is only five miles.

There are only a few all-weather roads. A national highway runs from Dimāpur through the southern part of the state, via Kohīma to Imphāl, the capital of Manipur. The second important road runs from Amguri, in the Sibsāgar district of Assam, to Mokokchūng, via Tuli and Merangkong village. The third important road links Tuensang with Mokokchūng. And Mokokchūng, Zunheboto, Kohīma, and Phek are also connected with short all-weather roads. There is a total of 2,209 miles of road in Nāgāland, and more roads were under construction. Post and telegraph communications have developed rapidly since 1964. There are telephone exchanges at Kohīma, Dimāpur, and Mokokchūng.

Cultural life. Nāga society contains varying patterns of local organization, ranging from near dictatorship to ex-

treme democracy. Among the Semas and Changs there is a system of hereditary chieftainship. The Konyaks are ruled by powerful chiefs, or *āngs*, whose word is law and who are held to be sacred. The Aos have councils of elders representing the main family groups in the village. The Angamis, Lhotas, Rengmas, and others are democratically run.

Tribal institutions and customs

A prominent village institution is the *morung* (communal house, or dormitory), for young unmarried men. Some tribes have communal houses for unmarried girls also. Formerly, skulls and other trophies of war were hung in the *morungs*. The pillars are still carved with striking representation of tigers, hornbills, and human and other figures.

The basic interest of every Nāga is in his family, the clan, the *khel*, and the village, and the Nāgas are passionately attached to their land and traditional customs. A central feature of their lives is the Feast of Merit, consisting of a series of ceremonies that finally culminate with the sacrifices of oxen (called *mithun* by the Nāgas).

Although there is no matriarchy among the Nāgas, women hold a high and honourable position. They work on equal terms with men in the fields and make their influence felt in the tribal councils.

The people have a rich vein of poetic thought that is little known outside the state. They have a passion for beauty and colour that inspires them to decorate almost everything they use or wear, and their poetry, folk songs, tales, and folklore all express an exuberant concern for life, romance, and a love of chivalry.

BIBLIOGRAPHY. The *Census of India, 1961, Nagaland,* gives a general account of the state together with important data on geographic, economic, cultural, and social aspects. ANAND V.K. MAJOR, *Nagaland in Transition* (1967), is a traveller's account of Nāgāland. *The Gazetteer of India*, vol. 1, *Country and People* (1965), contains ten chapters written by eminent scholars on the physiography, climate, geology, flora, fauna, people, languages, religion, social structure, and life of all states of the Indian Union. See also ELWIN VERRIER, *Nagaland* (1961); and the *Statistical Handbook of Nagaland* (1967), an invaluable publication for statistical data.

(M.B.)

Nāgārjuna

From the 3rd century AD onward the Buddhist traditions in India, China, Korea, Japan, and Tibet have paid homage to a much-revered monk by the name of Nāgārjuna, whose spiritual insight and power to destroy evil were regarded as phenomenal. The fact that various texts ascribe different religious qualities to him and give dates for his life that range over 500 years suggests that the references available may pertain to several persons and may include some imaginary accounts. At the base of these traditions, however, stands an astute philosopher who clarified the meaning of the significant Buddhist term *śūnyatā* "emptiness" and became the founder of the Mādhyamika, or Madhyamaka ("Middle Path"), school of Buddhism. Such a feat was recognized not only as an intellectual but also as a deep spiritual achievement, so it is not surprising that within a hundred years after his death there were reports of devotees honouring him as a Buddha. In later centuries he was recognized as a patriarch by several Buddhist schools.

Traditional accounts of life

The earliest available biographical account of Nāgārjuna is in Chinese, supplied about AD 405 by a renowned Buddhist translator, Kumārajīva. It agrees with other Chinese and Tibetan accounts that Nāgārjuna was born in South India into a Brahmin (Hindu priestly caste) family. The stories of his boyhood are contradictory in historical terms but indicate that he possessed an extraordinary intellectual capacity and underwent a spiritual conversion when he learned the profound meanings of the doctrines of Mahāyāna (Great Vehicle) Buddhism, now the predominant form of Buddhism in East Asia. According to Kumārajīva's account, after Nāgārjuna had learned some basic Buddhist views, but without complete satisfaction, a "Mahānāga *bodhisattva*"—"a chief Nāgā [literally "serpent"; also the name of a hill tribe of South India], who was on the path to enlightenment"—took pity on him and presented him with the most profound Mahāyāna verses. Nāgārjuna mastered these in a short time and propagated the truth (*dharma*) in India, successfully defeating many opponents in scholastic philosophical debates. Traditional accounts also suggest that he lived to an old age and then decided to end his life.

Various elements of Nāgārjuna's biographies are supported by some historical materials. Present scholarship indicates that Nāgārjuna could have lived as early as AD 50 and as late as AD 280. A common consensus gives his dates as AD 150–250. The claim that he lived in South India is supported by some archaeological evidence: by a letter (*Suhṛllekha*, "Friendly Letter") credited to him, written to a king of the Sātavāhana dynasty, possibly Yajñaśrī (*c.* 173–202); and perhaps by his name, which includes the name of the Nāgā people, who lived in South India. While there is no evidence regarding Nāgārjuna's parents or his boyhood training, his knowledge of several Hindu philosophical positions expressed in his later critical expositions supports the tradition of a Brahmin family background. It is not known how widely accepted his critique of Hindu and early Buddhist philosophical-psychological interpretations was, but the fact that he had disciples who carried on his critique indicates some pervasive influence of his effort.

Influence and works

Nāgārjuna's influence has continued through the followers of the Mādhyamika school of Buddhism, whose critical examination of philosophical positions and whose didactic expositions have been studied down to the present day in many Far Eastern schools as part of the Chinese Buddhist Canon (*Ta-ts'ang Ching*). Similarly, 17 Mādhyamika treatises are contained in the Tanjur part of the Tibetan Buddhist Canon. Not all of these are credited to Nāgārjuna, and some traditionally credited to him are probably not by him. Two basic writings that are substantially his are available at present in Sanskrit; they are *Mūlamadhyamakakārikā* (*Mādhyamika Kārikā*, "Fundamentals of the Middle Way") and *Vigrahavyāvartanī* ("Averting the Arguments"), both critical analyses of false views about how existence arises, the means of knowledge, and the nature of reality. Three important Mādhymika texts attributed to Nāgārjuna and extant only in Chinese are *Ta-chih-tu-lun* (*Mahāprajñāpāramitā-śāstra*, "Great Perfection of Wisdom Treatise"), *Shi-chu-p'i-p'o-sha-lun* (*Daśabhūmi-vibhāṣā-śāstra*, "Illumination of the Ten Levels Treatise"), and *Shih-erh-men-lun* (*Dvādaśa-dvāra-*[*nikāya-*]*śāstra*, "Twelvefold Entrance Treatise"). The following three texts are found only in the Tibetan canon and are accepted by most scholars as compositions of Nāgārjuna: *Rigs pa drug ca paḥi tshig leḥur byas pa shes bya ba* (*Yukti-ṣaṣṭikā*, "Sixty Verses on Union"), *Ston pa ñid bdun cu paḥi tsig leḥur byas pa shes bya ba* (*Śūnyatā-saptati*, "Seventy Verses on Emptiness"), and *Shib mo rnam par ḥtag pa shes bya baḥi mdo* (*Vaidalya-sūtra*, "Sacred Text of the Vaidalya Type").

Besides the verses of Mādhyamika analysis there are a large number of Tantric (magical) and medical works attributed by Tibetan tradition to a "Nāgārjuna." There are also references in late Indian materials to a great *siddha*, or sorcerer, by the name of Nāgārjuna, who acquired his magical power through Tantric practices; *e.g.*, employment of magical spells and diagrams, and gaining access to nonphysical planes of existence through diet and meditation. Closely allied with these tales are stories of a powerful alchemist who, among other accomplishments, discovered the elixir of immortality. Due to the lack of concrete historical data, or because of conflicting accounts, the reports of a great sorcerer are generally not accepted outside the Tibetan tradition as applying to the 2nd-century philosopher.

Something of the Mādhyamika philosopher's life and attitude can be gleaned from his writings. His critical analytic verses, his didactic treatises, letters and hymns indicate his deep concern to practice "nonattachment" (*i.e.*, perceiving the emptiness of all things and hence becoming detached from them) in engagement with people. Through rigorous logical argumentation, as found in the *Mādhyamika Kārikā*, he criticized both Buddhist and Hindu views on existence. Most of his polemics, however,

were directed toward the explanations of existence offered by the Buddhist schools of Sthaviravāda and Sarvāstivāda. Nāgārjuna's position is closely allied to, and probably dependent on, that found in the early Mahāyāna literature known as the *Prajñāpāramitā-sūtras* ("Perfection of Wisdom Verses"), in which texts the notion of *śūnyatā* (emptiness, void) is an important term for the wayfarer on the path to enlightenment and becomes the distinguishing term in the Mādhyamika school. It indicates that the nature of existence is relational: there is no soul, no thing, no concept independent of its context; all things are empty of an absolute reality and exist only in relation to conditions. There is no eternal reality behind changing forms; even unconditioned Nirvāṇa (enlightenment) is not independent of the changing forms of existence. Those beings who have perfected wisdom perceive Nirvāṇa and the changing flux of existence together. This perfection transforms the other "perfections" that are part of the *bodhisattva*'s (Buddha-to-be's) course, such as morality, endurance, and meditation. The perfection of wisdom releases the ardent striver for perfection from even attachment to "perfection" that fails to perceive the relational character of the moment. From Nāgārjuna's point of view, then, participating in scholarly debates, writing explanations of the Buddha's teaching, and celebrating freedom from illusion and pain in hymns were considered to be consistent with the highest truth that does not separate wisdom from compassion.

BIBLIOGRAPHY. F.J. STRENG, "Life and Works of Nāgārjuna," in *Emptiness: A Study in Religious Meaning* (1967), an annotated bibliography of Western-language material, containing information on related subjects of Mādhyamika and Indian Māhāyana Buddhism; K.K. INADA, *Nāgārjuna: A Translation of His Mūlamadhyamakakārikā with an Introductory Essay* (1970), information on editions of Mādhyamika works in Sanskrit, Tibetan, Chinese, and Japanese, as well as on modern works in Western languages and Japanese.

For a discussion of the historical documentation on Nāgārjuna and an analysis of works credited to him, see K.V. RAMANAN (ed.), *Nāgārjuna's Philosophy as Presented in the Mahā-Prajñāpāramitā-śāstra* (1966); R.H. ROBINSON, *Early Mādhyamika in India and China* (1967); and T.R.V. MURTI, *The Central Philosophy of Buddhism* (1955).

(F.J.S.)

Nagoya

Nagoya, a city of Japan and the capital of Aichi Prefecture (*ken*), is one of that country's leading industrial centres and ports. It is located at the head of Ise-wan (Ise Bay) on the Pacific coastal strip of the Chūbu (midland) area of Honshu, 160 miles (260 kilometres) west of Tokyo and 90 miles east of Ōsaka. With an area of 126 square miles (326 square kilometres), Nagoya's population in the early 1970s was over 2,000,000, making it the fourth largest city of Japan. The city stretches for about 10 miles from east to west and approximately 14 miles north to south.

History. The history of old Nagoya began in 1610, when a great castle was erected by the first shogun of the Tokugawa family for his son Yoshinao as daimyo (feudal lord) of Owari, one of the three powerful collateral houses of the Tokugawa shogunate. Around the castle were houses of samurai (warriors). South of them, a new castle-town for commerce was built, to which the merchants and craftsmen of Kiyosu, three miles northwest of Nagoya, were ordered to move. From that time Nagoya grew as a castle-town of the Tokugawa family and became the greatest city in the Tokai region. By the end of the 17th century, Nagoya had a population of 100,000, and at the beginning of Meiji era (1868), it had increased to 130,000.

Beginnings as a castle-town

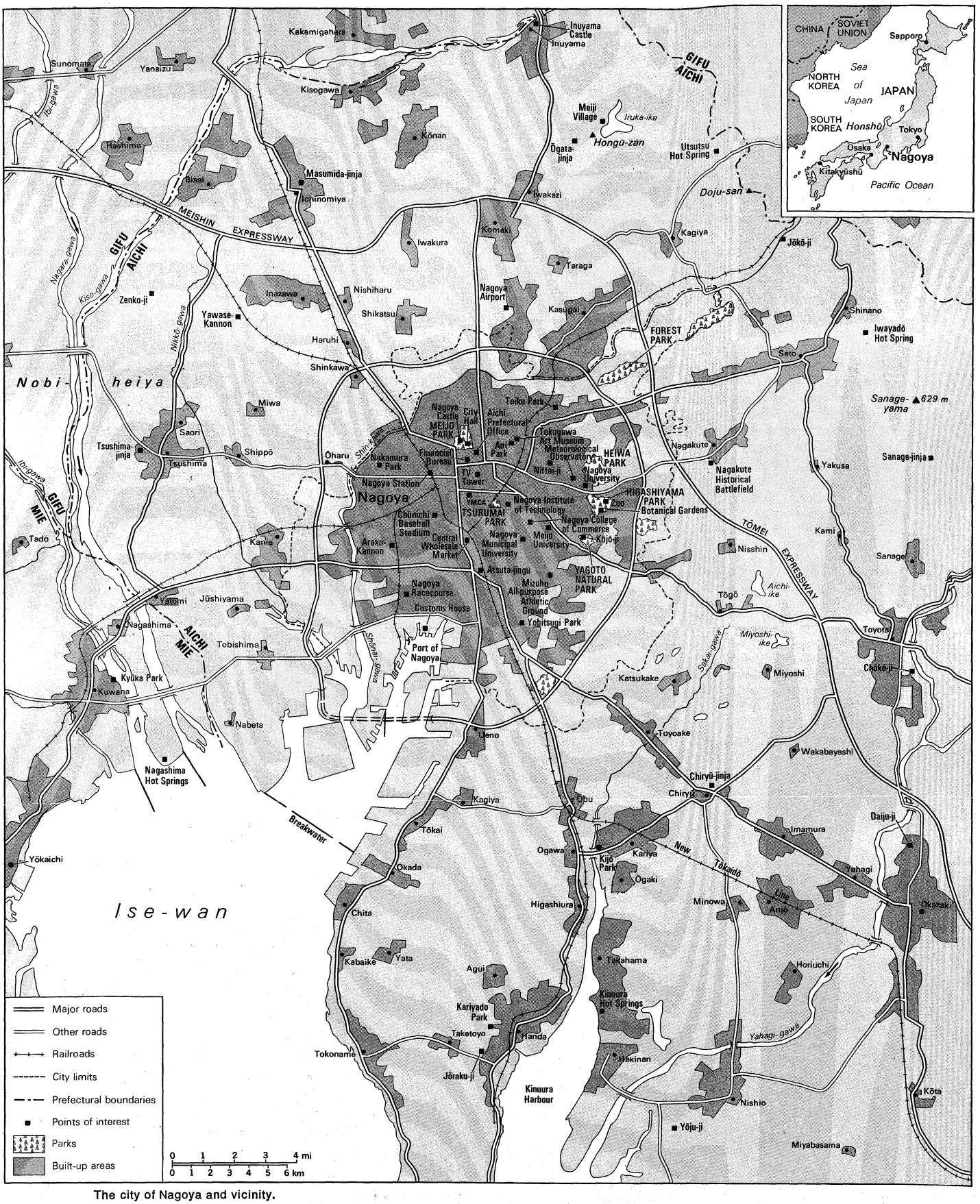
Atsuta, a part of modern Nagoya and to the south of old Nagoya, was an ancient political centre of Owari Province. Atsuta-jingū (Atsuta Shrine) and the Ise-jingū (Grand Shrine of Ise) are the oldest and most highly esteemed shrines in Japan. Especially in the Middle Ages, Atsuta was a famous stage town on the Tōkaidō Highway, the main overland transportation artery of Japan.

After the Meiji Restoration (1868), Nagoya experienced a recession resulting from the abolition of castle and stage towns. In the 1880s the economy began to develop, and many of the present factories, banks, and firms were established. In 1889 Nagoya became a municipality with a population of 157,000 and an area of five square miles; in the same year the National Railways' Tōkaidō Main Line was completed. The Kansai Main Line and the Chūō Main Line were constructed in 1895 and 1900, respectively, giving a further impetus to the city's expansion; and modern industries manufacturing textiles, ceramics, and timepieces, cement, machinery, woodwork, flour, and beer were established.

The area of the city grew larger through the consolidation of neighbouring towns and villages. In 1907 the entire southern region, including the port of Atsuta, whose initial phase of construction had been completed in the same year, was merged into the city. Heavy industry was developed after World War I. During World War II, Nagoya became the centre of Japan's aircraft

By courtesy of the International Society for Educational Information Tokyo, Inc.

Plaza and office buildings in downtown Nagoya.



The city of Nagoya and vicinity.

industry, and by 1941 the population of the city had reached 1,379,000.

Nagoya suffered serious damage from massive air raids in 1945. Nearly half of the entire city area was destroyed, and by the end of the war the population had been reduced to 500,000. In the postwar period, however, industry was re-established. Littoral districts were reclaimed on a large scale, and many large plants were built. The

population began to increase and by 1957 numbered about 1,400,000. With further consolidation of neighbouring cities, towns, and villages, the population grew to over 2,000,000 by 1970.

The "Central Capital"

The contemporary city. Nagoya, sometimes called Chū-kyō ("Central Capital"), lies between Tokyo and Ōsaka. The city developed on a low, level plain, the Nobi-heiya, that faces Ise-wan on the south and extends toward the Japanese Alps to the north. Most of the city is 26 feet to 34 feet above sea level, out of danger from floods. In the eastern part, much new residential construction has taken place; Owari Hill in the northeast rises to 650 feet (198 metres). Lower land in the west is less urbanized. The southern part of the city (former Atsuta), facing on Ise-wan and including the industrial area and harbour, is for the most part reclaimed land, much of it at or below sea level.

Climate. Nagoya's climate is generally temperate and is especially mild in spring and fall. The average annual temperature is about 59° F (15° C). In winter the temperature sometimes falls as low as 19° F (−7° C), and in summer it often rises to over 97° F (36° C). Nagoya averages 61 inches (1,525 millimetres) of annual precipitation, 75 percent of which falls in a seven-month period from April through October.

Environmental pollution in Nagoya has become a significant issue in regard to the people's health. Fish sometimes die from water pollution in the Kiso-gawa (Kiso River), which supplies the city's drinking water.

City plan. The planning area of Greater Nagoya encompasses 195 square miles (510 square kilometres). Nagoya and 16 neighbouring towns and villages are included in the area, which is divided into an urbanization promotion area (covering 93 percent of the total area) and an urbanization control area. Within the promotion area, definite plans for urban facilities, including parks, streets, and sewerage systems, have been formulated. In the control area, private construction is severely restricted.

Transportation. Nagoya Station is the centre of the transportation network, upon which the railways, the subway lines, and the bus lines that provide transport within the city all converge. A system of highways serves the metropolitan area, and two new high-speed city subway lines were completed in 1971. The city is served by the Japanese National Railways (JNR); the New Main (Trunk) Line between Ōsaka and Tokyo stops at Nagoya. A rail network runs northward to Toyama on the Sea of Japan and southward to the city of Ise and the Kishū region. The city is linked to other countries through its port, from which canals lead inland to the industrial areas, and by Nagoya Airport, on the north side of the city.

Population density and distribution

Demography. The population of the city of Nagoya in 1970 was 2,036,000. Population density was 16,193 persons per square mile (6,252 per square kilometre). There has been a large-scale emigration from the city to the suburbs and surrounding areas. Since the rate of increase of the population as a whole is diminishing, the population of the city is likely to increase only gradually, or to remain static.

Housing. The city has undertaken urban renewal programs and is enforcing strict adherence to the land-use plan. The construction of high-rise apartments is being promoted in the built-up areas, and large housing projects are being built on the fringes of the city.

Industrial production

Economic life. Production of timepieces, bicycles, and sewing machines began in the early stages of Nagoya's economic development, and textile and woodworking machinery industries also grew. The development of Nagoya's port, especially after World War II, and the advantages of the city's central location and abundant hydroelectric power from the rivers of central Honshū stimulated the development of heavy industry. Steel mills were constructed, which were followed by new plants turning out special steels, chemicals, oil, and petrochemicals, as the area's automobile, aviation, and shipbuilding industries flourished.

With the movement of population to the suburbs, many big stores (supermarkets) have been built. Since the mid-1960s, manufacturing plants have tended to move outside the city. In spite of its modern industrial development, cottage industries and traditional handicrafts continue to flourish in the area. Nagoya cloisonné and Seto pottery, to cite only two examples, are famous throughout Japan.

Accompanying the city's industrial growth has been its development as a commercial and financial centre.

Political and governmental institutions. The Nagoya City Assembly consists of 76 assemblymen, whose term of office is four years; they are elected directly by the citizens from 14 constituencies that correspond to administrative wards. The assembly is convened by an elected mayor, who exercises general supervision over the city as its administration chief and principal representative.

Public utilities. The city's water supply is obtained from three purification plants and the Aichi Prefectural Water System. The city's industrial water is taken from the Shōnai-gawa, from the treated water from the Chitose sewage-disposal plant, and from the Oharu purification plant; the prefecture's industrial water is from the Aichi Waterway.

Municipal sewage is handled by 12 sewage-disposal plants, two sludge-treatment plants, and 24 pumping stations. Electricity is supplied by Chūbu Electric Power Company.

Health and safety. The death rate decreased after World War II as a result of the improvement and expansion of public health services and facilities. Main causes of death are adult diseases and traffic accidents.

Many buildings, especially dwellings, are still constructed of wood. In spite of Nagoya's modern street plan, traffic congestion often delayed fire engines, and the Fire Bureau, therefore, increased the number of fire brigades and substations throughout the city. The increase of tall buildings and an underground shopping area have introduced new fire-fighting problems.

Nagoya is notorious for the frequent occurrence of traffic accidents. In traffic deaths it ranks first or second in the nation every year.

Education. Although there were only 260 schools and 192,000 students in 1935, there were 800 schools and more than 500,000 students in the 1970s. The number of students in higher education also increased dramatically: in the 1970s there were 55 colleges or universities in Aichi Prefecture, including 33 in Nagoya. Nagoya Medical College, founded in 1931 and based on a medical school of the Meiji era, became Nagoya Imperial University in 1939. After World War II it became Nagoya University. Other universities include Nagoya Institute of Technology, Aichi Prefectural University, and Nagoya Municipal University.

Historic sites and cultural institutions

Cultural life. Nagoya abounds in cultural assets and has many fine works of art and buildings of architectural value. One of them was Nagoya Castle, built in 1610–12, which was reduced to ashes in the fire raid in May 1945. In 1959 the castle was rebuilt, in ferroconcrete, and is now a historical museum and the city's most famous landmark. The Tokugawa Art Museum, opened in 1945, is the only museum that has preserved the collection of the Tokugawa family; it owns about 12,000 art treasures. It is located northeast of the city, adjacent to Hōsa Library and Tokugawa Park. The sacred Atsuta-jinjū is in the south of the city. The Nagoya Municipal Science Museum includes an astronomical building with a planetarium.

Nagoya Public Hall at Tsurumai Park is used for music concerts and similar functions. Citizen Hall, used for various kinds of entertainment, was completed in 1972. Other institutions include Aichi Cultural Center, Aichi Art Museum, Chūnichi Hall, and Misono Theatre.

The media. Nagoya is the home of one of Japan's foremost local newspapers, the *Chūnichi shimbun*. The *Asahi shimbun* and the *Mainichi shimbun* also have offices in the city. In the 1970s there were six television and four radio broadcasting companies. The Chūbu-Nippon Broadcasting Company, established in 1951, was the first private broadcasting company in Japan.

Recreation. Nagoya has about 300 parks, among which the Higashiyama Park, in the eastern hill district, is particularly noteworthy for its zoological and botanical gardens. After World War II, the Heiwa (Peace) Park was established nearby; it contains about 188,500 tombs from 280 temples in the downtown district.

Mizuho All-Purpose Athletic Ground, in the southeast of the city, has an athletic ground and facilities for many sports and games. Aichi Gymnasium and Nagoya Municipal Gymnasium are open to the public. Chūnichi Baseball Stadium, the home field of the locally popular Chūnichi Dragons, a professional baseball team, has a seating capacity of 30,000.

BIBLIOGRAPHY. *Nagoya City* (1971), provides information on the contemporary city; *City Planning for the City of Nagoya* (1962), deals with past city planning programs—both are issued by the city of Nagoya and are available in English. *The Outline of Municipal Government*, published annually by the city, describes its administration. There are two historical works (in Japanese) of significance—"The History of Nagoya City," 12 vol. (1916), which covers the city's history from ancient times to the Meiji era; and "The History of Nagoya City in the Taisho-Showa Era," 8 vol. (1946–53), edited by EIJIRO HONJO, an economic historian.

(T.Ma.)

Nairobi

Nairobi, the capital city of the Republic of Kenya, is located in the highlands of south central Kenya and is situated on the Athi Plains near the foot of the Gikuyu Escarpment. It lies about 300 miles (480 kilometres) northwest of the Indian Ocean port of Mombasa and about 160 miles due east of Lake Victoria. With a population of more than 500,000, it is one of the major cities of East Africa. In 1963 its area was increased from 35 to 264 square miles.

The city originated in the late 19th century as a colonial railway settlement, taking its name from a small stream known to the Masai people as Enkare Nairobi (Cold Water). The modern city includes within its limits Nairobi National Park—a 44-square-mile game reserve containing giraffes, gazelles, lions, zebras, and other animals. Nairobi is a centre for trade and transport, as well as for administration (for an associated physical feature, see EAST AFRICAN MOUNTAINS).

History. Nairobi was originally the site of a water hole used by Masai herdsmen, but a railway encampment established there in the late 1890s became the nucleus of the modern city. When the railhead arrived there in 1899, the British colonial capital of Ukamba Province was transferred from Machakos to the site, and Nairobi became the capital of the British East Africa Protectorate in 1905. From about 1900 onward, when a small Indian bazaar was established at Nairobi, the city was a trading centre.

From its modest beginnings, Nairobi's development has paralleled the transformation that Kenya has experienced in the 20th century. The city has attracted a stream of migration from rural Kenya that has made it one of the largest cities in tropical Africa. As early as 1901 a municipal committee was established to administer the affairs of the settlement. It was declared a municipality in 1919 and was then governed by a municipal corporation. In 1928 a town council was created with an elected mayor, deputy mayor, and other appointed officers. In 1954 Nairobi was granted city status. Kenya's 1963 constitution expanded the municipal area, as mentioned, from 35 to 264 square miles; the enlarged municipality, known as the Nairobi Area, is an independent unit administered by the Nairobi City Council.

The city's expansion

The rapid changes in the area and status of Nairobi were paralleled by its population growth. In 1919, when the city covered an area of about ten square miles, its population was about 10,000. The 1948 census reported the population to be almost 120,000. The present population, as mentioned, is more than a half million. The Nairobi greater planning area, furthermore, today extends beyond the city's boundaries into the adjoining districts of the neighbouring provinces.

The natural setting. Nairobi is located approximately 80 miles to the south of the Equator, at a height of about 5,000 feet. To the northwest rise the highlands that form part of the Aberdare Range, which is bounded farther to the west by the East African Rift Valley. To the southeast lie the Athi-Kapiti Plains, over which the city has spread. Much of the western half of the city is built on the northwest to southeast ridges more than 5,000 feet above sea level. In the extreme east the land drops to altitudes of between 4,000 and 5,000 feet.

Climate. Nairobi enjoys a stimulating climate that reflects the moderating influence of its high altitude. The noonday temperature rarely rises above 80° F (27° C). The mean annual rainfall varies from less than 30 inches in the drier and lower parts to well over 50 inches in the wetter fringes to the north.

The four seasons

Four seasons may be distinguished. The season from December to March is characterized by northeasterly winds accompanied by sunny and warm weather with occasional showers. From late March to May the rainy season occurs, accompanied by light and variable northeasterly to southeasterly winds. Occasional dry periods may also occur. From about June to October southeast moisture-laden monsoon winds predominate, bringing cloudy and overcast skies. Low temperatures prevail, cold may be experienced at higher altitudes, and fog and drizzle are common. A transitional season follows in November, when less rain falls.

Vegetation and animal life. The vegetation of Nairobi reflects the transitional nature of its location. The higher and western half of the city site, standing on rich volcanic soil, formerly had a forest cover that is still represented by a semideciduous type of vegetation. Elsewhere the most prevalent form of vegetation is the savanna type (park-like grassland) found on the plains, which are dotted by various species of thorn (acacia) trees.

Nairobi National Park

The grasslands and forest of the game reserve (Nairobi National Park) form an ideal habitat for some of East Africa's mammals. The herbivores (animals feeding on plants) that thrive there, and the predators that feed on them, such as the lion, leopard, and cheetah, are major attractions for visitors.

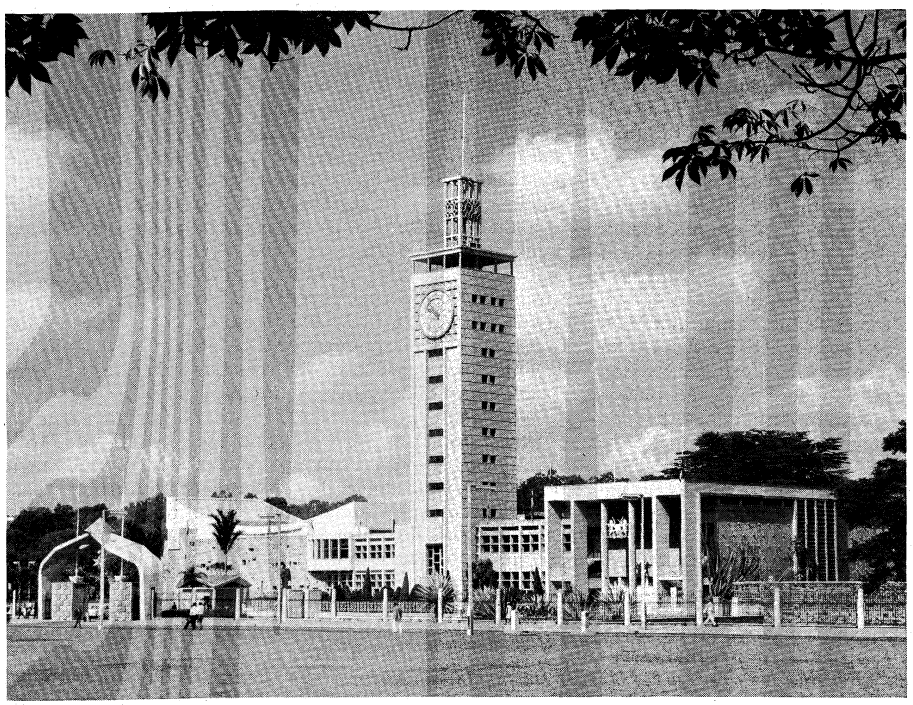
The contemporary city. *Demography.* Nairobi is one of the most rapidly growing urban centres in Africa. The 1969 census showed that the total population had increased from 344,000 in 1962 to 509,000. The annual growth rate between these years was almost 5 percent.

In 1948 the proportion of Africans to non-Africans was 54 percent to 46 percent. In 1962 the proportion of Nairobians within the city's pre-1963 boundaries had changed to 59 percent Africans and 41 percent non-Africans. The 1969 census showed that 83 percent of the population was African and 17 percent was non-African. Of the African population 97 percent was Kenyan. The majority of the non-Africans were aliens.

Ethnic groups

Among the Kenyan Africans the Kikuyu ethnic group accounted for 47 percent of the population, the Luhya 16 percent, and the Luo and Kamba each 15 percent. Various other ethnic groups constitute the remaining 8 percent.

Approximately 59 percent of the total population of the city in 1969 was concentrated in the central area, which included Eastlands, Eastleigh, and the census divisions of Parklands, Ngara, Juja, Racecourse, Mathare, and Central. Eastlands alone had approximately 24 percent of the total. This inequality in the distribution of population is reflected in the density pattern. The 1969 census showed an outer region of very low density, mostly under 1,300 persons per square mile; a medium to low density area extending roughly from the industrial area just south of the city centre to the extreme west of the city limits with densities ranging from 3,200 to 5,800 persons per square mile; and the central area, which contained well over half of the total population, with average densities that were usually well over 32,000 persons per square mile. The Racecourse, Bahati, and Jericho districts had 51,000 to 64,000 persons per square mile, and the housing estates of Maisha, Mbotela, Maka-



Parliament buildings, Nairobi, Kenya.
Authenticated News International

dara, and Maringo in Eastlands well over 64,000 persons per square mile.

The central part of the city and some of the main housing estates conform to the rectangular pattern with parallel streets and right angle intersections common to many Western cities. In other parts of the city this pattern has been considerably modified by the nature of the land. The pattern of the ridges in the extreme west and north determines the direction of the main streets that converge on the central core of the city.

Upper Nairobi, which lies to the west and north, includes most of the residential land. The Dagoretti area in the extreme west is a former rural part of Kiambu District. The remainder of the northern part of Nairobi is made up of a number of low-density housing estates that were formerly European residential areas. Close to the central part of the city are the Parklands and Eastleigh residential areas. These constituted one of the main Asian residential areas until about 1960; they still have a large proportion of Asians, as well as some African residents. Other Asian areas included what are now Nairobi West and the districts known as South C and South B.

Most of the city's population is to be found in the 15 housing estates located to the east of the city centre and the main industrial area to the south. The eastern part of the city, commonly known as Eastlands, was formerly reserved for Africans, and the majority of the African population still lives in this area. New housing estates continue to develop on the unused land that extends to Embakasi Airport, nine miles to the southeast.

The industrial district

A considerable area of flat land southeast of the city centre is reserved for industrial development. The industrial area lies between Eastlands and the main Nairobi–Mombasa road, which runs to the southeast, also extending southeastward to the Embakasi Airport. Part of the area is undeveloped land. It is provided with railway sidings and access roads. Considerable areas of agricultural land also lie within the city's boundaries: in the east, north of the airport, are large ranches; in the extreme north are rich farmlands that have been invaded by built-up areas.

The city centre is a well-defined area bounded by Uhuru (Freedom) Highway to the west, the railway station to the south, and Government Road to the north. In the centre is the city square with the principal public buildings and ministerial offices. To the north and east are two of the main retail shopping areas.

Housing and architectural features. Most city dwellers are accommodated in rented houses or in those supplied by employers. The city council, the government, and various large employers have played an important part in the provision of accommodations. The city has been fortunate in the availability of excellent local building stone. Most of the houses are built of stone or brick, and the majority have running water and toilets. Due to the influx of population, however, the city is far from meeting the total demand for housing.

Architectural landmarks

Among the city's architectural landmarks in the central area are the Parliament Building, the law courts, the city hall, the ministerial offices (including the office of the president), and the Roman Catholic cathedral. There is also a well-planned commercial centre.

Economic life. As well as being pre-eminent in the economic life of Kenya and the seat of national government, Nairobi also plays a significant role in the community of East African states. It is the headquarters of the East African Community's railways, harbours, and airways corporations. In addition, the city accommodates some United Nations regional services, as well as a number of foreign diplomatic missions.

Public services employ the largest proportion of wage earners (28 percent of the city total), followed by commerce (19 percent) and industry (18 percent). Nairobi is the principal industrial centre of the country. The railway workshop is the largest single industrial employer. Light manufacturing industries produce beverages, cigarettes, and processed foods. The city's commercial and industrial pre-eminence exerts a strong attraction upon the population in other parts of Kenya and beyond. Nairobi has modern banking facilities, including the National Bank of Kenya Ltd. and a stock exchange.

Transport and communication. Located to the southeast of East Africa's agricultural heartland, Nairobi is connected to Mombasa on the coast as well as to Uganda in the interior by the East African Railways Corporation. Paved all-weather roads also run to Kisumu on Lake Victoria in west Kenya, northward to the Gikumbo (Mt. Kenya) region, southeastward to Mombasa, and northwestward to Uganda. A paved highway to Namanga and Arusha in Tanzania is under construction. The increasing volume of both local and through traffic poses a constant planning problem. Nairobi has also become a focus for Eastern Africa air traffic. Embakasi is one of the principal international airports in Africa.

Governmental institutions. As the national capital, Nairobi has come to dominate much of the political activity in the rest of the country. In addition to parliamentary institutions, it contains the headquarters of the ruling party, the Kenya African National Union (KANU). The city government is composed of a mayor, aldermen, and councillors.

Power generation

The city possesses a modern piped water supply and has a waterborne sanitation system that is in advance of many other African cities. Electricity is obtained from Uganda's hydroelectric source on the Nile at Jinja and is conveyed to the city by high-tension transmission lines; the East African Power and Lighting Company, Ltd. is the distributing agency. The increasing electrical needs of Nairobi are also being supplied by the Upper Tana Hydro-Electric Scheme in southeastern Kenya.

In the planning of the city, emphasis has been placed on maintaining a high standard in public health. The sewer system is being extended, but many homes are still dependent on the use of septic tanks.

The city population is served by the Kenyatta National Hospital, which also serves as the teaching hospital for the National University in Nairobi. There are three other large independent hospitals, as well as a number of smaller private nursing homes.

There is a well-equipped central fire station. Law and order are maintained by the Nairobi Area Police Division, which is supported by a mechanized unit that assists in the patrolling of the expanding urban area.

Education. The rapid rate of growth in the city has demanded constant expansion of educational facilities. There are a large number of primary schools; the schools that formerly enrolled only non-Africans are now integrated into the national system and provide education for all children. The total number of enrolled primary school pupils in 1969 was 56,000, of which 30,000 were boys and 26,000 were girls. In addition, there are schools for handicapped children. There are also a number of well-staffed secondary schools. While some privately run secondary schools still exist, most now form part of the integrated system of education. The National University, of which Kenyatta College is a constituent part, is also located in the city, as is the Kenya Polytechnic and Kabete Technical and Trade School.

The National University

Cultural and other services. The University College Library, the Kenya National Archives, and McMillan Memorial Library are located in the centre of the city. A museum close to the centre accommodates treasures of Kenya's past. Also in the area are the cultural centre, the main campus of the university, the Kenya National Theatre, the East African Conservatoire of Music, and the Sorsbie Art Gallery.

Mass-communication media include two English and one Swahili daily newspaper and two Swahili weeklies. The Voice of Kenya radio and television services—broadcast in Swahili, English, and vernacular languages—are administered by the Ministry of Information and Broadcasting.

BIBLIOGRAPHY. W.T.W. MORGAN (ed.), *Nairobi City and Region* (1967), a comprehensive collection of essays; R.W. WALMSLEY, *Nairobi: The Geography of a New City* (1957), an account of the capital following its elevation to city status; W.T.W. MORGAN, *The Location of Nairobi* (1968), S.H. OMINDE, "Internal Migration of Economically Active Age Groups in Kenya," and R.B. OGENDO, "Industrial Concentration and Diversification in Kenya," *East African Studies* (1968), works on aspects of the city's growth; UNITED NATIONS, *Report of Mission to Kenya on Housing* (1965) includes consideration of the housing problem in Nairobi; KENYA, ECONOMICS AND STATISTICS DIVISION, *Kenya Population Census, 1969*, vol. 1 (1970), gives demographic details on Nairobi, including a statistical breakdown by tribe, nationality, sex, and age.

(S.H.O.)

Names

A name is a word or group of words used to refer to an individual entity (real or imaginary); the name singles out this entity by directly pointing to it, not by specifying it as a member of a class. For instance, it is possible to refer to the same entity, in this case a river, in two distinct ways: (1) "The Colorado is a beautiful river," and (2) "The river that flows through Austin is beautiful." Because there is only one river that flows through Austin, Texas, the subject of sentence (2) is unambiguously identified, and the reference of the sentence is fully individual. The subject of sentence (2), however, is not a name but rather a nominal (noun) phrase that specifies one member of the whole class of rivers by indicating a unique property of it. The word Colorado in sentence (1), on the other hand, is a name, because it directly points to the specific river. The fact that there is more than one river called Colorado, and that more specific information is sometimes needed to identify the one being discussed (*e.g.*, "I prefer the Texan Colorado to the California one"), does not change the status of "Colorado" as a name, because each of the two rivers is referred to in the way required by the definition.

Names and appellatives. A general appellative (*i.e.*, a common noun) capable of being used in reference to a whole class of entities can also be used with an individual reference. For instance, if an inhabitant of Austin, Texas, says, "Let's go swimming today, not in the pool but in the river," there is no doubt that the word river has a unique, individual reference to one single river, namely, the Colorado. This fact, however, does not make a name out of it; "river" is here a common noun, but its reference is specified by the extralinguistic context of the situation in which the sentence was said. Some names seem to belong more to the category of appellatives, or common nouns, than to the category of names like Colorado in "the Colorado River." For example, names like Big River, Red River, Stony Brook, and Cedar Hill may sometimes have their origin in a specific use of a general noun. If a sentence like "After five days of marching, we had to cross a river, the big one, not one of the smaller ones" is used very often, the name Big River may result. Such names are, however, much more frequently given as directly descriptive names. The similarity of names of this type with expressions like those exemplified in sentence (2) above is deceptive. There is, after all, more than one big river, so that the specification "the big river" is not complete. The full identification of one single river as the reference of this sentence is given by the situational context. Therefore, apart from certain special expressions (like "the big one, not one of the smaller ones"), names like Big River, Red River, and so on have the same status as names like Colorado.

Formal means for marking names

In some languages, a name is differentiated from an appellative (common noun) by formal means. The difference is sometimes indicated by the script; *e.g.*, languages using alphabets such as the Latin, Greek, Cyrillic, Armenian, and Georgian use a capital letter at the beginning of a name. (But, on the contrary, in German all nouns, not only names, are written with an initial capital.) There are examples of a purely grammatical differentiation of names as well, such as the usual absence of the articles "a" or "an" in English; *e.g.*, "Yesterday I saw an archer practicing his art" and "Yesterday I saw (Bill) Archer practicing his art."

The distinction between names and appellatives (common nouns) is generally clear: names are used in individual reference, appellatives can be used in reference to all members of a class or to any number of them (*e.g.*, river, hill, man, girl, car, table, virtue, and so on). Nevertheless, there are some borderline cases. For instance, a nation can be conceived as an individualized entity, so that "Americans," "Englishmen," and "Spaniards" are names; on the other hand, it is clear that other groups of people are not conceived in this way, so that expressions like "soldiers," "sailors," and "clergy" are not names. It is difficult to decide on the status of expressions like the "Baptists," "Adventists," and "spiritists." In a similar way, if all vehicles produced by Henry Ford are Fords and if one can buy an individual Ford as well, is "Ford" a name? It probably is, or approaches that status; but names of this type frequently lose the character of names and develop into common nouns. Expressions like "the Roman Catholic Church," "the Ministry of Education" (of a specific state) also have a dubious position as to

their status as real names. The uncertainty in this respect is indicated by the vacillation in the use of capital letters in various languages. This overlapping has a long history and is reflected in modern terminology. The Greeks used the term noun (*onoma*) for both the common noun and the name; when they wished to make a distinction, they specified the name as a proper noun (*onoma kyrion*). It is in this tradition that the term proper noun, or proper name, is used for a name, and noun, general noun, or common noun, for an appellative.

THE SCIENCE OF ONOMASTICS

Categories of names. The science that studies names in all their aspects is called onomastics (or onomatology—an obsolete word). The subject of this science is broad because almost everything can have a name and because the study of names theoretically encompasses all languages, all geographical and cultural regions, and all historical epochs. For practical purposes, some divisions of the subject are necessary; *e.g.*, by language (as the study of Kiowa or Provençal names) or by geographical, historical, or similar partitions (the study of the names in India, of the Levant at the time of the Crusades, and so forth). Another division (usually combined with the preceding ones) is given by the character of the names themselves; in a very broad categorization, names of persons, or personal names, are discerned on the one hand, and names of places, or place-names, on the other. In the most precise terminology, a set of personal names is called anthroponymy and their study is called anthroponomastics. A set of place-names is called toponymy, and their study is called toponomastics. In a looser usage, however, the term onomastics is used for personal names and their study, and the term toponymy is used for place-names and their study. The term toponymy itself can be understood in two ways, even in the exact terminology: either it is taken in the broadest possible way as including inhabited places, buildings, roads, countries, mountains, rivers, lakes, oceans, stars, and so on, or it is restricted to inhabited places (cities, towns, villages, hamlets). If the latter alternative is the understanding of the term toponymy, then the uninhabited places (*e.g.*, fields, small parts of forests) are called microtoponymy; names of streets, roads, and the like are called hodonymy; names of bodies of water, hydronymy; names of mountains, oronymy. Additional terms are not generally used (though one occasionally hears words like chrematonymy—names of things).

Divisions within onomastics

In any case, different categories of names frequently must be studied together, because there are transitions. For instance, many place-names are derived from personal names (*e.g.*, Washington), many names of planets and stars are derived from the names of mythological characters (*e.g.*, Venus, Mars, Alpha Centauri), and many personal names are derived from place-names, names of nations, and other such names (*e.g.*, Austerlitz, Napoleon's battlefield; French; Scott). There is also a division of names into primary and secondary ones. Neptune is primarily the name of a Roman god; transferred to the name of a planet, it is a secondary name.

Forms of personal names. There are many subdivisions and terms within the category of personal names. Originally, one name was given to a person at an early period of life—in Europe (and later in America), normally at Baptism. This is called either simply the name, the baptismal or Christian name, or the forename; in the U.S. and Canada it is usually called the first or given name. Because many people received the same name (given name), they were differentiated by surnames (of the type John Redhead, John Hunter, John Scott). Many of these surnames became fixed and hereditary in individual families. These are called either surnames or family names, and in the U.S. and Canada they are frequently known as last names. Thus the basic pattern is given + family name, together called the name or the personal name. There are exceptions concerning this sequence. Among the Chinese and Hungarians, for example, the family name precedes the given name: Mao Tse-tung, Nagy István. The Hungarians usually switch the order when they write English; thus, Nagy István becomes István (or Stephen) Nagy. The Chinese, however, maintain the order of family name first.

Surnames and family names

There are variations in the basic pattern. In the U.S. and Canada the usual practice is to insert another name (frequently expressed in writing only by the initial letter) between the given and the family name. This is the second, or middle, name. It may be the original family name of a married woman inserted between the first and the last name of her husband, the maiden name of one's mother, the family name of a girl retained as a second name after her marriage, as well as other names. In Europe, such a second name is less common and is usually acquired at Baptism (or, eventually, at Confirmation). In most European countries, the first baptismal name is the important one, and the second one (third, and so forth) can be omitted. In German usage, however, the baptismal name immediately preceding the family name is the most important one. For example, if one of the baptismal names in Johann Sebastian Bach or Johann Wolfgang von Goethe is to be omitted, it would be Johann. (But in a sequence like Johann Nepomucenus Nestroy, the shorter form is Johann Nestroy, because Nepomucenus is only an attribute discerning one of the numerous saints who had the name Johann.) British usage varies in this respect, but sometimes follows the German pattern—*e.g.*, W. Sidney Allen.

In a few areas, above all in Iceland and the Soviet Union, the so-called patronymic (*i.e.*, a name derived from the given name of the father) is inserted between the given and the family name. In Russian, if the father's name is Ivan Krylov, then the son's name, for example, will be Pyotr (given) Ivanovich (patronymic) Krylov (family), and the daughter's name will be, for example, Varvara Ivanovna Krylova. The usual form of address in Russian—among acquaintances, neighbours, colleagues at work, and inclusive superiors—is by the given name and the patronymic.

Use of patronymics

The terms maiden name and girl's name are sometimes used for the original family name of a married woman. Nickname is used in reference to surnames (which have not developed into family names), mainly of the jocose type; *e.g.*, a thickish Mr. John Smith is called Fatty. A surname, also called a byname or to-name (obsolete), can be used to differentiate persons with the same family names if they belong to different families and if given names are not used among them. In a village, there may be several families with the name Jones; if they are not called or referred to by first names, they may be known as Jones at the Pond, Jones the Redhead, and so forth. Hypocoristic forms of names are those that are used in familiar, friendly, or intimate situations (usually shortened or otherwise modified); *e.g.*, Tom for Thomas, Jim for James. Some of these forms are also used as given names, particularly in the U.S.

THE NAMING PROCESS

One of the most important elements of the naming process concerns the meaning and associations of the name. In this case the term meaning is radically different from that in the case of common nouns, in which the "meaning" is their ability to be used in reference to a class of entities, to denote or designate them. As was noted above, the absence of this ability to refer to a class of entities is typical of a name. If the meaning of a typical common noun, such as automobile, is considered, it can be seen that it denotes a certain type of vehicle. On the other hand, if the word automobile itself is considered, one can see that it consists of a Greek element, *auto* "self," and a Latin one, *mobilis* "movable," so that the sum of the meanings of the constituent parts of the word suggests a gloss like "self-movable," "self-mover." The meaning of a name involves that which the constituent parts suggest. In this sense, the "meaning" of a name like Red River is obvious. To get a "meaning" of a name like Philip, however, one must go back to its original Greek version, Philippos, which means "lover of horses." This "meaning" of names frequently gets lost, however. There are several causes for this, one being that the name may be

Loss of original meaning of names

accepted into another language, as were the Indian place-names in America (*e.g.*, Oshkosh, Chicago, Kankakee), and the Greek and other names transferred to Europe and America via Christianity. In addition, names may cease to be understood as a result of language change; *e.g.*, the place-name Birmingham was understandable in Old English as "habitation of Biorma's people," and the originally Germanic name Gerard was once understood as "strong spear" (*Ger-hardo*). Names also changed by shortening (*e.g.*, Los Angeles, from El Pueblo de la Reina de los Angeles de la Porciúncula "Town of the Queen of Angels of Porciuncula," one of the titles of the Virgin Mary in the Roman Catholic Church) and by scribal error (*e.g.*, Pria in France, a misread medieval abbreviation of Pradaria, "meadow"). Another cause of the loss of meaning in names is that the "meaning" simply fades out by constant use of the word as a name. No one thinks of the meaning "ford for oxen" when speaking about Oxford, and no one realizes a discrepancy if Mr. White has a dark complexion. Finally, it sometimes happens that a name has no "meaning" from the beginning. For instance, the place-name Tonolo and the family name Bréal were created from random sequences of sounds.

Choice of personal names. Names that have no "meaning" (above all not for the person who chooses the name) still can have associations. Although "Mary" and "John" may have no specific "meaning," they were the names of important persons in the Christian religion and therefore have been used very frequently. An association may be so strong that it overwhelms the "meaning" of a name, even a disagreeable meaning; *e.g.*, the association with the cult of St. Demetrios made the name Demetrios one of the most popular in the Greek Orthodox Church, though its meaning is "belonging to [the pagan goddess] Demeter." On the other hand, such an association may more or less completely fade out and be combined with or replaced by other associations, such as a national tradition (Patrick in Ireland, Yves in Brittany, István in Hungary, Ivan in Russia) or with a family tradition (Louis in the Bourbon family, Wilhelm among the Hohenzollerns, Henry in the Ford family). On a less elevated level, there is the example of a rich uncle making a given name more than eligible. A name can be associated, correctly or not, with various prestige factors, or its choice may be influenced simply by fashion. Another source of names, often extraordinary ones, was the occasional habit in Roman Catholic countries of giving a child the name of the patron saint whose day of celebration coincides with the child's day of birth (or Baptism); many names like Hyacinthus X, or Narcissus Y, were produced in this way.

"Good" and "bad" names

In the majority of cases, children are given "good," likable, and propitious names. In some cultures (*e.g.*, in some parts of sub-Saharan Africa, formerly in China, and sporadically in ancient Greece), however, the children are (or were) sometimes given "bad" names with meanings like "ugly," "disagreeable," or "crippled." The purpose of such names, which are called apotropaic names, is to make the child undesirable to demons.

The choosing of a given name is a highly private and individual matter. All the circumstances just mentioned can be motives for the choice, in addition to many other personal reasons, such as a consideration for the relatives' names or a simple liking of the phonetic shape of a given name. This wish to give a likable name may go so far that a sequence of sounds is chosen that sounds pleasant to the person who makes the choice but that has no relation to the existing stock of names or to the words of the language; *e.g.*, "Gwyned" was invented as a name of a girl and has no "meaning" or associations. This phenomenon is relatively common in the U.S.

Choice of place-names. Place-names are less personal, less intimate, and a matter of public concern. The usual pattern is that the national Ministry of the Interior (or its equivalent) keeps an official list of place-names, particularly of place-names that form administrative units, together with lists of districts, counties, and the like. This function may also be performed by the ministry or agency that supervises the postal service. Bodies endowed with authority provide or choose new place-names if there is a need to create them on a greater scale; *e.g.*, the U.S. Board on Geographic Names.

International cooperation (performed above all by the Universal Postal Union) is necessary, because names of identical places may vary from language to language. Particularly difficult are place-names originally written in scripts other than the Latin (Roman) one (the official script of the Universal Postal Union), such as Arabic, Chinese, Cyrillic, and the Indic writing systems. But even within the Latin script, there are two basic types of difficulties. First, one place can have different names (or modifications of a name) in various languages—*e.g.*, French Nice and Italian Nizza; German München, English and French Munich, and Italian Monaco. A very difficult situation arises when a place is generally better known by its international name than by its original one; *e.g.*, Dublin is Baile Átha Cliath in Irish. Confusion may also be caused by names that are translated; *e.g.*, the Rocky Mountains are in German Felsengebirge and in Russian Skalistye Gory. There are also names with the same written form but with varying pronunciations; *e.g.*, for Paris, the accent, pronunciation of vowels, and pronunciation of consonants change from French to German to English.

The second difficulty involves the actual printing of all the letters with diacritical marks that are necessary for different alphabets of the Latin script. Because many printing firms lack the various marks, some possibly confusing omissions or modifications can hardly be avoided; *e.g.*, the dot over the *I* in Turkish İstanbul, and the bar through the *l* in Polish Kołobrzeg are frequently omitted. International cooperation is also necessary and is developing in connection with the choice of place-names in outer space, particularly on the surface of the moon.

HISTORICAL AND CROSS-CULTURAL DEVELOPMENT OF NAMES

Legal aspects of naming. While place-names are considered a public matter, personal names also seem to be getting more regimented by various laws and regulations. Great Britain and the United States are practically the only countries that adhere to the principle of Roman law that a person has the right to use and change his name as he pleases, except for fraudulent purposes. The first important regulation concerning given names was the decision of the Council of Trent (1563), which specified that the Roman Catholic priest administering Baptism should make certain that children are given names of Catholic saints; if the parents were to insist on another name, the priest should administer Baptism in that name but add the name of a saint as the second baptismal name. This regulation, still a valid part of Canon Law, was directed against the Protestant custom (spreading at that time) of giving children names of important persons from the Old Testament otherwise unconnected with Christianity (*e.g.*, Abraham, Samuel, Rachel). In this respect the regulation was successful in Catholic countries, but it did not succeed in stopping the use of given (baptismal) names like Cesare in Italy (from Latin Caesar).

Naming regulations of the Council of Trent

The next important law was passed in France. The French Revolution first gave complete freedom in naming; the result was some very fanciful given names like Mort aux Aristocrates, Racine de la Liberté, or even Café Billard. To stop this, a law was passed in 1803 that restricted given names to "names of persons known from ancient history" and "names used in various calendars." Again, the law was successful in its main intention; in addition, it prevented the spread of controversial given names such as Marat and Robespierre and of literary names such as Aramis, d'Artagnan, and Romeo. Very reasonably, the law never was interpreted too narrowly, so that feminine given names such as Jeanette and Henriette, for example, have been admitted, though they were not legal because no calendar contains them. This law is still valid in France.

Similar laws were passed, at various times, in the states of eastern Europe, where the given name can be chosen only from names known and established as such, the ex-

act formulation varying from country to country. One of the results of this legislation is that the repertory of Russian given names today is more or less exactly the same as that of 60 years ago, when the Orthodox Church was paramount. In non-Russian areas of the Soviet Union, Catholic (in Lithuania), Muslim (in the Uzbek S.S.R. and Tadzik S.S.R., etc.), and other names are used, sometimes without any ecclesiastical association whatsoever. In the Caucasus, for example, there are given names such as Soslan and Dzerassa, drawn from Caucasian mythology.

In regard to family names, the most important regulation was made at the Council of Trent (1563); it was decreed that every parish must keep complete registers of baptisms, with the names of the child and those of his parents and grandparents. This had been done before, but not so systematically. The new practice (soon followed in Protestant parishes) helped to establish the family names. There is not much legislation concerning family names, because two basic assumptions are taken for granted: that the bride will accept the bridegroom's family name by marriage and that their children will automatically have the family name of the parents. Legal institutions like the West German one permitting the bride to add her original family name to her new one in a hyphenated form (Inka Schmidt, when married to Karl Neumann, may become Inka Neumann-Schmidt), or the practice of outstanding British families to combine the family names of the married couple in a hyphenated form (Beatrix Curzon and Frederic Cholmondeley become Beatrix Cholmondeley-Curzon and Frederic Cholmondeley-Curzon) are rather rare.

Assumptions concerning family names

In the majority of cases, the law is concerned with names mainly in cases of divorce, adoption, and illegitimacy. After a divorce, the wife is usually allowed to eventually reassume her maiden name, and in Germany, for example, can be forced to do so if judged as the guilty party and if her former husband so desires. In adoption procedures, either the family name of the adopting person is accepted or a hyphenated form is created. Illegitimate children usually receive the family name of their mother.

In the eastern European states, in which legislation changed the basic assumptions concerning the family name, a different and unusual situation has developed. When a Czech girl, Anna Klímová, for example, marries a Josef Novák, both may retain their original family names, or the girl may become Anna Nováková, or, more remarkably, the man may become Josef Klíma, accepting the bride's family name. This must be decided by mutual consensus, and such a consensus is also the basis for the children's names. The purported reason for this legislation is the full equality of women. (There is, however, one loophole in the system, namely, the Russian patronymic: this is automatically derived from the father's name, whereas equality understood in this way would demand a choice between the father's or the mother's name.)

The development of personal names. *European patterns of naming.* The development of personal names was a complicated one. In the old Indo-European system, a person had one name, which could be of two types: a compound substantive or a non-compound one. Noncompound names may originally have been given to less outstanding members of the tribe and their children. The compound names frequently associate the bearer with a god (they are called theophoric names) or state his virtues, abilities, skills, possessions, and so forth. The association of the meanings of the parts of the compound was sometimes only a loose one, as is particularly observable in German anthroponymy (see below). Examples of compound names include the Sanskrit Viṣṇuputra "son of Viṣṇu," Devadatta "given by god," Devarāja "god-king." From old Iran come the Avestan name Hōrmizāfrīd "benediction of Ahura Mazdā" and the Old Persian name Mithradates "given by Mithra" (two Iranian gods). Among Greek names there are also many theophoric names, such as Herodotos "given by Hera," Isidoros "given by Isis" (modern Isidore), or Theodoros (modern Theodore) and Dorotheos (modern only in the feminine form, Dorothy), both "given by god." There are many other similar Greek names; *e.g.*, Astyanax "lord of the city," Perikles "very famous," and Demosthenes "strength of the people." Platon "broad (in shoulders)" is a noncompound Greek name.

Compound and noncompound names

The compound names of the Celts include Vercingetorix "great king of warriors," Orgetorix "king of killers," Rextugenos "son of justice." Noncompound Celtic names were, for example, Artos "bear" and Galba "big." Examples of Germanic compound names include Heriberhto "army + resplendent" (modern Herbert), Huguberhto "resplendent by thought" (modern Hubert), Godofrido "divine peace" (modern Gottfried and Friedrich and Geoffrey), Frideriko "peace + powerful" (modern Frederic), and Theodobaldo "people + valiant." Among the noncompound Germanic names is Karl (or Charles), in the Latin form Carolus, meaning "man." Vladimir "governs the world," Vladislav "rule + glory," and Miroslav "world + glory" are typical Slavic compound names.

The Latin system of personal names was quite different and probably developed under Etruscan influence. In the oldest times the Romans seemingly had only one name; *e.g.*, Romulus, Remus, Manius. From the beginning of historical times, however, the Roman personal name consisted of a praenomen (given name, forename) and a nomen (or *nomen gentile*). The use of the praenomen was limited to intimate contacts, and its choice was strictly restricted to fewer than 20 names, among them Gaius, Gnaeus, Marcus, Quintus, Publius, Tiberius, and Titus. The subsequent nomen was hereditary in each gens (a related group of families, like the Scottish clan); examples include Antonius, Aurelius, Claudius, Cornelius, Fabius, Horatius, Julius, Lucius, Maccius, Tullius, and some others. Because the choice of both the praenomen and the nomen was restricted, the patrician families and later all families started using a hereditary name, called a cognomen. These cognomina developed from original surnames; *e.g.*, Cicero "bean," Pictor "painter," Plautus "flat foot," Tacitus "silent." In this way, the Roman name consisted of three parts: Marcus Tullius Cicero, Gaius Julius Caesar. In addition, a person might have acquired an individual surname, called an agnomen: Publius Cornelius Scipio Africanus was so named because of his successful war in Africa.

Roman system: praenomen, nomen, and cognomen

This system of naming was used during the whole republican epoch and later in the empire. Toward the end of the empire, however, the naming pattern began to change and subsequently was lost. One reason was that more persons used names lacking any real relation to themselves. For instance, a slave (and then his children) used the praenomen and the nomen of the master who set him free; *e.g.*, had Marcus Tullius Cicero freed a Syrian slave, the name of the latter might have been Marcus Tullius Syrus. The number of affranchised slaves grew constantly, particularly after the victory of the Christian religion. Also, a free inhabitant of the empire who was granted Roman citizenship acquired the praenomen and nomen of the magistrate who made the decision. Thus, in AD 212 all free noncitizens were given citizenship by the emperor Caracalla, with the result that hundreds of thousands of persons prefixed Marcus Aurelius before their name (be it Greek, Syrian, African, or any other). In this way, Roman names lost their significance. Another influence was introduced by the Christians, who belonged to social classes that were not particularly concerned with the habits of the Roman higher class and who preferred names connected with their own religion; *e.g.*, from its founders (Petrus, Paulus, Joannes, Maria, Timotheus) or from the new martyrs, frequently persons with simple Latin or Greek surname-like names (such as Stephanos "wreath" [modern Stephen], Laurentius "laurel" [modern Lawrence], and Sidonius "coming from Sidon in Phoenicia" [modern Sidney]). Simple names like these were sometimes called *signum*. The Christians, however, soon started creating their own names; *e.g.*, Benedictus "blessed," Desiderius "desiring (salvation)," Renatus "reborn (by Baptism)," modern René.

Christian and non-Christian given names

With the spread of Christianity, this stock of names spread into territories that did not belong to the Roman Empire, but the diffusion was slow. In both the Germanic and the Slavic sphere (half of which came under the influence of the Eastern Church), the use of many of the original non-Christian names was continued, partly by tradition and partly because some of the bearers of these names became saints themselves. In this way, the repertory of given names was set, in general, somewhere around the 12th century. A notable addition to it was the influx of Old Testament names brought by the Reformation (Adlai, Benjamin, and so on). Certain names have left no trace of their ephemeral existence—*e.g.*, Puritan names such as Humility, Be Faithful, Kill Sin; French revolutionary names; and Russian postrevolutionary names such as Mels (an acronym containing the initial letters of Marx, Engels, Lenin, and Stalin). American fanciful given names for girls, such as Claretta, Elizene, Gwyned, and Marilla, are also relatively insignificant in impact, though the group is growing.

Family names. Family names came into use in the later Middle Ages (beginning roughly in the 11th century); the process was completed by the end of the 16th century. The use of family names seems to have originated in aristocratic families and in big cities, where they developed from original individual surnames when the latter became hereditary. Whereas a surname varies from father to son, and can even be changed within the lifespan of a person, a hereditary surname that develops into a family name better preserves the continuation of the family, be it for prestige or for the easier handling of official property records and other matters. Family names frequently developed (via surnames) from hypocoristic forms of given names; *e.g.*, from Henry came Harry, Harris, Hal, Halkin; from Gilbert came Gibbs, Gibbons, Gibbin, Gipps, Gilbye, Gilpin; and from Gregory there developed Gregg, Grigg, Greggs, Griggs, Greig. Other sources of family names are original nicknames—Biggs, Little, Grant (grand, large), Greathead, Cruikshank, Beaver, Hogg, Partridge: occupations—Archer, Clark, Clerk, Clarkson (son of a clerk), Bond, Bonds, Bound, Bundy (bondman); and place-names—Wallace (man from Wales), Allington, Murray, Hardes, Whitney (places in England), Fields, Holmes, Brookes, Woods (from microtoponyms).

A great number of family names come from patronymic surnames; in English, they are usually formed by the suffixation of "-son." Patronymic surnames can be formed from the father's given name or from any of its variants. Therefore, there is not only the form Richardson, but also Dickson, Dixon, Dickinson; and Henryson, Harrison, Henderson; Gilbertson, Gibson; and Gregson, Grigson. Some English patronymics, particularly in old families, are formed with a prefixed "Fitz-" (*e.g.*, Fitzgerald), which goes back to Norman French *fis* "son." In contradistinction to English, the Scottish patronymics are formed by a prefixed "Mac" or "Mc" (McGregor), the Irish with "O" (O'Brien) or "Mc" or "Mac," and the Welsh with "P-" (Powell "son of Howel"). In Modern Greek, patronymics are formed by suffixation; *e.g.*, Dimitriopoulos "son of Dimitrios."

French family names

The development of family names is similar in all of Europe. For example, French names such as Jaquet, Jacquot, Jacotot, Jacotin, Cottet, Cottin, Cotin, Jacquin, Jacquinet, Jacquinot, Jacquart, Jacquier all derive from Jacques; Davignon, Decaen, Derennes, and Beauvais developed from the place-names Avignon, Caen, and so forth; Breton, Lebreton, Lenormand come from the names of districts; Clerk, Leclerc, Duclerc, Auclerk, Clergue (*cf.* English Clark), Boucher, Boulanger, Masson designate professions ("butcher," "baker," "mason"); and Roux, Leroux, Roussel, Rousseau, Lerouge, Roujon are all variants of "red" (*i.e.*, red hair). Roughly the same scheme exists everywhere in Europe. Some family names can be traced to nicknames that must have their origin in incidents and attitudes that cannot be understood now; *e.g.*, Czech family names such as Nejezchleba "Don't eat bread!" and Skočdopole "Jump into the field!"

The development is slighty different among Jews. While living in ghettos, they used only given names. After the end of the 18th and the beginning of the 19th century, they chose or were given family names. Many of these names (which vary in the individual languages) are derived from religious vocations: Cantor, Canterini, Kantorowicz (lower priest); Kohn, Cohen, Cahen, Kaan, Kahane (priest); Levi, Halévy, Löwy (name of the tribe of priests). Many are derived from place-names—Morpurgo (Marburg)—or from nicknames—Hirsch ("deer" in German). Frequently, particularly in Austria, the Jews were given derisory family names, such as Eberstark "strong as boar," Rosenduft "fragrance of roses," and Hitzig "hot," from Itzick, a mocking form of Isaac.

The only outstanding exception to this European pattern of naming occurs with the names of kings, who use one of their given names. Some royal families have what could be called family names; *e.g.*, the Hohenzollerns (more correctly, Hohenzollers) of Prussia. The British royal family accepted the name Windsor only in 1917 (this was changed to Mountbatten-Windsor for the future members of the family who will not enjoy princely status). The pope of the Roman Catholic Church abandons his personal name after his election and chooses a single name, sometimes associated with his intentions; *e.g.*, Pope Paul VI chose the name Paulus because of St. Paul's missionary activities and travels.

Other patterns of naming. Names and naming practices in other cultural areas show a strong similarity in the basic trends. Among the ancient Assyrian and Babylonian names are theophoric designations such as Ashurbanipal, meaning "Ashur [a god] created son," and Nabukudurriusur (Nebuchadrezzar of the Bible), translated as "Nabu [a god] protected the estate." The Phoenician (Carthaginian) name Hannibal means "grace of Baal" (a god). The Hebrew Yehonatan, Yonatan (*i.e.*, Jonathan) means "God gave," Rafa'el (Rafael) is translated as "God cured." There are also nontheophoric names such as Laban (from Hebrew *lavan* "white"). The Aramaic surname of the fisherman Simon, Kepha, meaning "stone," became famous in the New Testament as Petros (Peter), the Greek translation of the name (*petra* "rock, stone").

Theophoric names

The more complicated structure of Arabic society brought an independent development similar to the European one. Given names such as Muḥammad, Ibrāhīm (=Abraham), Manṣur "victor, ʿAlī "exalted," ʿAbd Allāh "slave of Allāh" are differentiated by surnames such as Ibn ʿAbbās "son of ʿAbbās," al-Baghdādī "from Baghdad," al-Ghazālī "the spinner." The Caucasian (*e.g.*, Ossetic) personal name consists of a given name preceded by the name of the tribe (gens) in the genitive plural; the name of the father may be inserted, thus giving Gaglojty Soslany fyrt Nafi "Nafi, son of Soslan, of [the gens of] Gaglo." Chinese society has had the institution of hereditary family names since the 4th century BC, but the number of these names has been reduced to some 200. Examples include Chan, Mao, and Lu. The choice of the given name was much freer, but recent legislation seems to have restricted it. In a similar way, there are not more than 300 Korean family names, but only three of them—Kim, Pak, and Yi—belong to the great majority of families in Korea. The given name is chosen, but its choice is limited by the practice that one of the two syllables of the name should be identical within a family for a generation; the whole given name should have an auspicious meaning.

By the 20th century, the originally European pattern of given name + family name had been introduced practically everywhere. Black Africa (*e.g.*, among the Yoruba) now has the "normal" pattern of personal names, but both the given and the family names are of vernacular stock. There are given names such as Olúṣolá "god [non-Christian] made greatness," Òṣunbúnmi "Osun [a river] gave me," and Adeyẹmí "crown befits me," and family names like Ajólore "who [is] a kind doer." Among the American Indians there are, surprisingly, practically no theophoric names. Instead, the Indians used names related to the totem, to animals indicated by omens or dreams, and to successful incidents in life. Those North American

Indians who did not accept English names now use the English translation of their names as last names (which sometimes are not hereditary); *e.g.*, John Sleeping Owl, Mary Little Bear.

Place-names. *Descriptive and commemorative place-names.* If the "meanings" of place-names and the motives for their choice are examined, several broad types may be discerned. Descriptive names indicate a characteristic feature of the entity; *e.g.*, Rocky Mountains, North Sea, Newcastle. The chosen feature is sometimes only illusory or brought by chance, as in the case of the Pacific Ocean (only a small part of it was calm, or pacific, when first seen and named). Honorific and commemorative names are another broad category. Examples include Constantinople (formerly called Byzantium, renamed Konstantinoupolis "city of Constantine" because that emperor made it the capital of the Roman Empire); Aphrodisias "[the town of] Aphrodite" (in Asia Minor), changed into Stauropolis "city of the cross" with the advent of Christianity; Cartagena, transferred to Colombia (South America) in commemoration of Cartagena in Spain (Cartagena in Spain was in turn developed from Latin Carthago Nova, a translation of the name given to the town by the Phoenician settlers in commemoration of Carthage, the Phoenician rival of Rome); Nieuw Amsterdam, commemorative of the Dutch capital, changed to New York (honorific for the Duke of York). Among the numerous benedictory, wishful names are the Russian Vladivostok "govern the East!" (founded and named by Russians as their main base on the Pacific coast), Cape of Good Hope (a renaming of a more descriptive Cape of Tempests), and Greek Pontus Euxinus (now the Black Sea) "hospitable sea" (a renaming of Pontus Axeinos "inhospitable sea"). In most cases, however, place-names do not have a "meaning" at all, particularly not for the general user.

Types of place-names

Place-names reflecting historical influences. Place-names are frequently accepted into the language of a new population. The toponymy of the United States illustrates this well. Spanish names are numerous in the South and Southwest; *e.g.*, Florida, San Antonio, Santa Fe, Las Cruces, and San Diego, all of which are Spanish names of Roman Catholic saints or holidays. French names occur frequently in the Southeast and the central U.S. (*e.g.*, La Nouvelle Orléans, changed into New Orleans; Baton Rouge; St. Louis; Louisiana); Dutch names are found in the East (Haarlem changed into Harlem); Indian names are interspersed everywhere; and, finally, English names are superimposed over all the rest. An examination of all these names, which are now used by a mostly English-speaking population, could not fail to yield some information about the colonization of the United States, even if the history were not known.

In the same way, the place-names of Britain reflect its history. Above all, there are Celtic names; *e.g.*, Eboracum (named for a tree), partly translated as Eoforwic, which changed into York. Roman names are also numerous; *e.g.*, Castra "military camp," changed into Chester, and Lindum Colonia "Colony Lindum" (which itself is Celtic, *linne* + *dunom* "town at the lake") is now Lincoln. There are, in addition, Anglo-Saxon names (*e.g.*, Whittingham "habitation of Hwita's people"), a few Scandinavian names (*e.g.*, Badby, in which *-by* is the Scandinavian element instead of English *-bury* "castle"), and Norman names (*e.g.*, Richmond, from a personal name consisting of *rīki* "rich, powerful" + *mond* "world").

Changes in place-names

Any country's toponymy consists of various layers. In France there are Celtic names such as Lucodunos "shining town" that became Latinized into Lugdunum and changed into the modern form Lyon; Greek names such as Agathe (Tyche) "good (luck)," which has become Agde; Roman names such as Forum Julii "marketplace of Julius," modern Fréjus; and old Germanic names such as Clarbec "clear brook" (*klar* + *Bach* in German). Most important is the fact that place-names can be used as a source of information themselves, even if other forms of evidence are lacking; for instance, because Moskva (=Moscow) is a Finnish name (Finnish *kva* "water"), and other Finnish toponyms are present in the central Soviet Union, the prehistoric presence of Finnish tribes in that location can be presumed. Names of rivers are particularly important for this purpose, because they are very conservative and usually are taken over by a new population. A considerable number of river names in western and central Europe show remarkable similarity (*e.g.*, Esera in Spain, Isère in France, Ysar in Belgium, Isar in Bavaria, Jizera in Czechoslovakia) and are the only evidence of a pre-Celtic Indo-European population of those regions.

In a similar way, names are also indicative of cultural and political trends. Singapore (Sanskrit Siṃhapura "castle of lions"), for example, testifies to the cultural influence of India in the area. Particularly significant in this respect are deliberate changes of names caused by changes in political power, ideology, and so forth. Changes such as Nieuw Amsterdam becoming New York, or Léopoldville (named after the king of Belgium who acquired the Congo) becoming Kinshassa, are very common. Some cases of renaming do not lack humour: La Roche-sur-Yon in France was rebuilt by Napoleon and renamed Napoléon-Vendée, which was changed to Bourbon-Vendée under the Restoration; during the 100 days of Napoleon's return from Elba and in the subsequent second Restoration, the cycle was repeated once more, and only the Third Republic restored the old name. Renaming also shows examples of cynicism: Lyon, which had rebelled against the revolutionary covenant, was punished by systematic demolition and a massacre of its inhabitants; the ruins of the city were renamed Commune-Affranchie "Liberated Commune." Changes of place-names are sometimes made systematically; when the territory called Alto Adige in Italian and Südtirol (South Tirol) in German became part of Italy after World War I, for instance, a systematic effort was made to give or to return to these territories the Italian character in both place-names and personal names. The Russian Revolution brought a change of names that were reminiscent of the old regime and ideology; *e.g.*, St. Petersburg (changed into a more Slavic Petrograd by the tsar during the war) became Leningrad, Caricyn (from *car* "emperor") became Stalingrad (and then Volgograd in the late 1950s), Yekaterinodar (from Yekaterina [=Catherine], an empress) became Krasnodar (from *krasny* "red"), but Arkhangelsk "town of archangels" somehow escaped notice.

BIBLIOGRAPHY

Books: C.W. BARDSLEY, *Dictionary of English and Welsh Surnames* (1901, reprinted 1967), good for quick reference concerning individual names; C.H. L'ESTRANGE EWEN, *A History of Surnames of the British Isles* (1931, reprinted 1968), contains rich information and a bibliography; E.G. WITHYCOMBE (comp.), *Oxford Dictionary of English Christian Names*, 2nd ed. (1950), indicates the provenience of names listed; L.G. PINE, *The Story of Surnames* (1965), readable and well informed; H.L. MENCKEN, *American Language*, suppl. 2 (1948), a survey of American naming; E. EKWALL, *Concise Oxford Dictionary of English Place-Names*, 4th ed. (1960), a handy reference that gives the provenience of place-names; J.B. JOHNSTON, *Place-Names of Scotland*, 3rd ed. (1934); P.W. JOYCE, *Origin and History of Irish Names of Places*, 7th ed., 3 vol. (1920), partly antiquated, but still of great value; G.R. STEWART, *Names on the Land*, 3rd ed. (1967), gives the origin of American place-names, readable and very instructive because it not only provides the etymology of names, but indicates why they were chosen; and E. PULGRAM, *Theory of Names* (1954; reprinted by the American Name Society from *Beiträge zur Namenforschung*, 1954), treats theoretical problems, such as the distinction between names and common nouns.

Journal: Names, quarterly journal of the American Name Society (1953–).

The study of names is plagued with imaginative and fantastic, but unfounded and unscholarly, publications; therefore, caution should be exercised in the choice of sources.

(L.Zg.)

Namib Desert

The Namib is a cool coastal desert extending for 1,200 miles (1,900 kilometres) along the Atlantic coast of southwestern Africa, from Moçâmedes in Angola southward across South West Africa to the Olifants River in the Cape

Province of South Africa. It reaches inland 80 to 100 miles to the foot of the Great Escarpment of southern Africa. Its southern portion merges with the Kalahari Desert on the plateau atop the escarpment. Its name is derived from the Nama language, implying "an area where there is nothing."

The Namib is arid and is almost totally uninhabited, except for several towns. It is important because of the trade routes which cross it, its mineral deposits, the fisheries of the bordering sea, and its increasing utilization for recreation.

Landscape. The desert basically consists of a relatively smooth platform of truncated bedrock of various types and ages. Mica-schist and other metamorphics and granite and similar intrusives predominate. The platform rises gradually from the coast to about 3,000 feet at the foot of the escarpment. Scattered isolated mountains rise steeply and abruptly above the platform, and, in the northern half, streams have carved deep, steep-walled gorges into it.

In much of the southern half of the desert the platform is surmounted by a vast expanse of sand—yellow-gray near the coast and brick-red inland—which is derived from the Orange River, as well as from other rivers that flow westward from the escarpment but never reach the sea. The dunes are arrayed in lines that run from north-northwest to south-southeast, with individual dunes having lengths of ten to 20 miles and reaching heights of 200 to 800 feet. The troughs between are interrupted by smaller transverse dunes. The extreme southern coastal area consists of wind-scoured bedrock and a few rapidly moving crescent-shaped barchan (*i.e.*, convex to the wind) dunes. The northern third—the Kaokoveld region—consists of gravel plains and rock platforms occurring between scattered, rugged mountains, interspersed with a few large dune fields.

Drainage and soils. Being an almost rainless area, the Namib has a poorly developed and fragmentary drainage pattern. Water from the interior plateau flows through or into the desert. In the northern half the larger streams reach the sea, but between the Kuiseb and the Orange rivers, every stream terminates in a vlei (salt pan or mud flat) against or among the dunes.

A portion of the water of major streams seeps through the sands of the stream beds. The underflow of the Kuiseb River has been tapped 25 miles inland to provide water supplies for Walvis Bay and Swakopmund. A pipeline 80 miles long supplies Lüderitz with water from the seepage of the Koichab, a stream that terminates in the dunes. Only the Cunene and the Orange rivers flow permanently on the surface. Other streams have surface flow only after heavy rainfall in the interior plateaus—they normally flow for no more than a few days in several years.

Large areas of the Namib are completely soilless, with bedrock at the surface. Other areas are covered with shifting sand. Soils that do occur are often highly saline, impregnated with gypsum, or cemented firmly by calcium carbonate, forming a calcrete layer just below the surface. Arable soils in the Namib are limited to floodplains and terraces of major rivers and are subject to occasional inundation.

Climate. The coastal area is almost totally rainless, yet its air is almost always at or near the saturation point. The cold Benguela Current flows northward along the coast, chilling the air above it, and thus producing fog. This cool air moves inland as a southwest sea breeze, creating a temperature inversion about 1,000 feet thick, with fog below and hot, dry air above.

Temperature range

At the coast, there is little difference in temperature between day and night or between winter and summer. Temperatures are usually in the 50s or 60s (Fahrenheit). Along the inland margins, summer temperatures normally reach the high 80s; only in areas sheltered from the cooling sea breeze (lee sides of mountains and bottoms of canyons) do temperatures frequently approach those expected in low-latitude deserts—*i.e.*, in excess of 100° F. Freezing temperatures occur occasionally along the inner edge of the desert.

On a few days each year, usually in fall or spring, "Berg," or mountain, winds, blowing from the east, bring high temperatures (over 100° F), together with dry air and clouds of dust, across the desert to the coast itself. The rare rains occur in spring and fall, usually as short-lived torrential thunderstorms.

Average annual precipitation at the coast is approximately one-half inch, increasing inland, until it reaches two inches at the foot of the escarpment. Some years, however, have no rain at all. Dew, on the other hand, is heavy, and for some types of vegetation is more important than the rainfall. In the extreme south, some winter precipitation occurs from frontal storms passing farther south over the Cape region; on rare occasions, snow may fall on the higher southern mountains.

Vegetation and animal life. Six vegetational regions may be identified in the Namib: (1) the coastal region, with highly succulent vegetation, which uses moisture derived from the fog; (2) the almost completely barren Outer Namib; (3) the steppes of the Inner Namib, which in many years are barren, but which in wet years are covered with short grasses, both annual and perennial; (4) the dunes of the Inner Namib, which produce a surprisingly rich flora of bushes and tall grasses; (5) the larger river channels, along which large trees, particularly acacias, grow; and (6) the southern winter rainfall area, where a succulent bush growth occurs. A curious Namib plant is *Welwitschia mirabilis*, whose two gigantic leaves sprawl over the surface of the ground from the crest of its huge root crown.

The plains and dunes of the Inner Namib support large numbers of antelope (especially gemsbok and springbok) and ostriches and some zebras. The dunes of the Outer Namib have various types of insects and reptiles, especially beetles, geckos, and snakes, but virtually no mammals. The shore area is densely populated by marine birds (notably flamingos, pelicans, and, in the southern part, penguins), as well as a few jackals, some rodents, and two colonies of seals. Large quantities of guano are scraped annually from the rocks of several offshore islands, and from one artificial platform.

The inhabitants. A few Bushmen roamed the Namib until early in the present century, gathering whatever was edible along the shore, hunting in the Inner Namib, and often depending for water upon the bitter juices of the tsama (a species of watermelon). Today a few hundred Ovahimba and Ovatjimba Herero tribesmen herd cattle and goats from waterhole to waterhole in the desert part of the Kaokoveld, living in their traditional manner. A few Topnaar Namas ("Hottentots") graze their sheep and goats on the riverine vegetation along the Kuiseb River.

Towns

Walvis Bay is a modern port city with a 1970 population of 23,500, composed of 8,320 whites (predominately Afrikaners, but including a considerable number of English and Germans); 3,210 Coloureds (mixed white, Malay, and Nama), chiefly from the Cape Town area and associated with the fishing industry; and 11,930 Africans, consisting partly of local Nama families, but largely of male Ovambos (Ambos) from Angola and the northern part of South West Africa, working at the port under short-term labour contracts. The port serves as a base for pilchard and whitefish fleets, which supply both shore canneries and cannery ships on the open sea. It is also the major transhipment port for South West Africa, exporting ore, hides, wool, tinned fish, fish oil and meal, and frozen lobster, and importing petroleum, foodstuffs, and manufactured goods from Europe, the United States, and Japan.

Lüderitz, a small town on a shallow, rock-strewn harbour, has a small trade with Walvis Bay and Cape Town and is the base of a crayfish (lobster) industry. Oranjemund, a company town of the Consolidated Diamond Mines, is a base for large-scale diamond mining in the alluvial gravels along the southern coast.

Swakopmund, the summer capital for South West Africa and a popular coastal resort town, still retains much of its former German atmosphere. The coastal area for nearly 100 miles north of Swakopmund is becoming increasingly

popular as a recreational area in general and for fishing in particular.

The innermost region of the desert in its southern two-thirds is divided into "farms" (*i.e.*, ranches), devoted chiefly to the raising of karakul sheep. Sheep-raising, however, suffers severely from recurrent droughts.

Transport. The desert is crossed by rail lines that link the ports of Lüderitz and Walvis Bay with South Africa. A tarred road from Walvis Bay also runs into the interior. Graded roads connect Oranjemund with Lüderitz, and thence with the interior, and a graded road runs from Swakopmund northward along the coast for about 150 miles. Two areas are completely roadless—the Kaokoveld in the north and the area between Walvis Bay and Lüderitz. The entire area between the Kuiseb and Orange rivers is a diamond area that remains closed completely to public entry except for the road and rail access to Lüderitz. There are airports with regularly scheduled passenger service located near Walvis Bay; at Lüderitz; and at Alexander Bay, to the south of the Orange River.

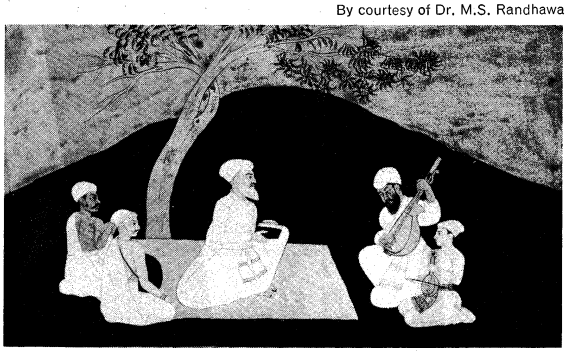
BIBLIOGRAPHY. E. KAISER, *Die Diamantenwüste Südwest-Afrikas*, 2 vol. (1926), a masterly compendium of geological and geomorphic information on the southernmost region; C. KOCH, "The Fauna of the Namib Desert," *Bull. Trans. Mus.*, vol. 2 (1955), a survey of animal life, with emphasis on lower forms; R.F. LOGAN, *The Central Namib Desert, South West Africa* (1960), a comprehensive study of the geography (physical, cultural, and economic) of the area from Walvis Bay and Swakopmund inland, and "Land Utilization in the Arid Regions of Southern Africa," pt. 2, "South West Africa," in L.D. STAMP (ed.), *History of Land Use in Arid Regions*, pp. 331–338 (1961), the historical development and present pattern of land use in South West Africa; H. WALTER, "Die Ökologischen Verhältnisse in der Namib-Nebelwüste," *Jb. Wiss. Bot.*, 84:58–222 (1936), a detailed study of plant ecology, with special reference to the utilization of scanty water supplies in the foggy coastal area; J.H. WELLINGTON, *Southern Africa: A Geographical Study*, 2 vol. (1955), a textbook of the geography of southern Africa, with a description of the Namib emphasizing physical aspects.

(R.F.Lo.)

Nānak

Nānak, a religious teacher of the Punjab (Pañjāb), commands a particular importance in the history of Indian devotional religion. Although born a Hindu, he marks the beginning of the Sikh community, and among modern Sikhs he enjoys a particular affection as the first of their Gurūs and as the supreme master of Punjabi devotional hymnody.

By courtesy of Dr. M.S. Randhawa



Nānak (centre), painting in the Mughal style, *c.* 1689. In the collection of Mahant Indresh Charan Dass, Dehra Dūn, Uttar Pradesh, India.

What little information there is about Nānak's life has been handed down mainly through legend and tradition. There is no doubt that he was born in 1469, probably on April 15th, in the village of Rāi Bhoi dī Talvaṇḍī, 40 miles southwest of Lahore. His father was a member of a subcaste of the mercantile Khatrī caste. The relatively high social rank of the Kahtrīs distinguishes Nānak from other Indian religious reformers of the period and may have helped promote the initial growth of his following. He married the daughter of a Khatrī, who bore him two sons.

For several years Nānak worked in a granary until his religious vocation drew him away from both family and employment, and in the tradition of Indian religious mendicants, he embarked on a lengthy journey, probably travelling to the Muslim and Hindu religious centres of India, perhaps even to places beyond its borders. Neither the actual route nor the places he visited can be positively identified.

References found in four of his hymns suggest that Nānak was present at attacks Bābur (an invading Mughal ruler) launched on Saidpur and Lahore, so it seems safe to conclude that by 1520 he had returned from his travels and was living in the Punjab.

Founder of Sikh community

The remaining years of his life were spent in Kartārpur, another village of central Punjab. Tradition holds that the village was actually built by a wealthy admirer to honour Nānak. It was presumably during this final period that the foundations of the new Sikh community were laid. By this time it must be assumed that Nānak was recognized as a Gurū, an inspired teacher of religious truth, and that, in accordance with the custom of India, disciples who accepted him as their Gurū gathered around him in Kartārpur. Some probably remained as permanent residents of the village; many more made periodic visits to obtain his blessing. All of them listened to the teachings expressed there in numerous devotional hymns intended for communal singing, many of which survive to this day.

The actual year of Nānak's death is disputed, tradition being divided between 1538 and 1539. Of these two possibilities, the latter appears to be the more likely. One of his disciples, Aṅgad, was chosen by Nānak as his spiritual successor, and following Nānak's death he assumed the leadership of the young Sikh community as Guru Aṅgad.

In view of the size of the following that Nānak attracted, numerous anecdotes concerning the deeds of the Gurū began to circulate within the community soon after his death. Many of these were borrowed from the current Hindu and Muslim traditions, and others were suggested by Nānak's own works. These anecdotes were called *sākhīs*, or "testimonies," and the anthologies into which they were gathered in rough chronological order are known as *Janam-sākhīs*. The interest of the narrators and compilers of the *Janam-sākhīs* has largely concentrated on the childhood of Nānak and above all on his travels. Amongst the earlier traditions are tales of visits he is supposed to have made to Baghdad and Mecca. Ceylon is a later addition, and later still the Gurū is said to have travelled as far east as China and as far west as Rome. Today the *Janam-sākhīs* offer a substantial corpus of hagiographical material, and the more important of these collections continue to be the basis of "biographies" of Nānak.

Doctrine

Nānak's message can be briefly summarized as a doctrine of salvation through disciplined meditation on the divine name. Salvation is understood in terms of escape from the transmigratory round of death and rebirth to a mystical union with God. The divine name signifies the total manifestation of God, a single Being, immanent both in the created world and within the human spirit. Meditation must be strictly inward, and all external aids such as idols, temples, mosques, scriptures, and set prayers are explicitly rejected. The Muslim influence is relatively slight; the influence of Hindu mystical and devotional beliefs is much more apparent. Always, though, the coherence and beauty of Nānak's own expression dominates early Sikh theology.

BIBLIOGRAPHY. W.H. MCLEOD, *Gurū Nānak and the Sikh Religion* (1968), an attempt to apply strict historical research procedures to the traditions concerning Nānak, his life and teachings; HARBANS SINGH, *Guru Nanak and the Origins of the Sikh Faith* (1969), more generous in its treatment of the traditional narratives, but not uncritical; J.S. GREWAL, *Guru Nanak in History* (1969), an analysis of the impact of Nānak upon the subsequent history of the Sikh community; GURBACHAN SINGH TALIB, *Guru Nanak: His Personality and Vision* (1969), a study of the teachings of Nānak; GANDA SINGH (ed.), *Sources of the Life and Teachings of Guru Nanak* (1969), a useful anthology of works dealing with Nānak from the earliest

traditions to the present day, including an English translation of a *janam-sākhī* (the *Mahima Prakas*) and a comprehensive bibliography; KHUSHWANT SINGH, *Hymns of Guru Nanak* (1969), a selection of Nānak's hymns in English translation.

(W.H.McL.)

Nanking

Nanking (Nan-jing in Pin-yin romanization), capital of Kiangsu Province in central eastern China, is a port on the Yangtze River (*q.v.*) and a major industrial and communications centre. Rich in history, it served seven times as the capital of regional empires, twice as the seat of revolutionary governments, once (during World War II) as the site of a puppet regime, and twice as the capital of a united China (the second time from 1928 to 1937). In the early 1970s its population was estimated at about 2,300,000. The name Nanking (Southern Capital) was introduced in 1421 during the Ming dynasty.

History. Nanking's recorded history dates back to the Eastern Chou dynasty (8th to 3rd centuries BC), when it was held successively by the states of Wu, Yüeh, and Ch'u. In 221 BC, when the Ch'in annexed it, it was called Chin-ling. Under the Ch'in (221–206 BC) and Han dynasties (206 BC to AD 220), Nanking—named, respectively, Mo-ling and Tan-yang—remained a prefecture within the unified empire.

Early history

Nanking—under the name of Chien-yeh—emerged as the political and cultural centre of Southeast China during the period of the Three Kingdoms, when Sun Chien and his son Sun Ch'üan made it the capital of the kingdom of Wu from 229 to 280. In 317 the Eastern Chin dynasty (317 to 419), fleeing foreign invaders in North China, again chose the city as a capital. Renamed Chien-k'ang, Nanking became a haven for northern families in exile. After the fall of the Eastern Chin, Nanking, under four successive dynasties—Liu-Sung (from 420 to 479); Southern Ch'i (479 to 502); Liang (502 to 557); and Ch'en (557 to 589)—was the seat of government of the regional empires south of the Yangtze.

These regimes were dominated by military men whose rivalries weakened the government. But in Nanking progress was made in areas other than politics. Bountiful harvests, coupled with tea and silk industries, supported a booming economy. Culturally, the Six Dynasties—as the dynasties that ruled from 220 to 589 are idiomatically known—produced a galaxy of scholars, poets, artists, and philosophers. The works of Wang Hsi-chih and Ku K'ai-chih set the canons of calligraphy and painting, respectively. The publication of *Wen hsüan* ("Literary Selections") and of *Wen-hsin tiao-lung* (a classic in literary criticism), the evolution of what has come to be known as the Six Dynasties essay style, a blending of poetry and prose, and the invention by Shen-yüeh (a 6th-century courtier) of the system of determining the four tones of the Chinese language (which permits one syllable to be spoken in four different ways, thus serving four different characters with four different meanings) were achievements of this period. In philosophy, the so-called *ch'ing-t'an* ("pure discourse") movement, spiritually akin to a form of Taoism, found many adherents who held themselves aloof from politics. Hundreds of Buddhist temples were built; voluminous Buddhist scriptures were edited and transcribed; and thousands, including Emperor Wu Ti of the Liang dynasty, who ruled in the 6th century AD, took monastic vows.

From 581 to 1368, under the successive unified empires of the Sui, T'ang, Sung, and Yüan dynasties, Nanking reverted to the status of a prefectural city. Various names were given the city: Chiang-chou and Tan-yang under the Sui; Chiang-chou, Chin-ling, and Pai-hsia in early T'ang; Sheng-chou in late T'ang; and Chin-ling again under the Five dynasties in the 10th century; Chien-k'ang under the Sung; and Chi-ch'ing under the Yüan. When the Southern T'ang briefly maintained a regional regime in the city from 937 to 975, Nanking enjoyed much intellectual creativity (the ruler Hou-chu himself being a poet of consummate skill) and was the scene of new construction, notably, the octagonal stone pagoda of the Ch'i-hsia-shih Temple and the crosstown channel of the Ch'in-huai Ho. Another period of prominence occurred during the Southern Sung dynasty (1126 to 1279 AD), when Yo Fei used the city as his base for resistance against the Tatars in North China.

In 1368 Emperor Hung-wu, founder of the Ming dynasty, made Nanking the capital of a united China. Naming the city Ying-t'ien-fu (Responding to Heaven), he built a grand Imperial palace and a gigantic city wall, made of brick, 60 feet high and 32 miles long, with 13 gates. In addition, earth ramparts were prepared to form the basis for a still larger outer wall, 60 miles in length, with 18 gates. In 1421, however, Hung-wu's son, the Yung-lo emperor, moved the capital to Peking. The city became a subsidiary capital and was renamed Nanking.

The growth of trade and industry, however, brought new wealth to Nanking, especially to Hsia-kuan, the city's foremost port and suburb. Weaving, pottery, printing, and brocade making were the leading industries. Oceangoing vessels used by Cheng Ho in his famous 15th-century expeditions to the South Seas were built in the shipyards to the northwest of the city. An Imperial college—the Kuo-tzu chien—attracted students from all over the empire, as well as from Japan, Korea, Okinawa, and Thailand. The scholars of this college helped compile the *Yung-lo ta-tien* ("Yung-lo Emperor's Encyclopaedia"); its printing plant issued fine editions of many classics, as well as such important works as *Pen-ts'ao kang-mu* ("Great Pharmacopoeia").

In the Ch'ing dynasty (1644 to 1911), Nanking, renamed Chiang-ning, became the government seat of the viceroy of Kiang-nan (who governed the provinces of Kiangsu, Kiangsi, and Anhwei). In 1842 the treaty ending the Opium War was signed there. A decade later, in 1853, the city was taken by the Taiping revolutionary forces under the leadership of Hung Hsiu-ch'üan. As the capital of *T'ai-p'ing T'ien-kuo* (Heavenly Kingdom of Peace), Nanking became a commune practicing universal brotherhood, equality of the sexes, and communal ownership of property. Despite the construction of palaces for Hung and his lieutenants, there was widespread destruction of public buildings, of temples, and of the city wall. When the Taipings were overthrown in 1864, the city was nearly prostrate.

The Taiping capital

Recovery took decades. Although it was sanctioned by the treaties of Tientsin concluded with France in 1858, foreign trade did not begin until 1899. By that time, modern industry and communication had reached the city. In 1908 the Shanghai–Nanking railroad was opened, followed by a railroad from the port city of Tientsin in Hopeh Province to P'u-k'ou, on the northern bank of the Yangtze opposite Nanking, which opened four years later. Such economic growth, however, was overshadowed by the revolution of 1911. After the uprising had begun upstream at Wu-ch'ang in Hupeh Province, the revolutionary leaders proclaimed Nanking the seat of the provisional government of the Republic of China, and the democratic constitution of 1912 was adopted there before the first president, Yüan Shih-k'ai, moved the capital to Peking. Thus, within 60 years, Nanking was twice at the centre of the revolution to overthrow the Manchu dynasty.

Under the infant Republic of China, Nanking was governed by warlords for over a decade. Sun Yat-sen, leader of the Kuomintang (Nationalist Party), embittered by politicians' intrigues centred in Peking, vowed to make Nanking the Nationalist capital. Accordingly, when his follower Chiang Kai-shek achieved national unification in 1928, the Nationalist government made Nanking once more the capital of a united China. Remarkable progress was made in developing communications, industries, and natural resources. Physically, too, the city acquired a new look; modern boulevards and government buildings were constructed; new railroad stations and airfields were built; and the Sun Yat-sen Mausoleum was erected.

Such achievements were, however, cut short by the war against Japan. Nanking fell in 1937. In the sack of the city that followed, more than 40,000 civilians were slaughtered. The city was then ruled by puppet governments until Japan's defeat in 1945. From 1946 to 1949,

Nanking resumed its status as the capital of Chiang Kai-shek's Nationalist government. It was then threatened by the rising power of the Chinese Communists, whose forces took the city in 1949. When the People's Republic of China was proclaimed on October 1, 1949, Nanking was once again abandoned in favour of Peking as the national capital. In the administrative reorganization of 1952, it was made the provincial capital of Kiangsu, replacing the nearby city of Chen-chiang.

The contemporary city. *Location, extent, and climate.* Nanking is situated on the southeast bank of the Yangtze, some 160 miles from Shanghai (215 miles by water, 193 miles by rail). In addition to the city proper, which occupies the area enclosed by the walls, the municipality of Nanking includes territory extending to the town of Lung-tan in the east and to the borders of Chiang-ning County (*hsien*) in the south, Pa-kua-chou Island in the Yangtze, and an area north of the Yangtze from P'u-k'ou to the Anhwei border between the county lines of Liu-ho and Chiang-p'u. Included in the municipality are both urban and rural districts.

The four seasons are clearly distinguishable. The hot summer months are from July to September, while winter lasts from December until March. Spring and autumn are both mild and pleasant. January and July mean temperatures are about 37° F (3° C) and 81° F (27° C), respectively. The average annual rainfall is 36 inches, the bulk of it falling between June and August.

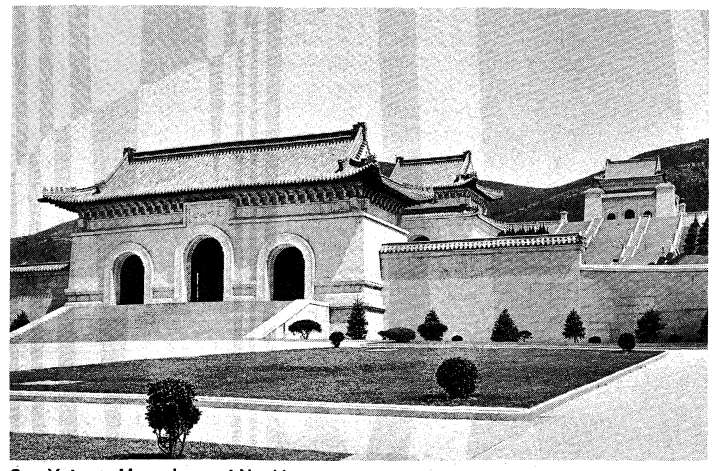
City layout and landmarks. The city of Nanking resembles a gourd, with its tip pointing northwest, toward Hsia-kuan, its base resting against the Ch'in-huai Ho (Ch'in-huai River) on the west and south and the foothills of the Tzu-chin Shan (Purple-Gold Mountain) on the east, and with its middle dented inward by Hsüan-wu Hu (Mystic Martial Lake) on the northeast and Moch'ou Hu (Lake of No Sorrow) on the southwest. The skyline suggests spaciousness and grandeur. Blue-glazed tiles adorning the gates, parklike scenery along the streets, lotus blossoms and tea pavilions on the lakes, and temples half hidden in the hills—all these are characteristic sights in Nanking.

The city's four districts

The city is comprised of four districts. The north district, from I-chiang Men (a city gate) to the Ku Lou (Drum Tower), traversed by Chung-shan Avenue as its axis, contains such landmarks as Pei Chi Ko (North Pole Pavilion) and Chi-ming Shih (Cockcrow Temple); the central district, surrounding Hsin Chieh K'ou (New Crossroads), is the business centre of the city; to the south is situated the populous old city—traditionally famous for entertainment; the east district, containing the former Ming palace and the municipal airfield, lies east of the central district.

Forming an integral part of the life of the city are its immediate outskirts. To the south is the Yü Hua T'ai (Terrace of the Rain of Flowers), noted for its five-colour pebbles and a Communist martyrs' memorial. In the eastern outskirts, on the magnificent Tzu-chin Shan range of hills, are the emperor Hung-wu's mausoleum, the mausoleum of Sun Yat-sen, and the Ling Ko Shih (Temple of the Valley of Spirits). Other scenic spots include Yen-tzu Chi (Swallow's Bluff) in the north and Ch'i Hsia Shan (Abode of Clouds Mountain) farther to the northeast.

Transportation. Nanking's major avenue of commerce is the Yangtze River, which is from 2,700 to 3,900 feet wide in the vicinity of Hsia-kuan, the hub of the city's river traffic. Hsia-kuan is also the terminus of the Shanghai–Nanking railroad and the Tientsin–P'u-k'ou railroad. In 1968 the ferry between P'u-k'ou and Hsia-kuan was replaced by the Yangtze bridge, 20,100 feet in length, with a double-track railroad on its lower deck and a four-lane highway on the upper deck. A third railroad leading south, the Nanking–Wu-hu railroad, which extends to T'ung-ling in Anhwei Province, will eventually connect with the Kiangsi–Chekiang railroad; a loop line through the eastern suburbs links it with the Shanghai–Nanking line. Nanking is also served by a municipal railway that runs eight miles from the centre of the city to Hsia-kuan.



Sun Yat-sen Mausoleum at Nanking.
Ewing Galloway

Major highways fan out from the East Gate to Chenchiang via Chü-jung, to Shanghai via Wu-hsi, and to Hangchow via I-hsing and Ch'ang-hsing. Other highways lead from P'u-k'ou to Ch'u-hsien, Pang-pu, and Ho-fei in eastern Anhwei and from the South Gate to Ma-an-shan, Tang-t'u, and Wu-hu in southeastern Anhwei, where there are connections with the Chekiang road system. Nanking lies on the routes of all major air services: Shanghai to Wu-han and Chungking; Canton to Tsinan, Peking, and Mukden; and Shanghai to Cheng-chou, Sian, and Lan-chou.

Demography and economic life. In 1970 Nanking was estimated by the United Nations to have a population of 2,300,000. Despite the industrial growth of the past two decades, the city retains a traditional feature—the existence of a substantial rural population within the municipality. The people speak Mandarin with a marked local accent that is pedestrian in character.

For centuries, Nanking was noted for its satins, pongees (home-woven silks), velvets, brocades, porcelainware, paper, and ink sticks. Modern industry was actively promoted by the Nationalists. The city's major industrial expansion, however, has taken place since the 1950s. In the early 1970s Nanking was the largest industrial centre in Kiangsu (excluding Shanghai, which is a special municipality). Its chief industries are iron and steel; machine building; farm equipment; motor vehicles (trucks, jeeps, autos); cement; oil refining; chemicals, fertilizers and insecticides; thermoelectric plants; arsenals; and coal mining. No less important are textiles, silk and silk piece goods, synthetic fibres, food processing, optical instruments, radios, pharmaceuticals, clocks, and fountain pens.

The city's industries

More than 80,000 acres of land in the municipality's rural districts are under cultivation, the major agricultural products being rice, wheat, corn, beans, vegetables, and watermelon. *Pai-ho* (lily bulbs), cherries, apples, water chestnuts, lotus roots, pine nuts, and peanuts (groundnuts) are grown on the hills and on the shores of the lakes. Along the canals and creeks, farmers raise large flocks of ducks; the Nanking Duck, preserved in salt and pressed, is one of China's food delicacies.

Imports from abroad consist primarily of industrial equipment for new plant installations. In addition to the city's manufactured goods, exports include eggs, poultry, tea, rice, wheat, beans, cotton, and peanuts.

Political institutions. Government authority is vested in the Nanking Municipal People's Council, elected by the Nanking Municipal People's Congress and composed of the mayor, deputy mayors, and some 40 council members. Serving under the council are numerous bureaus and commissions in charge of civil affairs, public security, justice, planning, finance, food, taxation, industry, commerce, labour, culture, education, public health, physical culture and sports, projects of municipal construction, and public services.

Administratively, the city is divided into districts (*ch'ü*).

Below the district, police substations supervise the population, while street mayoralties handle civil affairs in the same sub-area. Neighbourhood associations help mediate disputes, carry out literacy campaigns, and promote sanitation, welfare, and the continuous "weeding out of bourgeois elements." Inclusion of rural areas in the municipality reflects the government's policy to reduce town-and-country differences by maintaining integrated urban–rural units surrounding the city proper.

Health. Nanking made great progress in public health and medicine in the Nationalist period. Under the present regime, the Chinese Communist Party launches numerous drives to teach the population hygiene, disease-prevention methods, and the value of exercise for physical fitness. In 1967 there were 45 hospitals in Nanking; clinics and health stations are maintained by all neighbourhood associations. A significant recent trend is the designation of individual hospitals for specialization in specific branches of medicine. Nanking is also a centre for training doctors in traditional Chinese medicine—the complementary use of modern and traditional medicine being considered of prime importance.

Education and cultural life. Nanking inherited an excellent school system from Nationalist days. In 1949 there were 300 primary and middle schools, while the National Central University and the U.S. missionary colleges (the University of Nanking and Ginling Girls College) were among the leading institutions of higher education in the country.

Since the 1950s, considerable change has taken place in the city's educational system. Instruction in the primary schools stresses Chinese language, arithmetic, foreign language, geography, and history; Maoist thought; military and physical training; revolutionary art and literature; and labour. In the middle schools, mathematics, physics, chemistry, biology, geography, and proletarian ideology are high priority subjects. Since the Cultural Revolution of 1966 to 1969, the period of school attendance has been reduced, and students are sent on graduation to spend one or two years at physical labour in factories or communes. There has also been a notable increase in teacher-training schools, vocational–technical schools, and part-time agricultural middle schools.

The universities

Paramount among the universities are the Nanking University and the Nanking Technical University. There are about a dozen technical institutes in agriculture, geology, pharmacy, hydraulic engineering, aeronautical engineering, mechanical engineering, civil and chemical engineering, and the food industry. Admission is based not only on examination but also on practical experience and political orientation.

Numerous research agencies and scientific societies are located in Nanking. These include a branch of the Chinese Academy of Sciences, the Tzu-chin Shan Astronomical Observatory, and the Sun Yat-sen Botanical Garden, as well as various research institutes. Prominent among the museums are the Nanking Museum, the Museum of the Heavenly Kingdom of Great Peace, and the Kiangsu Provincial Museum. An important cultural centre for the general public is the Workers' Cultural Palace. The Nanking Library, having absorbed the Sinological Library of Kiangsu Province, contains about 2,300,000 volumes.

Nanking's principal newspaper is *Hsin-hua jih-pao* ("New China Daily"); the city is within easy reach of the newspapers of Peking and Shanghai, such as the *Jen-min jih-pao*, the *Kuang-ming jih-pao*, the *Ta-kung pao*, and the *Wen-hui pao*. While television is not yet available to everyone, radio broadcasts bring the day's news to every farm, factory, school, and home. *Ta-tze-pao*, or large-character wall posters, enjoyed great vogue during the Cultural Revolution.

Basketball, tennis, soccer, and track and field are popular sports, and gymnasiums and sports grounds are found in all parts of the city and suburbs. The Wu-t'ai Shan Stadium in the centre of the city provides accommodation for major sports events. The public at large, however, finds its recreation in the many beautiful parks and resorts.

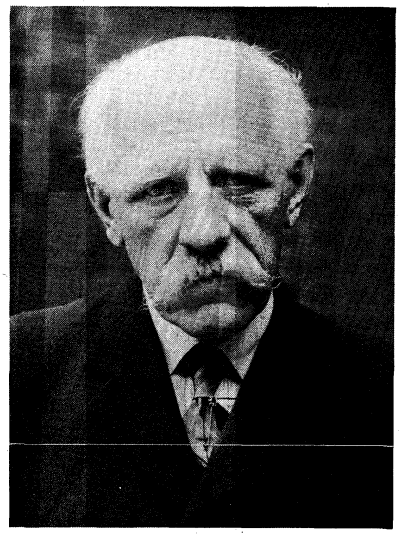
BIBLIOGRAPHY. *Nagel's Encyclopedia-Guide to China* (1968), though a tourist guide, is useful for general reference. For geography and economic development, see T.R. TREGEAR, *An Economic Geography of China* (1970). TSUI CHI, *A Short History of Chinese Civilization* (1942), and IMMANUAL C.Y. HSU, *The Rise of Modern China* (1970), contain general references to Nanking's history. For accounts of specific epochs, see A.F. LINDSAY, *Ti-ping Tien-kwoh: The History of the Ti-ping Revolution*, 2 vol. (1866); PAUL M.A. LINEBARGER, *The China of Chiang Kai-shek* (1941); and HAROLD J. TIMPERLEY (ed.), *What War Means: The Japanese Terror in China* (1938). Glimpses of contemporary Nanking may be found in HUGO PORTISCH, *So soh ich China* (1965; Eng. trans., *Red China Today*, 1965); and in REWI ALLEY'S article on Nanking in *Eastern Horizon*, vol. 10, no. 2 (1971). Helpful information may also be found in two current magazines—*China Reconstructs* and the *Far Eastern Economic Review*.

(P.–C.K.)

Nansen, Fridtjof

One of the most imaginative and intrepid of Arctic explorers and an oceanographer of note, Fridtjof Nansen was also an artist and statesman. But today he is best remembered for his work for the relief of suffering peoples. Nansen was born at Store-Frøen near Christiania (now Oslo), Norway, on Oct. 10, 1861. He went to school in Christiania where, in 1880, he passed his entrance examination to the university. He chose to study zoology in the expectation that fieldwork would give him the chance of an outdoor life and enable him to make use of his artistic talents. Although scientific work was always closest to his heart, he first attained fame as an explorer.

EB Inc.



Nansen.

Explorer. As a young man Nansen was a great outdoor athlete, an accomplished skater and skier, and a keen hunter and fisherman. He thus developed an excellent physique, which gave him the stamina and endurance needed later on his journeys to the Arctic. In 1882 when he joined the sealing ship "Viking" for a voyage to the Greenland waters, Nansen first saw at a distance Greenland's mighty ice cap. It occurred to him that it ought to be possible to cross it, and gradually he developed a plan, which he announced in 1887. Instead of starting from the inhabited west coast he would start from the east coast, and by cutting off his means of retreat would force himself to go forward. The expedition of six left Norway in May 1888. In Iceland the group joined a sealing ship bound for east Greenland waters, and after various difficulties the crossing started on August 15. After enduring storms and intense cold, they reached the highest point of the journey (8,920 feet) on September 5 and struck the west coast at Ameralik fjord on September 26. They were forced to winter at the settlement of Godthaab, where Nansen took the opportunity to study the Eskimos and gather material for his book *Eskimoliv* (1891; English

Crossing Greenland's ice cap

translation *Eskimo Life*, 1893). The party returned home in triumph in May 1889.

In 1890 Nansen presented before the Norwegian Geographical society a plan for an even more hazardous expedition, and in 1892 he laid it before the Royal Geographical society in London. Having collected evidence showing that the ice of the polar sea drifted from Siberia toward Spitsbergen, he proposed to build a ship of such a shape that it would be lifted but not crushed when caught by the ice and to let this ship freeze in off eastern Siberia in order to be carried from there by the currents. Though his plan was severely criticized by contemporary Arctic explorers, the Norwegian Parliament granted two-thirds of the estimated expenses, and the rest was raised by subscriptions from King Oscar II and private individuals. His ship, "Fram" (*i.e.*, "Forward"; now preserved outside Oslo), was built according to his ideas.

With a complement of 13 men, the "Fram" sailed from Christiania on June 24, 1893. On September 22 she was enclosed by the ice in 78°50′ N, 133°37′ E; she froze in, and the long drift began. She bore the pressure of the ice perfectly. On March 14, 1895, Nansen, being satisfied that the "Fram" would continue to drift safely, left her in 84°4′ N, 102°27′ E, and started northward with dogsleds and kayaks, accompanied by one F.H. Johansen. On April 8 they turned back from 86°14′ N, the highest latitude then yet reached by man, and headed toward Franz Josef Land. As they approached the northern islands progress was hampered by open water and, because of the advanced season, they wintered on Frederick Jackson Island (named by Nansen after the British Arctic explorer) where they stayed from Aug. 26, 1895, to May 19, 1896. They built a hut of stone and covered it with a roof of walrus hides and lived during the winter mainly on polar bear and walrus meat, using the blubber as fuel. On their way to Spitsbergen they encountered Frederick Jackson and his party of the Jackson-Harmsworth expedition, on June 17, and returned to Norway in his ship "Windward," reaching Vardø on August 13. The "Fram" also reached Norway safely, having drifted north to 85°57′. Nansen and his companions on board the "Fram" were given a rousing welcome, which reached its climax on arrival in Christiania on September 9. His two-volume account of the expedition, *Fram over Polhavet* (English translation *Farthest North*), appeared in 1897.

Scientist. Nansen's success as an explorer was due largely to his careful evaluation of the difficulties that might be encountered, his clear reasoning, which was never influenced by the opinions of others, his willingness to accept a calculated risk, his thorough planning, and his meticulous attention to detail. Many of these traits can be recognized in his scientific writings. In 1882 he was appointed curator of zoology at the Bergen museum. He wrote papers on zoological and histological subjects, illustrated by excellent drawings. For one of his papers, "The Structure and Combination of Histological Elements of the Central Nervous System" (1887), the Christiania university conferred upon him the degree of doctor of philosophy. Though the paper contained so many novel interpretations that the committee that had to examine it accepted it with doubt, it is now considered a classic.

Work in oceanography

On his return from the "Fram" expedition in 1896, a professorship in zoology was established for Nansen at the Christiania university, but his interests shifted from zoology to physical oceanography; and in 1908 his status was changed to professor of oceanography. During 1896–1917 he devoted most of his time and energy to scientific work. He edited the report of the scientific results of his expedition and himself wrote some of the most important parts. He participated in the establishment of the International Council for the Exploration of the Sea and for some time directed the council's central laboratory in Christiania. In 1900 he joined the "Michael Sars" on a cruise in the Norwegian Sea. In 1910 he made a cruise in the "Frithjof" through the northeastern North Atlantic; in 1912 he visited the Spitsbergen waters on board his own yacht "Veslemoy," and in 1914 he joined B. Helland-Hansen on an oceanographic cruise to the Azores in the "Armauer Hansen." In 1913 Nansen travelled through the Barents Sea and the Kara Sea to the mouth of the Yenisey River and back through Siberia. He published the results of his cruises in numerous papers, partly in cooperation with Helland-Hansen. His lasting contributions to oceanography comprise improvement and design of instruments, explanation of the wind-driven currents of the seas, discussions of the waters of the Arctic, and explanation of the manner in which deep- and bottom-water is formed.

Nansen also dealt with other subjects: for instance, his *Nord i Tåkeheimen* (2 vol., 1911, English translation *In Northern Mists*, 1911) gives a critical review of the exploration of the northern regions from early times up to the beginning of the 16th century. Nansen's papers are all stimulating, but he wrote so fluently that he was likely to be verbose and his attention to detail made his presentation cumbersome. His great literary talent is, on the other hand, evident in his popular accounts of his journeys, many of which he himself illustrated.

Statesman and humanitarian. As Nansen grew older he became more and more interested in the relations between individuals and nations. In 1905 he took a lively part in the discussion about the dissolution of the union between Norway and Sweden. His attitude may be summarized by his words: "Any union in which the one people is restrained in exercising its freedom is and will remain a danger." On the establishment of the Norwegian monarchy, Nansen was appointed its first minister in London (1906–08). In 1917, during World War I, he was appointed head of a Norwegian commission to the United States and negotiated a satisfactory agreement with the U.S. government about the import of essential supplies to Norway.

At the first assembly of the League of Nations in 1920, the Norwegian delegation was headed by Nansen, who was to remain one of the outstanding members of the assembly until his death. With his tall and bony frame, his snow-white hair, his flowing moustache, and his broad-brimmed hat at a rakish angle, he made a striking figure in Geneva.

Work with the League of Nations

In April 1920 the council of the League of Nations gave Nansen his first great task, appointing him high commissioner responsible for the repatriation from Russia of about 500,000 prisoners of war from the former German and Austro-Hungarian armies. The Soviet government would not recognize the League of Nations but negotiated with Nansen personally, and in September 1922 he reported to the third assembly of the League that his task was completed and that 427,886 prisoners of war had been repatriated.

In August 1921 Nansen was asked by the International Committee of the Red Cross to direct an effort to bring relief to famine-stricken Russia. He accepted, and on August 15 a conference in Geneva, at which 13 governments and 48 Red Cross organizations were represented, appointed him high commissioner of this new venture. On August 27 he concluded an agreement with the Soviet government authorizing him to open in Moscow an office of the "International Russian Relief Executive." Nansen's request to the League for financial assistance was turned down, but by appealing to private organizations and by addressing large public meetings he succeeded in raising the necessary funds.

On July 5, 1922, on Nansen's initiative, an international agreement was signed in Geneva introducing the identification card for displaced persons known as the "Nansen passport." In 1931 the Nansen International Office for Refugees was created in Geneva (after Nansen's death); it cared mainly for the "White" Russians, for the Armenians from Turkey and, later, for the Jews from Germany.

In 1922 Nansen was awarded the Nobel Peace Prize; he used the prize money for the furtherance of international relief work.

Nansen died at his home at Lysaker near Oslo on May 13, 1930.

BIBLIOGRAPHY. JON SØRENSEN, *Fridtjof Nansens saga* (1931; Eng. trans., *The Saga of Fridtjof Nansen*, 1932), is a good, relatively brief biography. See also LIV NANSEN HØYER, *Nansen og verden* (1955; Eng. trans., *Nansen: A Family Por-*

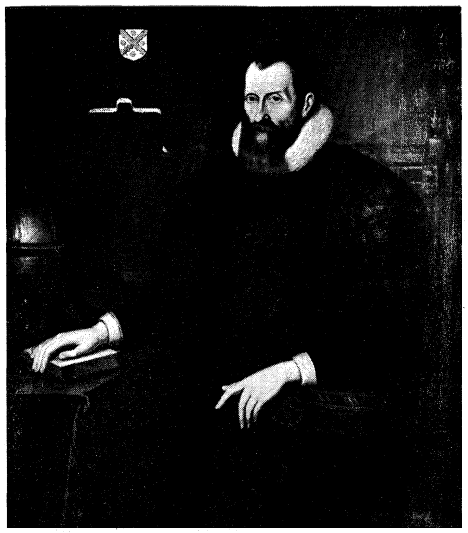
trait by His Daughter, 1957); E.E. REYNOLDS, *Nansen*, 2nd ed. (1949); and A. DE SELINCOURT, *Nansen* (1957).

(H.U.S.)

Napier, John

The invention of logarithms by the distinguished Scottish mathematician, John Napier (Neper), provided science and mathematics with a vastly improved and rapid method of notation and calculation. Active in church-state affairs all his life, he also wrote an influential work in support of Protestantism. He was born at Merchiston Castle, near Edinburgh, in 1550. At the age of 13, he entered the University of St. Andrews, but his stay appears to have been short, and he left without taking a degree.

By courtesy of the University of Edinburgh; photograph, the Scottish National Portrait Gallery



Napier, oil painting by an unknown artist, 1616. In the collection of the University of Edinburgh.

Little is known of his early life, but it is thought that he travelled abroad, as was then the custom of the sons of the Scottish landed gentry. He was certainly back home in 1571, and he stayed either at Merchiston or at Gartness for the rest of his life. He married the following year. A few years after his wife's death in 1579, he married again.

Protestant partisan

Napier's life was spent amid bitter religious dissensions. A passionate and uncompromising Protestant, in his dealings with the Church of Rome he sought no quarter and gave none. It was well known that James VI of Scotland hoped to succeed Elizabeth I to the English throne, and it was suspected that he had sought the help of the Catholic Philip II of Spain to achieve this end. Panic stricken at the peril that seemed to be impending, the general assembly of the Scottish Church, a body with which Napier was closely associated, begged James to deal effectively with the Roman Catholics, and on three occasions Napier was a member of a committee appointed to make representations to the king concerning the welfare of the church and to urge him to see that "justice be done against the enemies of God's Church."

In January 1594, Napier addressed to the king a letter that forms the dedication of his *Plaine Discovery of the Whole Revelation of Saint John*, a work that, while it professed to be of a strictly scholarly character, was calculated to influence contemporary events. In it he declared:

> Let it be your Majesty's continuall study to reforme the universall enormities of your country, and first to begin at your Majesty's owne house, familie and court, and purge the same of all suspicion of Papists and Atheists and Newtrals, whereof this Revelation forthtelleth that the number shall greatly increase in these latter daies.

The work occupies a prominent place in Scottish ecclesiastical history.

Following the publication of this work, Napier seems to have occupied himself with the invention of secret instruments of war, for in a manuscript collection now at Lambeth Palace, London, there is a document bearing his signature, enumerating various inventions "designed by the Grace of God, and the worke of expert craftsmen" for the defense of his country. These inventions included two kinds of burning mirrors, a piece of artillery, and a metal chariot from which shot could be discharged through small holes.

Mathematical contributions

Napier devoted most of his leisure to the study of mathematics, particularly to devising methods of facilitating computation, and it is with the greatest of these, logarithms, that his name is associated. He began working on logarithms probably as early as 1594, gradually elaborating his computational system whereby roots, products, and quotients could be quickly determined from tables showing powers of a fixed number used as a base.

His contributions to this powerful mathematical invention are contained in two treatises: *Mirifici Logarithmorum Canonis Descriptio*, which was published in 1614 (*Description of the Marvelous Canon of Logarithms*, 1857), and *Mirifici Logarithmorum Canonis Constructio*, which did not appear until 1619, two years after his death (*Construction of the Marvelous Canon of Logarithms*, 1889). In the former, he outlined the steps that had led to his invention.

Logarithms were meant to simplify calculations, especially multiplication, such as those needed in astronomy. Napier discovered that the basis for this computation was a relationship between an arithmetical progression—a sequence of numbers in which each number is obtained, following a geometric progression, from the one immediately preceding it by multiplying by a constant factor, which may be greater than unity (*e.g.*, the sequence 2, 4, 8, 16 . . .) or less than unity (*e.g.*, 8, 4, 2, 1, ½ . . .).

In the *Descriptio*, besides giving an account of the nature of logarithms, Napier confined himself to an account of the use to which they might be put. He promised to explain the method of their construction in a later work. This was the *Constructio*, which claims attention because of the systematic use in its pages of the decimal point to separate the fractional from the integral part of a number. Decimal fractions had already been introduced by the Flemish mathematician Simon Stevin in 1586, but his notation was unwieldy. The use of a point as the separator occurs frequently in the *Constructio*. Joost Bürgi, the Swiss mathematician, between 1603 and 1611 independently invented a system of logarithms, which he published in 1620. But Napier worked on logarithms earlier than Bürgi and has the priority due to his prior date of publication in 1614.

Although Napier's invention of logarithms overshadows all his other mathematical work, he made other mathematical contributions. In 1617 he published his *Rabdologiae, seu Numerationis per Virgulas Libri Duo* (*Study of Divining Rods, or Two Books of Numbering by Means of Rods*, 1667); in this he described ingenious methods of multiplying and dividing of small rods known as Napier's bones, a device that was the forerunner of the slide rule. He also made important contributions to spherical trigonometry, particularly by reducing the number of equations used to express trigonometrical relationships from ten to two general statements. He is also credited with certain trigonometrical relations—Napier's analogies—but it seems likely that the English mathematician Henry Briggs had a share in these. He died at Merchiston on April 4, 1617.

BIBLIOGRAPHY. The most exhaustive account of the life and work of John Napier is the *Napier Tercentenary Memorial Volume* (1915), ed. by C.G. KNOTT, which consists of articles on different aspects of Napier's work by a number of writers, each expert in his particular field. There is an English translation of the *Mirifici Logarithmorum Canonis Constructio* by WILLIAM RAE MACDONALD (1889). It contains a catalogue of all Napier's works and is compiled with great care and thoroughness. See also the article "Logarithms," by J.W.L. GLAISHER, in the *Encyclopædia Britannica*, 11th ed., vol. 16; and *An Account of the Life, Writings and Inventions of John Napier of Merchiston by* DAVID STEWART, EARL OF BUCHAN, and WALTER MINTO (1787).

(J.F.S.)

Naples

Naples (Italian Napoli, ancient Neapolis, or "New Town"), the great port city, intellectual centre, and financial capital of southern Italy, lies on the western shore of the Italian peninsula. It is 120 miles southeast of Rome and 210 miles northwest of the country's southernmost point. Situated between two furnaces of volcanic activity, Mt. Vesuvius (*q.v.*) on the west and the Campi Flegrei (Phlegraean Fields) on the east, the city curves around one of the most beautiful and celebrated bays of the world and climbs the encircling hills. The 18-mile-wide entrance to the bay from the Tyrrhenian Sea is flanked by the islands of Ischia (to the west) and Capri (to the east).

Naples, sometimes characterized as the most mysterious of European cities, has a multinational, multicultural, multilevel heritage, the complexity of which can be discerned in its streets, buildings, social attitudes, and customs. The past continually thrusts through the surface of the present, and history is an essential guide to modern Naples, which was the home of some 1,259,000 people at the 1971 census.

HISTORY OF NAPLES

Foundations and the Roman era. It seems probable that Neapolis, the "new town," was founded in about 600 BC by Greek colonists, but the community was captured by the Romans in 326 BC and subsequently played an important role as a Roman ally in the wars of the period. Under the Roman Empire the settlement was enriched with temples, gymnasiums, baths, aqueducts, hippodromes, and arenas; during the time of Marcus Tullius Cicero, the great Roman man of letters, many upper-class Romans were attracted to Neapolis, not least because of its Greek manners and way of life. It was also the favourite residence of many of the emperors: Nero made his first appearance on the stage in one of its theatres, and Marcus Aurelius lived in the city and frequented its schools. Poets who were associated with the city included Virgil, who composed his *Georgics* there. The numerous catacombs outside the city were built in the later centuries of Roman rule, with the earliest Christian foundations there dating from the 4th century AD.

Classical associations

The medieval and modern periods. In the wake of the barbarian invasions that accompanied the collapse of Rome, Naples was conquered in 536, acknowledged Byzantine suzerainty after 552, and, gradually experiencing a return to prosperity, became a Byzantine dukedom in the 7th century. The city was soon fighting to preserve its independence against attacks by Lombards, Saracen pirates, and Normans. It finally capitulated to the feudal Norman kingdom of Sicily in 1139, and its intellectual climate (where Latin had now replaced Greek), as well as its prosperity, subsequently suffered. A period of dynastic manoeuvring ensued, culminating in the city's asserting its independence again by welcoming Charles I of Anjou in 1266. He made Naples his capital, and, during his reign and those of his Angevin and Aragonese successors, it increased in size, population, and prosperity, becoming famous for the number and beauty of its palaces and churches. Following further succession disputes, Naples became a possession of the Spanish Habsburgs, whose viceroys ruled, often autocratically, from 1503 to 1704. The city developed considerably during this period, its street pattern was restructured, and it weathered both epidemics (notably that of 1656) and unrest (notably the dramatic 1647 uprising led by Tomaso Aniello, a fisherman better known as Masaniello). In 1734 Naples became the capital of an independent kingdom of Sicily, undergoing further embellishments at the hand of Prince Charles of Bourbon (later Charles III of Spain). In 1799 his son Ferdinand was ousted by a French revolutionary army, and the Parthenopean Republic was proclaimed. In 1806 Napoleon's brother Joseph Bonaparte was crowned king of Naples, but in 1815 a turbulent period was ended with the restoration of Ferdinand under the title "king of the two Sicilies." The city played its part in the movement toward Italian unity, finally becoming a part of the larger entity in 1860. The 19th and 20th centuries were marked by increasing industrialization, with associated symptoms of urban squalor and unrest, and the city suffered heavily in air attacks and ground fighting associated with the campaigns of World War II. The postwar period witnessed a number of attempts to improve conditions of life in the city, which, nevertheless, retains much of its traditional, crowded exuberance and an architectural heritage reflecting its complex past.

THE CONTEMPORARY CITY

The Neapolitan character. Generations of observers have described Naples as a vast popular theatre, and although there have been changes behind the scenes since the late 1960s, the outward spectacle has not altered much. Against the dramatic backdrop of Vesuvius and the bay, with the seafront embellished by two fortified castles, the decor is still one of moldering palaces and oft-rebuilt churches. The narrow streets (occasionally transformed into staircases) clamber up hillsides covered increasingly by new buildings. Neapolitans still live intensively in the street, which remains the market, playground, workshop, and living room for a good part of the population. The spectacle is colourful and extraordinarily animated. The casual visitor could regard it as a high-decibel comic opera, but to the actors it was, in the past, a real-life melodrama, often a tragedy. Today suffering is less wide-spread and less acute. Until well past the middle of the 20th century the mass of Neapolitans lived the way the urban poor all over Europe had lived for centuries. They were poor without much hope of changing their lot, except through emigration or crime, both of which flourished.

Poverty as a traditional way of life

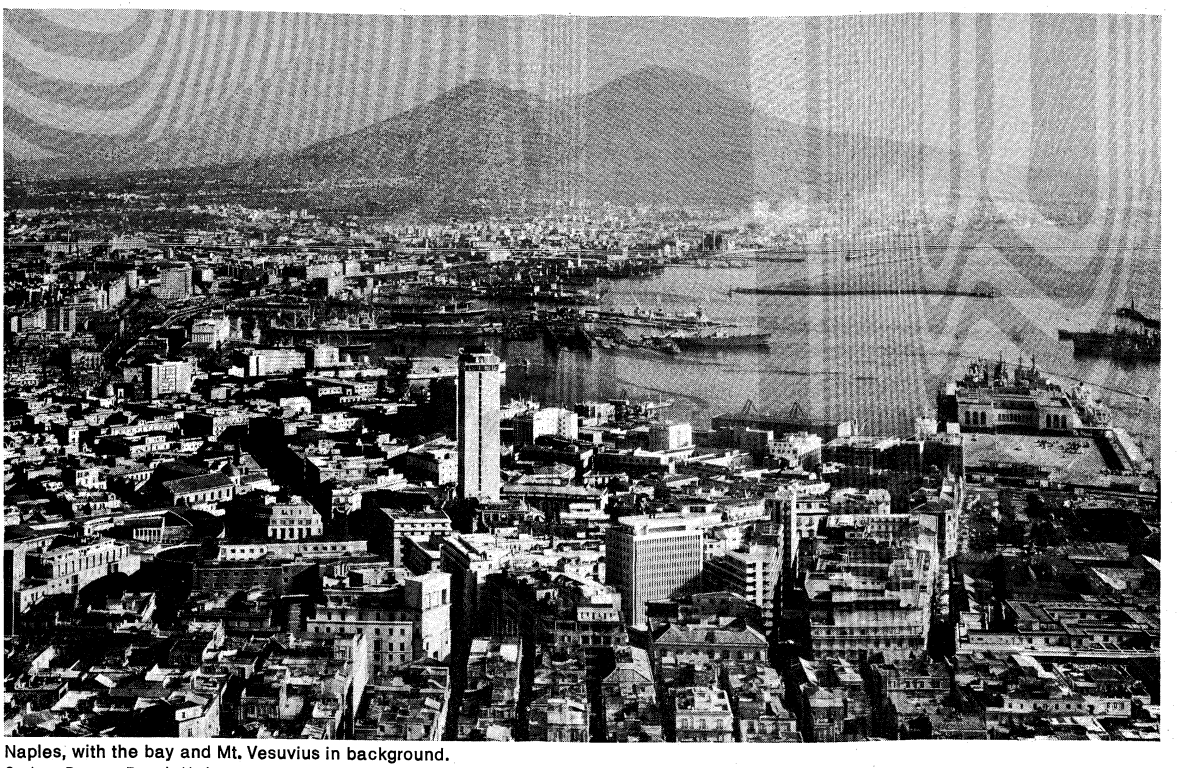
Two refuges from suffering were music and religion; the music that the Neapolitans created themselves was often gay, and their religion was passionate and personal. Their strength came from the family and spread to embrace the neighbourhood and finally the whole community. Their local patois became so developed that it was virtually a separate language. The little marvels of local cuisine—different kinds of pizza, pasta, and shellfish dishes—were all ways of making poverty more digestible.

Most of the time workers made ends meet somehow without regular employment. As late in the 20th century as 1951, when the population was 1,011,919, only 200,431 Neapolitans had regular jobs. This meant that there were fewer wage earners in the city than there were families (222,047). Since that time, things have improved noticeably, not to the point of general prosperity but at least above the level of general despair.

As prosperity and progress expand, the picturesque evaporates. Now that so many people can afford radios, television, and tickets to the cinema and now that the tourist industry is an important part of the economy, the Neapolitan "language" is disappearing. The gaily painted market cart is vanishing in favour of the three-wheeled delivery van, the street musician is rarely seen, and the hurdy-gurdy man is almost extinct.

Most neighbourhoods still string garlands of coloured electric bulbs, festoon the church, and open the festival of the parish patron saint with a parade of relics and banners through the streets; but the celebrations are less fervent and of shorter duration. The festival of our Lady of Piedigrotta still occupies six days of the calendar, and the streets around the church are illuminated, but most of the merrymaking and fireworks are confined to a single night. The parade of floats and the song and dance contests have been removed to a public park two miles away to keep the main street clear.

Neapolitans seem to revel in human contact. What could be deplored elsewhere as overcrowding has long been normal there. In their centuries of bare survival Neapolitans learned how to make inhuman conditions tolerable by preserving respect for the individual. In a society of nobodies, everybody, it appears, is somebody. No one is excluded as unworthy: children are to be seen and heard, and the elderly are welcome to act as if they were still in charge.



Naples, with the bay and Mt. Vesuvius in background.
Gaetano Barons—Bavaria-Verlag

In maintaining their mutual respect, in keeping their family and neighbourhood structure, in maintaining their gaiety and enterprise, in continuing a lively dialogue with their saints, the people of Naples may have found the solution to a problem that bedevils other metropolitan populations: how to live happily under modern urban conditions.

Another folk tradition, disrespect for authority, is still lively. Now that Neapolitans have automobiles and motor bicycles, they roar up one-way streets the wrong way and make forbidden turns, taking care only to avoid squashing street urchins even more heedless than they. On street corners, properly clad matrons hold market baskets of contraband cigarettes for sale.

The role of crime

Thievery, most of it petty, is another Naples tradition with many adepts. The local tourist board has warned visitors that parked automobiles are likely to be rifled, if not stolen. Handbag snatching in the early 1970s was epidemic during the tourist season. Yet, in contrast with some other great Western cities, violent crime against persons is fairly rare in Naples. It was not always thus. Murder in the streets was common until the way was lighted in about 1770 by hundreds of street-corner shrines, privately installed following a campaign by Father Rocco, a Dominican priest whose objectives were as much civic as religious.

The blind violence and criminality of the slums were organized in the 19th century by the Camorra, a secret society somewhat similar to the Sicilian Mafia. The liberation of Naples in 1943 was the occasion of the birth of a new Camorra, dominating what may have been the biggest black-market operation in Europe.

Planning and the urban economy. The improvement visible in Naples is the result of the concerted action of state planning and fiscal agencies and of companies either owned or controlled by the national government. Local enterprise is associated with the programs. New roads, port installations, modernized transport and communications, and expansion of energy production furnished the base for restructuring the economy of the region. A wedge of land between the freight yards and the waterfront in the east end of town was developed as an industrial zone. Existing industry was modernized and enlarged, and new industries were introduced.

A small but significant example was the Municipal Milk Central, which, after modernization of methods and machinery in 1966, was able to double production and sales and eliminate losses in three years. This implies an improvement in the health and purchasing power of the community, as well as an expansion of dairy farming in the region.

Under the Bourbons, Naples had an early start in industry, with foundation of the porcelain factory at the royal palace of Capodimonte in 1740 and the development of silk and other textile production soon after. In 1818 the first steamboat on the Mediterranean was launched at the royal shipyards. Remodelling of the port was begun in 1826. In 1839 Italy's first railway chugged along the five miles from Naples to the court resort of Portici. The first funicular railway on the peninsula was opened to the heights of Vomero in 1880 to the lilt of the new, prize-winning song from the Lady of Piedigrotta festival, "Funiculì-Funiculà." The first steel mill opened at the end of the 19th century, but the industry did not add significantly to national production until the 1970s. Among the newer industries in the region are electronics and petroleum products. The tourist industry, aided by state financing and planning, also expanded at the end of the 1960s. The crescent of the bay is set with tourist gems—Capri, Ischia, Vesuvius, Herculaneum, Pompeii, and Sorrento —and just to the south on the Gulf of Salerno are Amalfi and Paestum. The greatly augmented number of ferries, high-speed boats (*alsiscafi*), helicopters, and motor coaches serving tourists add to the air pollution already heavy from industrial fallout. The famous view is veiled with smog, except when the wind veers from its prevailing direction.

Water pollution in Naples is as bad as it is in every major port city in Italy. The waters that were once the siren Parthenope's home are closed to swimmers, and the celebrated shellfish are infected with hepatitis.

Topography and architecture. The old volcanic site of the city is pierced with man-made galleries, some of which have yet to be fully explored and some of which were opened by the Romans. During World War II the tunnels served as bomb shelters. Today some uncharted passage occasionally gives way, tumbling houses and leaving gaping craters in the streets: from 1959 to 1971 some 200 Neapolitans lost their lives in such collapses. The best known of the grottoes are the catacombs of S.

The catacombs

Gennaro and S. Gaudioso—some of them sculptured and frescoed—in the northern slopes, and the traffic tunnels (of very early origin) passing under Collina di Posillipo (Posillipo Hill) from the western edge of town.

At the foot of this hilly peninsula is the little port of Mergellina, from which the four miles of urban coastline curve westward to the giant commercial port. Along the first stretch toward the centre of town lies the 1,000-yard-long Villa Comunale, a somewhat seedy but happy public park that contains the Aquirio (Aquarium) and several outdoor musical cafés, which are augmented by scores of ambulant vendors of food, drink, and toys. Families promenading beneath the dusty trees amid the cheery clamour bring a turn-of-the-century innocence to the 1970s. The seafront is bordered by the sweeping Via Caracciolo, while along the inland border of the park runs the Riviera di Chiaia, which denotes the name of the whole neighbourhood. The name *chiaia* derives from the Latin *plaga* ("beach, flatland"), from which the French extracted *plage* and the Italians *spiaggia,* a development too simple for Naples.

The Chiaia was a favourite residence of foreigners, particularly the British, in the late 18th and early 19th centuries. The Museo Principe Diego Arangona Pignatelli Cortes, full of period furnishings, was begun as the residence of Sir John Acton, prime minister for the last Bourbon king.

Today one of the choice residential sections is Vomero, on the heights just above. Three funiculars descend to different quarters of the older part of the city. A spur off the Vomero heights was fortified in 1329 by King Robert of Anjou, and the installation was rebuilt in the form of a six-pointed star by the Spanish in 1537. This vast mass of masonry, known as the Castel Sant'Elmo, which still dominates Naples, was a key to the city's defense for centuries. The adjacent Museo di S. Martino was built in the 14th century as a monastery and largely remade in the 16th and 18th centuries. Its collections are concerned mostly with the city itself (of which there is a magnificent view from Room 25), including a Nativity scene peopled by hundreds of the traditional, locally made clay figurines.

From Chiaia the waterfront road, changing its name to Via Partenope, swings through the ancient *borgo marinaro* of Sta. Lucia below the headland of Pizzofalcone. Its picturesque squalor has given way to luxury hotels that look out over the tiny harbour bordered by yacht clubs and the renowned dockside restaurants. The west side of the harbour is formed by a causeway, the south side by the island the ancients called Megarus (Isola di Megaride), where Lucullus had one of his several Neapolitan villas. In the 12th century the Normans built a fortified castle there, expanded by their successors, the Hauenstauffens, and further enlarged and embellished by the Angevin kings of the 13th century. The hulking, sheer-walled, adobe-coloured bulk is today a military installation.

A little farther along, past the bathing beach and another public park (Giardini Litoranei), is the major port for the ferries that crisscross the bay, and just beyond is the New Maritime Terminal for ocean liners. The importance of the passenger port (until then Italy's largest) was greatly reduced in the 1970s with curtailment of many shipping services.

The Castel Nuovo

Up on the shore, to the west of the big Piazza Municipio, is the moated, turreted Castel Nuovo (New Castle), built in 1279–82 by Robert I of Anjou. It is also often called the Maschio Angioino (Angevin Fortress), one of the many local landmarks to bear two names, used interchangeably. In the mid-15th century the castle was extensively rebuilt by Alfonso I of Aragon, who also added a sculptured marble triumphal arch to the entry, one of the most distinguished Renaissance works in Naples. Still loyal to the Angevins, the barons of Apulia in 1486 mounted a conspiracy against the Aragonese. Ferdinand I, also called Ferrante, who reigned from 1458 to 1494, announced the betrothal of his niece to the son of one of the plotters and invited them all to a banquet at the castle, from which they did not emerge alive. The banquet room, called the Barons' Hall, is now the meeting place of the Naples Municipal Council. The municipal offices are at the opposite end of the oblong Piazza Municipio, in a building erected in 1819 for the ministerial offices of Ferdinand I (Ferdinand IV of Naples).

The castle grounds extend westward to abut the garden of the 17th-century Palazzo Reale (Royal Palace), through which lies the entrance to the Biblioteca Nazionale (National Library). Having grown out of the Farnese collection, the library contains many prizes, including the papyrus discovered at Herculaneum in 1752. Also in the palace are government offices, some restored state apartments, and a 1768 court theatre. The palace, which was built by Domenico Fontana for the Spanish viceroy, was damaged by fire in 1837, again in 1942 by aerial bombs, and again, more seriously, by Allied military occupation. The facade on the Piazza del Plebiscito remains almost as it was originally designed, with the addition of statues representing eight of Naples' ruling dynasties, from Roger Guiscard to Victor Emmanuel II of Savoy and Italy.

The piazza is now devoted in large part to service as an open-air bus terminal. The entire semicircle of the western side is occupied by a colonnade erected by Murat in 1810, in the centre of which King Ferdinand I placed, seven years later, the church of S. Francesco di Paola, fulfilling a vow he had made when praying for the return of his kingdom.

North of the piazza is the smaller Piazza Trieste e Trento. It is also called the Piazza S. Ferdinando because of the 17th-century church (like most Neapolitan churches, extensively rebuilt later) of that name, where the "Stabat Mater" of Pergolesi, written for the archconfraternity of the church, is performed every Good Friday. Below the church and adjacent to the Royal Palace is the renowned Teatro S. Carlo, home of Neapolitan opera. It was built between March and November 1737; the facade was added in 1810, and the interior was rebuilt after a fire in 1816. The acoustics of the six-tiered interior are considered perfect for bel canto: Rossini, Bellini, and Donizetti all wrote operas for the S. Carlo.

The Teatro S. Carlo

Across the street is one of the entrances to the cruciform Galleria Umberto I. The arcades of shops and offices, adrip with carvings, rise to a barrel-vaulted roof of glass on iron trellises. The mosaic pavements are dotted with busy café tables.

One of the galleria's entries opens onto the Via Roma, still often called the Toledo, the name the street bore for centuries after its construction by the Spanish viceroy in 1536. To the west of this main avenue he constructed the dozen tiny streets of quarters for his soldiers. Today's inhabitants of the *quartieri* overflow into the narrow, stepped streets, the *gradini.*

At the top of the street and a little to the right is the Museo Archeologico Nazionale, a red brick pile thrown up in 1586 as a cavalry barracks, used by the university in 1616–1777, and in 1790 remodelled to receive the royal collections, the nucleus of which were the Farnese collection, inherited by Charles III from his mother, Isabella of Spain, born Elisabetta Farnese. Charles had begun construction of Capodimonte palace on the hills above the city in 1738 to house the family treasures; but the discoveries at Herculaneum, Pompeii, and other nearby sites soon overflowed the palace, and everything was moved to the museum in town, remodelled for its new function. The Biblioteca Nazionale was removed to the Palazzo Reale and the National Gallery of Paintings to Capodimonte (1957). The Farnese marbles, most of them looted from the public baths of Rome, remain, as do the evocative murals from the villas of Pompeii, Herculaneum, and Stabbia, the Pompeiian mosaics, the Herculaneum bronzes, and a superb sampling of Greek and Roman art.

The Bourbons built the Piazza Dante in the mid-18th century halfway up the Toledo, where a gateway had been erected in 1625 and rebuilt in 1797. The Via Porta Alba behind the gate was once crowded with pizzerias, and many an expatriate establishment in the Italian section of American cities still bears the nostalgic name.

Directly eastward the Via S. Pietro a Maiella begins, named for the recently restored church (late 13th–early 14th century) dedicated to S. Pietro Angeleri, a cleric from the Montagna della Maiella (Maiella Mountains) who became Pope Celestine V in 1294, abdicated after six months, and died in 1296 as a prisoner of his successor, Boniface VIII. The former monastery adjacent to the church became the famous Conservatory of Music in 1816, comprised of four other existing schools of music. Its origins ultimately trace back to 1537, making it the oldest conservatory in the Occident. In the Teatro Bellini (1874), a few streets to the north, a local tenor named Enrico Caruso made his debut in 1894.

Going east the street changes its name to the Via Tribunali, at the far end of which is the Naples tribunal, where local lawyers give performances perhaps surpassed only by the singers at the S. Carlo. The courts building is known as the Vicaria and as the Castel Capuano, for it was founded by the Norman William I, enlarged and fortified by Frederick II, and frequently used as a royal residence by King Ladislas and by Queen Joan II. Behind the castle is the beautiful Renaissance gateway (1484)—one of Italy's finest—designed by Giuliano da Maiano. The piazza between the castle and the gate is one of the big outdoor markets for the farmers who bring in their produce from nearby Poggioreale. The next gateway to the south, Porta Nolana, marks the beginning of the fish market along the Via Carmignano, which terminates at the church of Sta. Maria del Carmine, a 12th-century foundation since restored many times (campanile, 1631; facade, 1766). Connected with the Masaniello revolt of 1647, it is still the church of simple fisherfolk and sailors, its cloisters hung with crude and touching ex-votos.

Gateways and the markets

The Via Tribunali was the middle one of the three *decumani* of ancient Neapolis. The upper, just to the north, is now the Via dell' Anticaglia. South of the Via Tribunali the lower *decumanus* changes its name many times in its east–west course of about one mile—the vias Pasquale Scura, Domenico Capitelli, Benedetto Croce, Largo Corpo di Napoli, S. Biagio dei Librai, Vicaria Vecchia, and Forcella—but is usually referred to by one name, as a single entity: Spaccanapoli, the street that sliced ancient Naples in two. At right angles to the three axes, small streets ran down toward the harbour. The layout unchanged, this antique, original city—little expanded in area until the Angevin epoch—is still the heart of old Naples. After the cholera epidemic of 1884, Premier Agostino Depretis decreed the *sventramento* ("gutting") of the slums of the Porto and Mercato districts, and from the Piazza Bovio a straight, ugly street was driven to slant up to the central railroad station. It was named the Corso Umberto I, but the people call it the Rettifilo.

The site of Neapolis is sprinkled with churches (more than 50 of the city's 135 ancient churches) and palaces (perhaps 100). Along the Via Tribunali there are, for example, seven churches. One of these is the recently restored French Gothic church of S. Lorenzo Maggiore, established by Charles I of Anjou on the site of a 6th-century church and built over a corner of the Greco-Roman forum of Neapolis, remains of which lie underneath. Another is S. Paolo Maggiore, 1591–1603, which incorporates in its facade Corinthian columns of a Greek temple to the Dioscuri that once stood on the site.

Crossing Tribunali and the other *decumani*, the Via del Duomo runs down to the bay. The cathedral Duomo was built 1294–1323 on the site of the 5th-century basilica of Stefania. It adjoins the 4th-century basilica—the first in Naples—of Sta. Restituta, containing the small baptistery of S. Giovanni in Fonte, which has some 5th-century mosaics. Both the cathedral and Sta. Restituta have suffered earthquake and restoration, and a good deal of what can be seen is of 18th- and 19th-century origin. The facade retains its original doorways, however. In the summer of 1971 remains of a 5th-century BC Greek temple were found under the floor.

The cathedral is dedicated to S. Gennaro (St. Januarius), the patron saint of the city. His two festivals, the first Saturday in May and the seven days that follow September 19, the day of the saint's decapitation in AD 305, are genuinely popular festivals and genuinely festive. The saint's reliquary is paraded through the decorated streets, and the church is jammed for the liquefaction of his blood. If the congealed blood, contained in two phials, fails to liquefy, hordes of Neapolitans expect ill fortune for their city. They recall that in 1527 there was a plague, in 1569 a famine, and in 1941 eruption of Vesuvius and aerial bombardment.

The festivals of San Gennaro

Along the Spaccanapoli there are four important churches and an unusually curious chapel. The Gesù Nuovo was, in the 15th century, the imposing palace of the Sanseverini, princes of Salerno, its facade an unbroken wall of white volcanic basalt and each stone shaped into a raised point. Its conversion into a Jesuit church was begun in 1584 and was completed in the 17th century. Across the way is the monastery of Sta. Chiara (now public offices), built (1310–28) for the Friars Minor as part of the Franciscan enclave established by Angevin King Robert and Queen Sancia of Majorca. Next to the church is the convent of the Poor Clares, the 14th-century cloister decorated with 18th-century majolica tiles. The church itself, approached through a 14th-century arch, burned for 48 hours after an incendiary bombing by the Allies in 1943. Reconsecrated in 1953, it has been restored to its original austere French Gothic. Some Angevin tombs, magnificent although bomb damaged, are still to be seen.

Farther east along the Spaccanapoli is the Dominican monastery where St. Thomas Aquinas, whose cell still exists, studied as a youth and returned to teach (1272–74). Among the students of a later age were the humanist Giovanni Pontano and the philosopher Giordano Bruno. In the adjacent church of S. Domenico Maggiore the main chapel holds the 13th-century painting of the crucifix that, it is said, spoke to St. Thomas, asking what he desired as a reward for his religious labours; he replied, "None other than thee." The 1289 church, incorporating a much older house of worship, has been rocked by earthquakes, gnawed by fires, and mutilated by restorers; it possesses a Titian "Annunciation" and Caravaggio's "Flagellation."

The much-restored apsidal end of the church presents its crenellations, patched piers, and walls to the Piazza S. Domenico, which, like the space before Gesù Nuovo, is decorated with a guglia (memorial column), this one to mark the end of the 1656 plague. Also on the piazza is one of the most noted coffee bars of Naples, a city that offers the best coffee in the world, according to natives, who, unlike the Viennese, drink it down without ceremony. The piazza is surrounded by four palaces: one built for the Del Balzo family in the 15th century (Del Balzo is Neapolitan for the noble Provençal name des Baux), the others for branches of the Sangro family, whose burial chapel is just off the piazza. Built at the end of the 14th century, the Capella Sansevero, also called Sta. Maria della Pietà dei Sangro, is noted for its implacably inspirational sculptures of the 18th century. Equally celebrated in Naples are the seven life-size, realistic terra-cotta figures grouped about a full-size, outstretched Christ (the 1492 "Pietà" of Guido Mazzoni) on the pavement in the right transept of Sant'Anna dei Lombardi (or Monteoliveto) to the southeast. The church, erected in 1414 and favoured by the Aragonese, contains much fine Renaissance sculpture. The Monteoliveto monastery is today a police barracks.

The Piazza S. Domenico

The Via Mezzocannone, which leads south out of the Piazza S. Domenico to the Rettifilo, borders the university, which was founded in 1224 by Emperor Frederick II for the purpose, as one critic wrote, of training "good Ghibelline clerks who could counter the influence of Popes and northern Guelfs." In addition to the original 18th-century installations in an old Jesuit monastery and the new (1887–1908) extensions, the University has some of its ten faculties located elsewhere. The Palazzo Giusso, near the church of S. Giovanni Maggiore, houses the Istituto Universitario Orientale (Oriental Institute), and the Palazzo Gravina (1539–41) is the School of Architecture.

The Palazzo Filomarino, rebuilt in 1512 on that part of

the Spaccanapoli that bears the name of Benedetto Croce, was the home of the scholar until his death in 1952. He bequeathed his excellent library and the palace to the Istituto Italiano per gli Studi Storici (Italian Institute for Historical Studies). To the right of the Filomarino is the 15th-century Palazzo Carafa della Spina, and farther down the Spaccanapoli are two other palaces of the same period belonging to members of the same family, the Carafa di Montorio and the Palazzo di Diomede Carafa. In the Via del Duomo the Renaissance Palazzo Duomo is now the Museo Civico Filangieri (Filangieri Civic Museum), with a collection of paintings, sculpture, and porcelains.

The housing problem

Scores of other palaces have been renovated by the municipality and by such local firms as the Bank of Naples. Many of the noble frontages are still pricked with little workshops that are literally holes in the wall, and most palaces are carved up into apartments, some among the most desirable in the city and others dark and dank with whole families jammed into *bassi* (one-room ground-floor apartments). It is thus that the historic heart of the city still beats vigorously, with few signs of the morbidity visible in other cities. It would seem to indicate that Neapolitans accept privations to cling to the old (and "old" is counted in millennia here) neighbourhood, but these are the same people who make 2,000 applications for every vacancy in modern public-housing developments.

Some of the new public housing has been erected on the hills that rise behind the city and that were until the end of the 1960s largely wooded areas and green fields. At the farthest and highest point, above the observatory (1819) and the catacombs, is the palace of Capodimonte in its splendid park. Started by Charles the III in 1738, it was finished a century later, a favourite royal residence. The art collections are mainly mediocre, possibly enhancing the impression made by the few masterpieces: Titian's "Danae with Nursemaid," Pieter Brueghel the Elder's "The Parable of the Blind," Jan Brueghel's "Winter Scene," a Masaccio "Crucifixion," two Goya portraits, and drawings by Michelangelo, Tintoretto, Raphael, and Rembrandt. The Farnese collection of arms and armour and the porcelain exhibitions are also of first quality.

The hilltop route from Capodimonte to Vomero once wound along country roads under chestnut trees through rural settlements; since the late 1960s it has been built up, including luxury housing erected illegally in green-belt reserves. It was reported at the beginning of 1972 that the desecration of the heights by profit seekers had been in vain: 80,000 new, high-priced rooms remained vacant.

BIBLIOGRAPHY

General: E. HUTTON, *Naples and Campania Revisited* (1958), a thorough treatment of the various districts of the city and its museums, galleries, churches, and palaces, with a sketch of the historical background; PETER GUNN, *The Companion Guide to Southern Italy* (1969), of which the early part deals with Naples and its environs, including Vesuvius, Herculaneum, and Pompeii.

The past: PETER GUNN, *Naples: A Palimpsest* (1961), largely a popular history, although there is some discussion of contemporary Naples; HAROLD ACTON, *The Bourbons of Naples, 1734–1825* (1956) and *The Last Bourbons of Naples, 1825–1861* (1961), a scholarly and detailed history of the Kingdom of Naples before it became part of the united Italy.

Chiefly photographic: KURT OTTO-WASOW, *The Bay of Naples,* with introduction and commentaries by WILLIAM SANSON (1960), colour photographs that stress the beauty of the city; JAN LUKAS, *Naples,* with an introduction by PETER GUNN (1965), buildings, streets, and monuments, with scenes from the daily life of the people; CAS OORTHUYS, *This Is Naples, Capri, Pompeii, Paestum* (1964), a compact paperback with evocative photographs of the city and its surroundings that includes explanatory notes.

Special topics: MORRIS WEST, *Children of the Sun* (1957), life in the slums of Naples and charitable work among the street urchins by Father Mario Borrelli, founder of the Casa dello Scugnizzo (House of Urchins); MARIO BORRELLI, *A Street Lamp and the Stars* (1963), his autobiography, written with the help of ANTHONY THORNE.

(B.E.)

Napoleon

Napoleon Bonaparte, a French general, as first consul and then as emperor of the French, carried out many reforms that left a lasting mark on the institutions of France and of much of western Europe. But his driving passion was the military expansion of French dominion, and, though at his fall he left France smaller than it had been at the outbreak of the Revolution in 1789, he was almost unanimously revered during his lifetime and until the end of the Second Empire under his nephew Napoleon III as one of history's great heroes.

Napoleon was born at Ajaccio on the island of Corsica on August 15, 1769, shortly after the cession of Corsica to France by the Genoese. He was the fourth, and second surviving, child of Carlo Buonaparte, a lawyer, and his wife, Letizia Ramolino. His father's family, of ancient Tuscan nobility, had emigrated to Corsica in the 16th century.

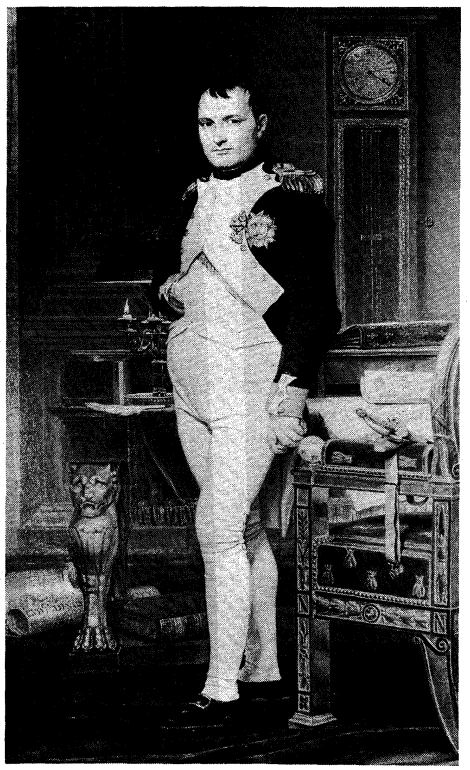
Carlo Buonaparte had married the beautiful and strong-willed Letizia when she was only 14 years old; they eventually had eight children to bring up in very difficult times. The French occupation of their native country was resisted by a number of Corsicans led by Pasquale Paoli. Carlo Buonaparte joined Paoli's party, but when Paoli had to flee, Buonaparte came to terms with the French. Winning the protection of the governor of Corsica, he was appointed assessor for the judicial district of Ajaccio in 1771. In 1778 he obtained the admission of his two eldest sons, Joseph and Napoleon, to the Collège d'Autun.

Early years and education in France

A Corsican by birth, heredity, and childhood associations, Napoleon continued for some time after his arrival in Continental France to regard himself a foreigner; yet from the age of nine he was educated in France as other Frenchmen were. While the tendency to see in Napoleon a reincarnation of some 14th century Italian *condottiere* is an overemphasis on one aspect of his character, he did, in fact, share neither the traditions nor the prejudices of his new country: remaining a Corsican in temperament, he was first and foremost, through both his education and his reading, a man of the 18th century.

Napoleon was educated at three schools: briefly at Au-

Giraudon



Napoleon, portrait by Jacques-Louis David, 1812. In the Samuel H. Kress Collection, National Gallery of Art, Washington, D.C.

tun, for five years at the military college of Brienne, and finally for one year at the military academy in Paris. It was during Napoleon's year in Paris that his father died of a stomach cancer in February 1785, leaving his family in straitened circumstances. Napoleon, although not the eldest son, assumed the position of head of the family before he was 16. In September he graduated from the military academy, ranking 42nd in a class of 51.

He was made second lieutenant of artillery in the regiment of La Fère, a kind of training school for young artillery officers. Garrisoned at Valence, Napoleon continued his education, reading much, in particular works on strategy and tactics. He also wrote *Lettres sur la Corse*, in which he reveals his feeling for his native island. He went back to Corsica in September 1786 and did not rejoin his regiment until June 1788. By that time the agitation that was to culminate in the French Revolution had already begun. A reader of Voltaire and of Rousseau, Napoleon believed that a political change was imperative, but as a career officer he seems not to have seen any need for radical social reforms.

THE REVOLUTIONARY PERIOD

The Jacobin years. When in 1789 the National Assembly, which had convened to establish a constitutional monarchy, allowed Paoli to return to Corsica, Napoleon asked for leave and in September joined Paoli's group. But Paoli had no sympathy for the young man, whose father had deserted his cause and whom he considered to be a foreigner. Disappointed, Napoleon returned to France; and in April 1791 he was appointed first lieutenant to the 4th regiment of artillery, garrisoned at Valence. He at once joined the Jacobin club, a debating society initially favouring a constitutional monarchy, and soon became its president, making speeches against nobles, monks, and bishops. In September 1791 he got leave to go back to Corsica again for three months. Elected lieutenant colonel in the national guard, he soon fell out with Paoli, its commander in chief. When he failed to return to France, he was listed as a deserter in January 1792. But in April France declared war against Austria and his offense was forgiven.

Apparently through patronage, Napoleon was promoted to the rank of captain but did not rejoin his regiment. Instead he returned to Corsica in October 1792, where Paoli was exercising dictatorial powers and preparing to separate Corsica from France. Napoleon, however, joined the Corsican Jacobins, who opposed Paoli's policy. When civil war broke out in Corsica in April 1793, Paoli had the Buonaparte family condemned to "perpetual execration and infamy," whereupon they all fled to France.

Napoleon Bonaparte, as he may henceforth be called (though the family did not drop the spelling Buonaparte till after 1796), rejoined his regiment at Nice in June 1793. In his *Souper de Beaucaire*, written at this time, he argued vigorously for united action by all republicans rallied round the Jacobins, who were becoming progressively more radical, and the National Convention, the revolutionary assembly that in the preceding fall had abolished the monarchy. At the end of August 1793, the National Convention's troops had taken Marseilles but were halted before Toulon, where the royalists had called in British forces. With the commander of the National Convention's artillery wounded, Bonaparte got the post through the commissioner to the army, Antoine Saliceti, who was a Corsican deputy and a friend of Napoleon's family. Bonaparte was promoted to major in September and adjutant general in October. He received a bayonet wound on December 16, but on the next day the British troops, harassed by his artillery, evacuated Toulon. On December 22 Bonaparte, aged 24, was promoted to brigadier general in recognition of his decisive part in the capture of the town.

Capture of Toulon

Augustin de Robespierre, the commissioner to the army, wrote to his brother Maximilien, by then virtual head of the government and one of the leading figures of the Reign of Terror, praising the "transcendent merit" of the young republican officer. In February 1794 Bonaparte was appointed commandant of the artillery in the French Army of Italy. Robespierre fell from power in Paris on 9 Thermidor (July 27, 1794). When the news reached Nice, Bonaparte, regarded as a protégé of Robespierre, was arrested on a charge of conspiracy and treason. He was freed in September but was not restored to his command. The following March he refused an offer to command the artillery in the Army of the West, which was fighting the counter-revolution in the Vendée. The post seemed to hold no future for him, and he went to Paris to justify himself. Life was difficult on half pay, especially as he was carrying on an affair with Désirée Clary, daughter of a rich Marseilles businessman and sister of Julie, the bride of his elder brother, Joseph. Despite his efforts in Paris, Napoleon was unable to obtain a satisfactory command because he was feared for his intense ambition and for his relations with the "Montagnards," the more radical members of the National Convention. He then considered offering his services to the sultan of Turkey.

The Directory. Bonaparte was still in Paris in May 1795 when the National Convention, on the eve of its dispersal, submitted the new constitution of the year III of the First Republic to a referendum, together with decrees according to which two-thirds of the members of the National Convention were to be re-elected to the new legislative assemblies. The royalists, hoping that they would soon be able to restore the monarchy, instigated a revolt in Paris to prevent these measures from being put into effect. Vicomte Paul de Barras, who had been entrusted with dictatorial powers by the National Convention, was unwilling to rely on the commander of the troops of the interior; instead, knowing of Bonaparte's services at Toulon, he appointed him second in command. Thus, it was Napoleon who shot down the columns of rebels marching against the National Convention (13 Vendémiaire, year IV; October 5, 1795), thereby saving the National Convention and the republic.

Bonaparte became commander of the army of the interior and, consequently, was henceforth aware of every political development in France. He also became the respected adviser on military matters to the new government, the Directory. Lastly, he came to know an attractive Creole, Joséphine Tascher de La Pagerie, the widow of Gen. Alexandre de Beauharnais (guillotined during the Reign of Terror), the mother of two children, and a woman of many love affairs.

From every point of view, a new life was opening for Bonaparte. Having proved his loyalty to the Directory by dissolving a communistic group led by François Babeuf and an Italian, Filippo Buonarroti, whom Bonaparte had known in Corsica, he was appointed commander in chief of the Army of Italy in March 1796. He had been trying to obtain that post for several weeks so that he could personally conduct part of the plan of campaign adopted by the Directory on his advice. He married Joséphine on March 9 and left for the army two days later.

Commander in chief of Army of Italy

Arriving at his headquarters in Nice, Bonaparte found that his army, which on paper consisted of 43,000 men, numbered scarcely 30,000 ill-fed, ill-paid, and ill-equipped men. On March 28, 1796, he made his first proclamation to his troops:

> Soldiers, you are naked, badly fed . . . . Rich provinces and great towns will be in your power, and in them you will find honour, glory, wealth. Soldiers of Italy, will you be wanting in courage and steadfastness?

He took the offensive on April 12 and successively defeated and separated the Austrian and the Sardinian armies and then marched on Turin. King Victor Amadeus III of Sardinia asked for an armistice; and, at the peace treaty in Paris on May 15, Nice and Savoy, occupied by the French since 1792, were annexed to France. Bonaparte continued the war against the Austrians and occupied Milan but was held up at Mantua. While his army was besieging this great fortress, he signed armistices with the duke of Parma, the duke of Modena, and finally with Pope Pius VI.

At the same time, he took an interest in the political organization of Italy. A plan for its "republicanization" by a group of Italian "patriots" led by Buonarroti had to

be shelved when Buonarroti was arrested for complicity in Babeuf's conspiracy against the Directory. Thereafter, Bonaparte, without discarding the Italian patriots altogether, restricted their freedom of action. He set up a republican regime in Lombardy but kept a close watch on its leaders, and in October 1796 he created the Cisalpine Republic by merging Modena and Reggio nell'Emilia with the papal states of Bologna and Ferrara occupied by the French Army. Finally he sent an expedition to recover Corsica, which the British had evacuated.

Defeat of the Austrians and peace terms

Austrian armies advanced four times from the Alps to relieve Mantua but were defeated each time by Bonaparte. After the last Austrian defeat, at Rivoli in January 1797, Mantua capitulated. Next, he marched on Vienna. He was about 60 miles (100 kilometres) from that capital when the Austrians sued for an armistice. By the preliminaries of peace, Austria ceded the southern Netherlands to France and recognized the Lombard republic but received in exchange some territory belonging to the old Republic of Venice, which was partitioned between Austria, France, and Lombardy. Bonaparte then consolidated and reorganized the north Italian republics and encouraged Jacobin—radical republican—propaganda in Venetia. Some Italian patriots hoped that these developments would soon lead to the formation of a single and indivisible "Italian Republic" modelled on the French.

Meanwhile, Bonaparte grew uneasy at the successes of the royalists in the French elections in the spring of 1797 and advised the Directory to oppose them, if necessary, by force. In July it attempted a coup d'etat against the royalists and failed; thereupon Bonaparte sent Gen. Pierre Augereau to Paris, along with several officers and men. Augereau's successful coup d'etat of 18 Fructidor (September 4, 1797) eliminated the royalists' friends from the government and legislative councils and also enhanced Bonaparte's prestige. Thus, Bonaparte could conclude the peace Treaty of Campo Formio with Austria as he thought best. The Directory was displeased, however, because the Treaty had ceded Venice to the Austrians and did not secure the left bank of the Rhine for France. On the other hand, it raised Bonaparte's popularity to its peak, for he had gained victory for France after five years of war on the Continent.

Only the war at sea, against the British, continued. The directors, who wanted to launch an invasion of the British Isles, appointed Bonaparte to command the army assembled for this purpose along the English Channel. After a rapid inspection in February 1798, he announced that the operation could not be undertaken until France had command of the sea. Instead, he suggested that France strike at the sources of Great Britain's wealth by occupying Egypt and threatening the route to India. This proposal, seconded by Talleyrand, the foreign minister, was accepted by the directors, who were glad to get rid of their ambitious young general.

Failures in Egypt

The expedition, thanks to some fortunate coincidences, was at first a great success: Malta, the great fortress of the Knights Hospitallers of St. John of Jerusalem, was occupied on June 10, 1798, Alexandria taken by storm on July 1, and all the delta of the Nile rapidly overrun. On August 1, however, the French squadron at anchor in Abū Qīr Bay was completely destroyed by Adm. Horatio Nelson's fleet in the Battle of the Nile, so that Napoleon found himself confined to the land that he had conquered. He proceeded to introduce Western political institutions, administration, and technical skills in Egypt; but Turkey, nominally suzerain over Egypt, declared war on France in September. To prevent a Turkish invasion of Egypt and also perhaps to attempt a return to France by way of Anatolia, Bonaparte marched into Syria in February 1799. His progress northwest was halted at Acre, where the British withstood a seige, and in May Bonaparte began a disastrous retreat to Egypt.

The Battle of the Nile showed Europe that Bonaparte was not invincible, and Great Britain, Austria, Russia, and Turkey formed a new coalition against France. The French armies in Italy were defeated in the spring of 1799 and had to abandon the greater part of the peninsula. These defeats led to disturbances in France itself. The coup d'etat of 30 Prairial (June 18, 1799) expelled the men of moderate views from the Directory and brought into it men who were considered Jacobins. Yet the situation remained confused, and one of the new directors, Emmanuel Sieyès, was convinced that only military dictatorship could prevent a restoration of the monarchy: "I am looking for a sabre," he said. Bonaparte did not take long to make up his mind. He would leave his army and return to France—in order to save the republic, of course, but also to take advantage of the new circumstances and to seize power. The Directory had, in fact, ordered his return, but he had not received the order, so that it was actually in disregard of his instructions that he left Egypt with a few companions on August 22, 1799. Their two frigates surprisingly escaped interception by the British, and Bonaparte arrived in Paris on October 14.

By this time French victories in Switzerland and Holland had averted the danger of invasion, and the counter-revolutionary risings within France had more or less failed. A coup d'etat could therefore no longer be justified by any need to save the republic. Sieyès, however, had not given up his project, and now he had his "sabre." From the end of October he and Bonaparte were in league together planning the coup, and on 18–19 Brumaire, Year VIII (November 9–10, 1799), it was carried out: the directors were forced to resign, the members of the legislative councils were dispersed, and a new government, the Consulate, was set up. The three consuls were Bonaparte and two of the directors who had resigned, Sieyès and Pierre-Roger Ducos. But it was Bonaparte who was henceforth the master of France.

The coup of 18–19 Brumaire

THE CONSULATE

Consolidation of power. Bonaparte, now 30 years old, was thin and short and wore his hair cut close—*le petit tondu*, the "little crop-head," as he was called. Not much was known about his personality, but people had confidence in a man who had always been victorious (the Nile and Acre were forgotten) and who had managed to negotiate the brilliant Treaty of Campo Formio. He was expected to bring back peace, to end disorder, and to consolidate the political and social "conquests" of the Revolution. He was indeed exceptionally intelligent, prompt to make decisions, and indefatigably hardworking, but also insatiably ambitious. He seemed to be the man of the Revolution because it was due to the Revolution that he had climbed at so early an age to the highest place in the state. He was not to forget it: but more than a man of the Revolution, he was a man of the 18th century, the most enlightened of the enlightened despots, a true son of Voltaire. He did not believe in the sovereignty of the people, in the popular will, or in parliamentary debate. Yet he put his confidence more in reasoning than in reason and may be said to have preferred "men of talent"—mathematicians, jurists, and statesmen, for instance, however cynical or mercenary they might be—to "technicians" in the true sense of the word. He believed that an enlightened and firm will could do anything if it had the support of bayonets; he despised and feared the masses; and, as for public opinion, he considered that he could mold and direct it as he pleased. He has been called the most "civilian" of generals, but essentially he never ceased to be a soldier.

The military dictatorship

Bonaparte imposed a military dictatorship on France, but its true character was at first disguised by the Constitution of the Year VIII (4 Nivôse; December 25, 1799), drawn up by Sieyès. This constitution did not guarantee the "rights of man" or make any mention of "liberty, equality, and fraternity," but it did reassure the partisans of the Revolution by proclaiming the irrevocability of the sale of national property and by upholding the legislation against the émigrés. It gave immense powers to the first consul, leaving only a nominal role to his two colleagues. The first consul—namely, Bonaparte—was to appoint ministers, generals, civil servants, magistrates, and the members of the Council of State and even was to have an overwhelming influence in the choice of members for the three legislative assemblies, though their members

were theoretically to be chosen by universal suffrage. Submitted to a plebiscite, the constitution won by an overwhelming majority in February 1800.

Program of reforms. The Consulate's work of administrative reform, undertaken at Bonaparte's instigation, was to be more lasting than the constitution and so more important for France. At the head of the government was the Council of State, created by the first consul and often effectively presided over by him; it was to play an important part both as the source of the new legislation and as an administrative tribunal. At the head of the administration of the *départements* were the prefects, who carried on the tradition of the intendants of the *ancien régime*, supervising the application of the laws and acting as the instruments of centralization. The judicial system was profoundly changed: whereas from the beginning of the Revolution judges had been elected, henceforth they were to be nominated by the government, their independence assured by their irremovability from office. The police organization was greatly strengthened. The financial administration was considerably improved: instead of the municipalities, special officials were entrusted with the collecting of direct taxes; the franc was stabilized; and the Banque de France, owned partly by shareholders and partly by the state, was created. Education was transformed into a major public service; secondary education was given a semi-military organization, and the university faculties were re-established. Primary education, however, was still neglected.

Bonaparte shared Voltaire's belief that the people needed a religion. Personally, he was indifferent to religion: in Egypt he had said that he wanted to become a Muslim. Yet he considered that religious peace had to be restored to France. As early as 1796, when he was concluding the armistice in Italy with Pope Pius VI, he had tried to persuade the Pope to retract his briefs against the French priests who had accepted the Civil Constitution of the Clergy, which in practice nationalized the church. Pius VII, who succeeded Pius VI in March 1800, was more accommodating than his predecessor, and ten months after negotiations were opened with him a concordat was signed reconciling the church and the Revolution. The Pope recognized the French Republic and called for the resignation of all former bishops; new prelates were to be designated by the First Consul and instituted by the Pope; and the sale of the property of the clergy was officially recognized by Rome. The concordat, in fact, admitted freedom of worship and the lay character of the state.

Religious peace and concordat with Rome

The codification of the civil law, first undertaken in 1790, was at last completed under the Consulate. The code promulgated on March 21, 1804, and later known as the Code Napoléon, gave permanent form to the great gains of the Revolution: individual liberty, freedom of work, freedom of conscience, the lay character of the state, and equality before the law; but, at the same time, it protected landed property, gave greater liberty to employers, and showed little concern for employees. It maintained divorce but granted only limited legal rights to women.

The army received the most careful attention. The First Consul retained in outline the system instituted by the Revolution: recruitment by forced conscription but with the possibility of replacement by substitutes; the mixing of the conscripts with old soldiers; and the eligibility of all for promotion to the highest ranks. Nevertheless, the creation of the Academy of Saint-Cyr to produce infantry officers made it easier for the sons of bourgeois families to pursue a military career. Moreover, the École Polytechnique, founded by the National Convention, was militarized in order to provide officers for the artillery and engineers. Yet Bonaparte was not concerned about introducing new technical inventions into his army. He put his trust in the "legs of his soldiers": his basic strategic idea was a fast-moving army.

Military campaigns and uneasy peace. The First Consul spent the winter and spring of 1799–1800 reorganizing the army and preparing for an attack on Austria alone, Russia having withdrawn from the anti-French coalition. With his usual quick assessment of the situation, he saw the strategic importance of the Swiss Confederation, from which he would be free to outflank the Austrian armies either in Germany or in Italy as he might see fit. His past successes made him choose Italy. Taking his army across the Great St. Bernard Pass before the snow had melted, he appeared unexpectedly behind the Austrian army besieging Genoa. The Battle of Marengo in June gave the French command of the Po Valley as far as the Adige; and in December another French army defeated the Austrians in Germany. Austria was forced to sign the Treaty of Lunéville of February 1801, whereby France's right to the natural frontiers that Julius Caesar had given to Gaul, namely, the Rhine, the Alps, and the Pyrenees, was recognized.

Great Britain alone remained at war with France, but it soon tired of the struggle. Preliminaries of peace, concluded in London in October 1801, put an end to hostilities, and peace was signed at Amiens on March 27, 1802.

General peace was re-established in Europe. The first Consul's prestige increased still more; and his friends—at his suggestion—proposed that a "token of national gratitude" should be offered to him. In May 1802 it was decided that the French people should vote in referendum on the following question: "Shall Napoleon Bonaparte be consul for life?" In August an overwhelming vote granted him the prolongation of his consulate as well as the right to designate his successor.

Consul for life

Bonaparte's conception of international peace differed from that of the British, for whom the Treaty of Amiens represented an absolute limit beyond which they were under no circumstances prepared to go. The British even hoped to take back some of the concessions they had been forced to make. For Bonaparte, on the other hand, the Treaty of Amiens marked the starting point for a new French ascendancy. He was, first of all, intent on reserving half of Europe as a market for France without lowering customs duties—to the indignation of British merchants. To revive France's expansion overseas, he also intended to recover San Domingo (now Hispaniola, which had rebelled), to occupy Louisiana (ceded to France by Spain in 1800), perhaps to reconquer Egypt, and at any rate to extend French influence in the Mediterranean and in the Indian Ocean. Finally, on the Continent of Europe, he advanced beyond France's natural frontiers: incorporating Piedmont into France; imposing a more democratic, decentralized government on the Swiss Confederation; and in Germany compensating the princes dispossessed of territory on the Rhine under the Treaty of Lunéville with shares of the secularized ecclesiastical states.

Great Britain was alarmed by this expansion of France in peacetime and found it scarcely tolerable that one state should command the coastline of the Continent from Genoa to Antwerp. The immediate occasion of Franco-British rupture, however, was the problem of Malta. According to the Treaty of Amiens, the British, who had taken the island on the collapse of the French occupation, should have restored it to the Knights Hospitallers; but the British, on the pretext that the French had not yet evacuated certain Neapolitan ports, refused to leave the island. Franco-British relations became strained, and in May 1803 the British declared war.

THE EMPIRE

The peace settlement had brought about the life consulate; the return of war was to stimulate the formation of the empire. The British government, which would have been glad to see Bonaparte deposed or removed by assassination, renewed its subsidies to the French royalists, who resumed their agitation and plotting. When a British-financed assassination plot was uncovered in 1804, Bonaparte decided to react vigorously enough to deter his opponents from any more such attempts. The police believed that the real head of the conspiracy was the young Duc d'Enghien, a scion of the royal house of Bourbon, who was residing in Germany, a few miles across the frontier. Accordingly, with the agreement of Talleyrand and the police chief Joseph Fouché, the Duc was kidnapped on neutral soil and brought to Vincennes, where

British-financed assassination plot

he was tried and shot (March 21). This action provoked a resurgence of opposition among the old aristocracy but enhanced the influence of Fouché.

Founding the empire. In the hope of consolidating his own position, Fouché now suggested to Bonaparte that the best way to discourage conspiracy would be to transform the life consulate into a hereditary empire, which, because of the fact that there would be an heir, would remove all hope of changing the regime by assassination. Bonaparte readily accepted the suggestion, and on May 28, 1804, the empire was proclaimed.

Though there was little change in the organization of the government of France, Napoleon as emperor revived a number of institutions similar to those of the *ancien régime*. In the first place, he wanted to be consecrated by the pope himself, so that his coronation should be even more impressive than that of the kings of France. Pius VII agreed to come to Paris, and the ceremony, which seemed equally outrageous to royalists and to the old soldiers of the Revolution, took place in Notre-Dame on December 2, 1804. At the last moment, the Emperor took the crown from the Pope and set it on his own head himself.

The imperial regime also instituted its symbols and titles. Princely titles were brought back for the members of Napoleon's family in 1804, and an imperial nobility was created in 1808. As opposition was still lively, Napoleon intensified his propaganda and imposed an increasingly strict censorship on the press. A dictatorial regime allowed him to carry on his wars for years without worrying about French public opinion. Having been president of the Italian Republic (as the Cisalpine Republic was renamed) since January 1802, Napoleon in March 1805 was proclaimed king of Italy and crowned in Milan in May.

War with Britain. From 1803 to 1805 Napoleon had only the British to fight; and again France could hope for victory only by landing an army in the British Isles, whereas the British could defeat Napoleon only by forming a continental coalition against him. Napoleon began to prepare an invasion again, this time with greater conviction and on a larger scale. He gathered nearly 2,000 ships between Brest and Antwerp and concentrated his Grande Armée in the camp at Boulogne (1803). Even so, the problem was the same as in 1798: to cross the Channel, the French had to have control of the sea.

Still far inferior to the British Navy, the French fleet needed the help of the Spanish; and even then the two fleets together could not hope to defeat more than one of the British squadrons. Spain was induced to declare war on Great Britain in December 1804, and it was decided that French and Spanish squadrons massed in the Antilles should lure a British squadron into these waters and defeat it, thus making the balance roughly equal between the Franco-Spanish navy and the British. A battle in the entrance to the Channel could then be fought with some chance of success.

Franco-Spanish naval strategy

The plan failed. The French squadron from the Mediterranean, under Adm. Pierre de Villeneuve, found itself alone at the appointed meeting place in the Antilles. Pursued by Nelson and not daring to attack him, it turned back toward Europe and took refuge in Cádiz in July 1805; there the British blockaded it. Accused of cowardice by the angry Napoleon, Villeneuve resolved to run the blockade, with the support of a Spanish squadron; but on October 21, 1805, he was attacked by Nelson off Cape Trafalgar. Nelson was killed in the battle, but the Franco-Spanish fleet was totally destroyed. The British had won a decisive victory, which eliminated the danger of invasion and gave them freedom of movement at sea.

They had also succeeded in organizing a new anti-French coalition consisting of Austria, Russia, Sweden, and Naples. On July 24, 1805, three months before Trafalgar, Napoleon had ordered the Grande Armée from Boulogne to the Danube (thus ruling out an invasion of England even if the French had won at Trafalgar). In the week preceding Trafalgar, the Grande Armée won an outstanding victory over the Austrians at Ulm, and on November 13 Napoleon entered Vienna. On December 2, 1805, in his greatest victory he defeated the combined Austrian and Russian armies in the Battle of Austerlitz. By the Treaty of Pressburg, Austria renounced all influence in Italy and ceded Venetia and Dalmatia to Napoleon, as well as extensive territory in Germany to his protégés Bavaria, Württemberg, and Baden. The French then proceeded to dethrone the Bourbons in the kingdom of Naples, which was bestowed on Napoleon's brother Joseph. In July 1806 the Confederation of the Rhine was founded—soon to embrace all western Germany in a union under French protection.

Victories in Europe

In September 1806 Prussia entered the war against France, and on October 14 the Prussians armies were defeated at Jena and at Auerstädt. The Russians put up a better resistance at Eylau in February 1807 but were routed at Friedland in June. In Warsaw Napoleon fell in love with Countess Marie Walewska, a Polish patriot who hoped that Napoleon would resurrect her country. Napoleon had a son by her.

The Russian emperor Alexander I could have continued the struggle, but he was tired of the alliance with the British. He met Napoleon at Tilsit, in northern Prussia near the Russian frontier. There, on a raft anchored in the middle of the Niemen River, they signed treaties that created the Grand Duchy of Warsaw from the Polish provinces detached from Prussia and, in effect, divided control of Europe between the emperors, Napoleon taking the west and Alexander the east. Alexander even made a vague promise of a land attack against the British possessions in India.

Blockade and the peninsular campaign. As Napoleon could no longer think of invading England, he tried to induce capitulation by stifling the British economy. By closing all of Europe to British merchandise, he hoped to bring about a revolt of the British unemployed that could force the government to sue for peace. He forbade all trade with the British Isles, ordered the confiscation of all goods coming from English factories or from the British colonies, and condemned as fair prize not only every British ship but also every ship that had touched the coasts of England or its colonies.

For the blockade to succeed, it had to be enforced rigorously throughout Europe. But from the beginning, England's old ally Portugal showed itself reluctant to comply, for the blockade would mean its commercial ruin. Napoleon decided to break down Portuguese opposition by force. Charles IV of Spain let the French troops cross his kingdom, and they occupied Lisbon; but the prolonged presence of Napoleon's soldiers in the north of Spain led to insurrection. When Charles IV abdicated in favour of his son Ferdinand VII, Napoleon, seeing the opportunity to rid Europe of its last Bourbon rulers, summoned the Spanish royal family to Bayonne in April 1808 and obtained the abdication of both Charles and Ferdinand; they were interned in Talleyrand's château. After the bloody suppression of an uprising in Madrid, insurrection spread across the whole country, for the Spaniards would not accept Joseph Bonaparte, king of Naples, as their new king.

The subsequent defeat of his forces in Spain and Portugal were sensational blows to Napoleon's prestige. Soon the Iberian Peninsula, up in arms, became a bridgehead on the Continent for the British. Under the energetic Arthur Wellesley (later 1st duke of Wellington), in command from 1809, the Anglo-Spanish-Portuguese forces were to achieve decisive successes.

Defeats in Spain and Portugal

At the Congress of Erfurt (September–October 1808), a conference with Alexander I, Napoleon assembled a great concourse of princes to impress the Russian emperor in an attempt to extract promises of help. Whether impressed or not, Alexander would make no definite commitment. Alexander's refusal, furthermore, was partly prompted by Talleyrand, who had become dismayed by Napoleon's policies and was already negotiating with the Russian emperor behind his master's back.

By early 1809, however, with most of the Grande Armée thrown into Spain, Napoleon seemed on the point of overcoming the revolt. Then, in April, Austria launched an attack in Bavaria in the hope of rousing all Germany

against the French. Napoleon once again defeated the Habsburgs (July 6) and by the Treaty of Schönbrunn (October 14, 1809) obtained the Illyrian Provinces, thus rounding out the continental system.

Consolidation of empire. In 1810 Napoleon's fortunes were at their zenith, despite some failures in Spain and Portugal. He considered himself Charlemagne's heir. He repudiated Joséphine, who had not given him a child, so that he could marry Marie-Louise, daughter of the Austrian emperor Francis I. The birth of a son, the king of Rome, in March 1811 seemed to assure the future of his empire—now at its greatest extent, including not only the Illyrian Provinces but also Etruria (Tuscany), some of the Papal States, Holland, and the German states bordering the North Sea. The empire was surrounded by a ring of vassal states ruled over by the Emperor's relatives: the Kingdom of Westphalia (Jérôme Bonaparte); the Kingdom of Spain (Joseph Bonaparte); the Kingdom of Italy (with Eugène de Beauharnais, Joséphine's son, as viceroy); the Kingdom of Naples (Joachim Murat, Napoleon's brother-in-law); and the Principality of Lucca and Piombino (Félix Bacciochi, another brother-in-law). Finally, other territories were closely bound to the empire by treaties: the Swiss Confederation (of which Napoleon was the mediator), the Confederation of the Rhine, and the Grand Duchy of Warsaw. Even Austria seemed bound to France by Napoleon's marriage to Marie-Louise.

European unity and the roots of new nationalisms

The political map of Europe, which had been so complicated before 1796, was now greatly simplified. Yet the frontiers did not coincide either with geographical features or with "nationalities." Whatever he may later have said, Napoleon, while he was in power, was not interested in realizing either German or Italian unity. Yet by reducing the number of states, by pushing the frontiers about, by amalgamating populations, and by propagating institutions like those that the Revolution and nationalism had created in France, he prepared the ground for German and Italian unification. National feeling in Europe, stirred by French ideas and by contact with Frenchmen, in turn gave rise to the first resistance against French domination. From 1809 onward Spanish guerrillas, supported by British troops, were harassing the French, and the national Cortes, convened at Cádiz by the insurrectionaries, in 1812, promulgated a constitution inspired by the ideas of the French Revolution of 1789 and by British institutions.

Disaster in Russia and its aftermath. Since the Congress of Erfurt, the Russian emperor had shown himself less and less inclined to deal with Napoleon as a trusted partner. In the spring of 1812, therefore, Napoleon massed his forces in Poland to intimidate Alexander. After some last attempts at agreement, in late June his Grande Armée—about 453,000 men, including contingents extorted from Prussia and from Austria—began to cross the Niemen River. The Russians retreated, adopting a "scorched earth" policy. Napoleon's army did not reach the approaches to Moscow till the beginning of September. The Russian commander in chief, Mikhail I. Kutuzov, engaged it at Borodino on September 7. The fight was savage, bloody, and indecisive, but a week later Napoleon entered Moscow, which the Russians had abandoned. On that same day, a huge fire broke out, destroying the greater part of the town. Moreover, Alexander unexpectedly refused to treat with Napoleon. Withdrawal was necessary, and the premature onset of winter made it disastrous. After the difficult crossing of the Berezina River in November, fewer than 10,000 men fit for combat remained with Napoleon's main force.

This catastrophe heartened all the peoples of Europe to defy Napoleon. In Germany the news unleashed an outbreak of anti-French demonstrations. The Prussian contingents deserted the Grande Armée in December and turned against the French. The Austrians also withdrew their troops and adopted an increasingly hostile attitude, and in Italy the people began to turn their backs on Napoleon.

Even in France, signs of discontent with the regime were becoming more frequent. In Paris a malcontent general nearly succeeded in carrying out a coup d'etat after announcing, on October 23, 1812, that Napoleon had died in Russia. This incident was a major factor in Napoleon's decision to hasten back to France ahead of the Grande Armée. Arriving in Paris on December 18, he proceeded to stiffen the dictatorship, to raise money by various expedients, and to levy new troops.

Attempted coup d'etat in Paris

Thus, in 1813 the forces arrayed against France were no longer armies of mercenaries but were those of nations fighting for their freedom as the French had fought for theirs in 1792 and 1793; and the French themselves, for all their courage, had lost their former enthusiasm. The Emperor's ideal of conquest was no longer that of the nation.

In May 1813 Napoleon won some successes against the Russians and Prussians at the battles of Lützen and Bautzen, but his decimated army needed reinforcements. The armed mediation of Austria induced Napoleon to agree to an armistice, during which a congress was held at Prague. There, Austria proposed very favourable conditions: the French Empire was to return to its natural limits; the Grand Duchy of Warsaw and the Confederation of the Rhine were to be dissolved; and Prussia was to return to its frontiers of 1805. Napoleon made the mistake of hesitating too long. The congress closed on August 10 before his reply arrived, and Austria declared war.

The French were even worse off than in the spring. The allies were gaining new troops every day, as one German contingent after another left Napoleon to go over to the other side. The greatest debacle since Napoleon came to power was the Battle of Leipzig, or "Battle of the Nations" (October 16–19, 1813), in which the Grande Armée was torn to shreds. That defeat degenerated fast into collapse. The French armies in Spain, forced to retreat, had been defeated in June; and by October the British were attacking their defenses north of the Pyrenees. In Italy the Austrians took the offensive, crossed the Adige River, and occupied Romagna. Murat, now openly a traitor to the Emperor who had made him king of Naples, entered into negotiations with the Viennese court. The Dutch and the Belgians demonstrated against Napoleon.

Downfall and abdication. In January 1814 France was being attacked on all its frontiers. The allies cleverly announced that they were fighting not against the French people but against Napoleon alone, since in November 1813 he had rejected the terms offered by the Austrian foreign minister Metternich, which would have preserved the natural frontiers of France. The extraordinary strategic feats achieved by the Emperor during the first three months of 1814 with the army of young conscripts were not enough; he could neither defeat the allies, with their overwhelming numerical superiority, nor arouse the majority of French people from their resentful torpor. The Legislative Assembly and the Senate, formerly so docile, were now asking for peace and for civil and political liberties.

Four-nation pact against Napoleon

By the Treaty of Chaumont of March 1814, Austria, Russia, Prussia, and Great Britain bound themselves together for 20 years, undertook not to negotiate separately, and promised to continue the struggle till Napoleon was overthrown. When the allied armies arrived before Paris on March 30, Napoleon had moved east to attack their rear guard. The Parisian authorities, no longer overawed by the Emperor, lost no time in treating with the allies. As president of the provisional government, Talleyrand proclaimed the deposition of the Emperor and, without consulting the French people, began to negotiate with Louis XVIII, the brother of the executed Louis XVI. Napoleon had only reached Fontainebleau when he heard that Paris had capitulated. Persuaded by the majority of his marshals that further resistance was useless, he finally abdicated on April 6.

By the Treaty of Fontainebleau, the allies granted him the island of Elba as a sovereign principality with an annual income of 2,000,000 francs to be provided by France and a guard of 400 volunteers; also, he retained the title of emperor. After unsuccessfully trying to poison

himself, Napoleon spoke his farewell to his "Old Guard," and after a hazardous journey, during which he narrowly escaped assassination, he arrived at Elba on May 4.

ELBA AND THE HUNDRED DAYS

"I want from now on to live like a justice of the peace," Napoleon declared on his little island. But a man of such energy and imagination could hardly be expected to resign himself to defeat at the age of 45.

In France, moreover, the Bourbon Restoration was soon exposed to criticism. Though in 1814 the majority of the French people were tired of the Emperor, they had expressed no wish for the return of the Bourbons. They were strongly attached to the essential achievements of the Revolution, and Louis XVIII had come back "in the baggage train of the foreigners" with the last surviving émigrés who had "learnt nothing and forgotten nothing" and whose influence seemed to threaten most of the Revolution's achievements. The apathy of April 1814 quickly gave way to mistrust. Old hatreds were revived, resistance organized, and conspiracies formed.

From Elba, Napoleon kept a close watch on the Continent. He knew that some of the diplomats at Vienna, where a congress was deciding the fate of Europe, considered Elba, between Corsica and Italy, too close to France and to Italy and wanted to banish him to a distant island in the Atlantic. Also, he accused Austria of preventing Marie Louise and his son from coming to join him (in fact, she had taken a lover and had no intention of going to live with her husband). Finally, the French government refused to pay Napoleon's allowance so that he was in danger of being reduced to penury.

Napoleon's return to France

All these considerations drove Napoleon to action. Decisive as ever, he returned to France like a thunderbolt. On March 1, 1815, he landed at Cannes with a detachment of his guard. As he crossed the Alps, the republican peasants rallied round him, and near Grenoble he won over the soldiers dispatched to arrest him. On March 20 he was in Paris.

Napoleon was brought back to power as the embodiment of the spirit of the Revolution rather than as the emperor who had fallen a year before. To rally the mass of Frenchmen to his cause he should have allied himself with the Jacobins; but this he dared not do. Unable to escape from the bourgeoisie whose predominance he himself had assured and who feared above all else a revival of the socialist experiments of 1793 and 1794, he could only set up a political regime scarcely distinguishable from that of Louis XVIII. Enthusiasm ebbed fast, and the Napoleonic adventure seemed a dead end.

To oppose the allied troops massing on the frontiers, Napoleon mustered an army with which he marched into Belgium and defeated the Prussians at Ligny on June 16, 1815. Two days later, at Waterloo, he met the British under Wellington, the victor of the Peninsular War. A savage battle followed. Napoleon was in sight of victory when the Prussians under Gebhard Blücher arrived to reinforce the British, and soon, despite the heroism of the Old Guard, Napoleon was overthrown.

Back in Paris, Parliament forced Napoleon to abdicate; he did so, in favour of his son, on June 22, 1815. On July 3 he was at Rochefort, intending to take ship for the United States, but a British squadron prevented any French vessel from leaving the port. Napoleon then decided to appeal to the British government for protection. His request granted, he boarded the "Bellerophon" on July 15. The allies were agreed on one point: Napoleon was not to go back to Elba. Nor did they like the idea of his going off to America. It would have suited them if he had fallen a victim to the "White Terror" of the returned counter-revolutionaries or if Louis XVIII had had him summarily tried and executed. Great Britain had no choice but to send him to detention in a far-off island. The British government announced that the island of Saint Helena in the southern Atlantic had been chosen for his residence; because of its remote position Napoleon would enjoy much greater freedom than would be possible elsewhere. Napoleon protested eloquently: "I appeal to history!"

EXILE ON SAINT HELENA

On October 15, 1815, Napoleon disembarked in Saint Helena with those followers who were voluntarily accompanying him into exile: Gen. Henri-Gratien Bertrand, grand marshal of the palace, and his wife; the comte Charles de Montholon, aide-de-camp, and his wife; Gen. Gaspard Gourgaud; Emmanuel Las Cases, the former chamberlain; and several servants. After a short stay at the house of a wealthy English merchant, they moved to Longwood, originally built for the lieutenant governor.

Daily life at St. Helena

Napoleon settled down to a life of routine. He got up late, breakfasting about 10 AM, but seldom went out. He was free to go anywhere on the island so long as he was accompanied by an English officer, but he soon refused to comply with this condition and so shut himself up in the grounds of Longwood. He wrote and talked much. At first Las Cases acted as his secretary, compiling what was later to be the *Mémorial de Sainte-Hélène* (first published in 1823). From 7 to 8 PM Napoleon had dinner, after which a part of the evening was spent in reading aloud—Napoleon liked to hear the classics. Then they played cards. About midnight Napoleon went to bed. Some of his time was devoted to learning English, and he eventually began reading English newspapers; but he also had a large number of French books sent from Europe, which he read attentively and annotated.

Saint Helena has a healthful climate, and Napoleon's food was good, carefully prepared, and plentiful. His inactivity undoubtedly contributed to the deterioration of his health. The man who for 20 years had played so great a role in the world and who had marched north, south, east, and west across Europe could hardly be expected to endure the monotony of existence on a little island, aggravated by a self-imposed life of a recluse. He had also more intimate reasons for unhappiness: Marie-Louise sent no word to him, and he may have learned of her liaison with the Austrian officer appointed to watch over her, Graf Adam von Neipperg (whom she eventually married in secret without waiting for Napoleon's death); nor did he have any news of his son, the former king of Rome, who was now living in Vienna with the title of duke of Reichstadt. Finally, though the severity of Sir Hudson Lowe has been much exaggerated, it is certain that this "jailer," who arrived as governor of Saint Helena in April 1816, did nothing to make Napoleon's life easier. Napoleon from the start disliked him as the former commander of the Corsican rangers, a band of volunteers that had been recruited by the British and that was largely composed of enemies of the Bonaparte family. Obsessed by his responsibilities and always anxious to carry out his instructions exactly, Lowe came into conflict with Las Cases. He saw Las Cases as Napoleon's confidant and had him arrested and expelled. Thenceforward, relations between the governor and Napoleon were limited strictly to those stipulated by the regulations.

First signs of illness

Napoleon showed the first signs of illness at the end of 1817; he seems to have had an ulcer or a cancer of the stomach. The Irish doctor Barry O'Meara, having asked in vain for a change in the conditions under which Napoleon lived, was dismissed; so also was his successor John Stokoe, who was likewise thought to be well-disposed toward Napoleon. The undistinguished Corsican doctor who took their place, Francesco Antommarchi, prescribed a treatment that could do nothing to cure his patient. It is uncertain, however, whether Napoleon's disease was curable at all, even by 20th-century methods.

From the beginning of 1821, the illness became rapidly worse. From March, Napoleon was confined to bed. In April he dictated his last will:

> I wish my ashes to rest on the banks of the Seine, in the midst of that French people which I have loved so much . . . . I die before my time, killed by the English oligarchy and its hired assassins.

On May 5 he spoke a few coherent phrases: "My God . . . The French nation . . . my son . . . head of the army. . . ." He died at 5:49 PM on that day, not yet 52 years old. His body was dressed in his favourite uniform, that of the

Chasseurs de la Garde, covered by the gray overcoat that he had worn at Marengo. The funeral was conducted simply, but with due propriety, in the Rupert Valley, where Napoleon had sometimes walked, beside a stream in which two willows were reflected. The stone covering his tomb bore no name, only the words "*Ci-Gît*" ("Here Lies").

THE NAPOLEONIC LEGEND

Napoleon's fall had set loose a torrent of hostile books designed to sully his reputation. One of the least violent of these was the pamphlet *De Buonaparte, des Bourbons, et de la nécessité de se rallier à nos princes légitimes* (1814; *On Buonaparte and the Bourbons, and the Necessity of Rallying Around Our Legitimate Princes, for the Safety of France and of Europe*) by François de Chateaubriand, a well-known writer of royalist sympathies. But this anti-Napoleonic literature soon died down, while the task of defending Napoleon was taken up. Lord Byron had published his "Ode to Napoleon Buonaparte" as early as 1814; the German poet Heinrich Heine wrote his ballad "Die Grenadiere"; and in 1817 the French novelist Stendhal began his biography *Vie de Napoleon*. At the same time, the Emperor's most faithful supporters were working toward his rehabilitation, talking about him, and distributing reminders of him, including engravings. They idealized his life ("What a novel my life is!" he himself had said) and began to create the Napoleonic legend.

Early memoirs and chronicles

As soon as the Emperor was dead, the legend grew rapidly. Memoirs, notes, and narratives by those who had followed him into exile contributed substantially to it. In 1822 Dr. O'Meara, in London, had his *Napoleon in Exile, or a Voice from Saint Helena* published; in 1823 the publication of the *Mémoires pour servir à l'histoire de France sous Napoléon, écrits à Sainte-Hélène sous sa dictée (Memoirs of the History of France during the Reign of Napoleon, Dictated by the Emperor at St. Helena)* by Montholon and Gourgaud, began; Las Cases, in his famous *Mémorial*, presented the Emperor as a republican opposed to war who had fought only when Europe forced him to fight in defense of freedom; and in 1825 Antommarchi published his *Derniers moments de Napoléon*. Thereafter the number of books in Napoleon's honour increased continually; among them were Victor Hugo's "Ode à la Colonne," the 28 volumes of the *Victoires et conquêtes des Français*, and Sir Walter Scott's *Life of Napoleon Buonaparte, Emperor of the French*. Neither police action nor prosecutions could prevent books, pictures, and objects evoking the imperial saga from multiplying in France.

After the July Revolution of 1830, which created the bourgeois monarchy under Louis-Philippe, thousands of tricolour flags appeared in windows, and the government had not only to tolerate the growth of the legend but even to promote it. In 1833 the statue of Napoleon was put back on the top of the column in the Place Vendôme in Paris; and in 1840 the King's son François, prince de Joinville, was sent in a warship to fetch the Emperor's remains from Saint Helena to the banks of the Seine in accordance with his last wishes. A magnificent funeral was held in Paris in December 1840, and Napoleon's body was conveyed through the Arc de Triomphe in the Place de l'Étoile to entombment under the dome of the Invalides.

Napoleon's nephew Louis-Napoleon exploited the legend in order to seize power in France. Though his attempts at Strasbourg in 1836 and at Boulogne in 1840 were failures, it was chiefly because of the growth of the legend that he won election to the presidency of the Second Republic with an overwhelming majority in 1848 and was able to carry out the coup d'etat of December 1851 and make himself emperor in 1852.

The disastrous end of the Second Empire in 1870 damaged the Napoleonic legend and gave rise to a new anti-Napoleonic literature, best represented by Hippolyte Taine's *Origines de la France contemporaine* (1876–94). World Wars I and II, however, together with the experience of the 20th-century dictatorships, made it possible to judge Napoleon more fairly. Any comparison with Stalin or Hitler, for instance, can only be to Napoleon's advantage. He was tolerant, he released the Jews from the ghettoes, and he showed respect for human life. Brought up on the nationalist *Encyclopédie* and on the writings of the Philosophes of the Enlightenment, he remained above all a man of the 18th century, the last of the "enlightened despots." One of the gravest accusations made against Napoleon is that he was the "Corsican ogre" who sacrificed millions of men to his ambition. Precise calculations show that the Napoleonic Wars of 1800–15 cost France itself about 500,000 men; *i.e.*, about one-sixtieth of the population. The loss of these young men, furthermore, seems to have had a notably adverse effect on the birth rate.

Assessment and legacy

The social structure of France changed little under the First Empire. It remained roughly what the Revolution had made it: a great mass of peasants comprising three quarters of the population—about half of them working owners of their farms or sharecroppers and the other half with too little land for their own subsistence and hiring themselves out as labourers. Industry, stimulated by the war and the blockade of English goods, made remarkable progress in northern and eastern France, whence exports could be sent to central Europe; but it declined in the south and west because of the closing of the Mediterranean and the Atlantic. The great migrations from rural areas toward industry in the towns began only after 1815. The nobility would probably have declined more swiftly if Napoleon had not restored it; but it could never recover its former privileges.

Above all, Napoleon left durable institutions, the "granite masses" on which modern France has been built up: the administrative system of the prefects, the Code Napoléon, the judicial system, the Banque de France and the country's financial organization, the universities, and the military academies. Napoleon changed the history of France and of the world.

BIBLIOGRAPHY. Since the beginning of the 20th century, more than 200,000 volumes have been published on Napoleon and his times. It is, therefore, impossible to compile a complete bibliography; cited here are only the most important works.

Sources: An enormous mass of documents is dispersed throughout all the archives of Europe, but the essential source is the *Correspondance de Napoléon I*, 32 vol. (1858–69) published on command of Napoleon III, with the *Oeuvres de Napoléon à Sainte-Hélène*, 4 vol. (1870). In 1867, the *Oeuvres littéraires et écrits militaires* of Napoleon were published in 3 volumes. In English, one can read *The Bonaparte Letters and Dispatches . . . from the Originals in His Private Cabinet*, 2 vol. (1846); *The Confidential Correspondence of Napoleon Bonaparte with His Brother Joseph . . .*, 2 vol. (1855); and the *Unpublished Correspondence of Napoleon I, Preserved in the War Archives*, 3 vol. (1913). The principal ideas of the Emperor may be found in the following works: *The Corsican* (1910); J.M. THOMPSON, *Napoleon Self-Revealed* (1934); HEROLD J. CHRISTOPHER, *The Mind of Napoleon* (1955); ANDRE PALLUEL (ed.), *Dictionnaire de l'Empereur* (1969); and ADRIEN DANSETTE, *Napoléon, pensées politiques et sociales* (1969). For Napoleon's itinerary, see ALBERT SCHUERMANS, *Itinéraire général de Napoléon Ier* (1906); and LOUIS GARRES, *Itinéraire de Napoléon Bonaparte* (1947).

General biographies: PIERRE LANFREY, *Histoire de Napoléon Ier*, 5 vol. (1867–75; Eng. trans., *The History of Napoleon the First*, 5 vol., 1871–79), a work written at the end of the Second Empire—very hostile to Napoleon; AUGUST FOURNIER, *Napoleon I*, 3 vol. (1886; Eng. trans., 2 vol., 1911), an impartial study written by an Austrian historian; FREDERIC MASSON, *Napoléon et sa famille*, 13 vol. (1897–1919), furnishes numerous details on the everyday life of Napoleon and his relationships with his relatives; J. HOLLAND ROSE, *The Life of Napoleon I*, 2 vol. (1902), very apologetic; J.M. THOMPSON, *Napoleon Bonaparte: His Rise and Fall* (1952), excellent rectification; E.V. TARLE, *Bonaparte* (Eng. trans. 1937, reprinted 1961), the point of view of a Soviet historian; FELIX MARKHAM, *Napoleon* (1963), a very objective study; JEAN MISTLER, *Napoléon et l'Empire* (1968), a magnificently illustrated work, each chapter written by a specialist; ANDRE CASTELOT, *Napoléon* (1968), a vulgarized book that realized great success in France; JACQUES GODECHOT, *Napoléon* (1969), an essay and an empirical study (in French).

Specialized biographies: On Napoleon's youth, the basic work remains that of ARTHUR CHUQUET, *La Jeunesse de Napoléon*, 3 vol. (1897–99). On the more intimate life of Napoleon, ARTHUR LEVY, *Napoléon intime*, 7th ed. (1932); and FREDERIC MASSON, *Napoléon chez lui: la journée de l'Empereur aux Tuileries* (1894; Eng. trans., *Napoleon at Home*, 2 vol., 1894). On Napoleon's health, see JAMES KEMBLE, *Napoleon Immortal* (1959).

The Napoleonic legend and studies on Napoleon's captivity: MABEL BROOKES, *Saint Helena Story* (1960); JEAN LUCAS-DUBRETON, *Le Culte de Napoléon, 1815–1848* (1960), the only definitive study of the Napoleonic legend.

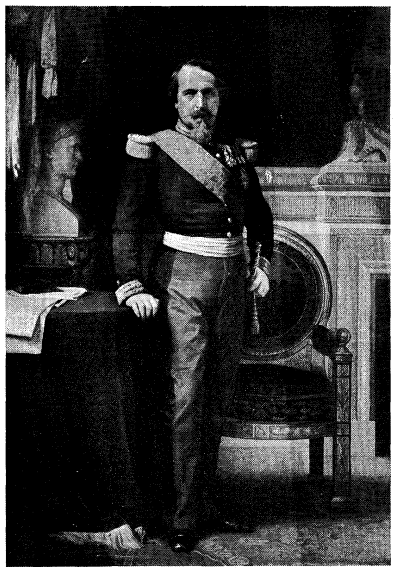
Films: Many films have been produced on the life of Napoleon. The best are by ABEL GANCE, *Napoléon* (1925) and *Austerlitz* (1960). The film *War and Peace* (1967), produced by Soviet director SERGEI BONDARTCHUK, gives a valid reconstruction of the imperial era. It was adapted from the novel by Tolstoy.

(J. Go.)

Napoleon III

Charles-Louis-Napoléon Bonaparte, the nephew of Napoleon I, was emperor of the French from 1852 to 1870. He successfully imposed two decades of stable, authoritarian government on France and encouraged industrial expansion. He also played a prominent role in European affairs, but finally led his country to defeat in the Franco-German War of 1870–71.

H. Roger-Viollet



Napoleon III, portrait by H. Flandrin (1809–64). In the Musée de Versailles.

He was born on April 20, 1808, in Paris, the third son of Napoleon I's brother Louis Bonaparte, who was king of Holland from 1806 to 1810, and his wife Hortense de Beauharnais Bonaparte, stepdaughter of Napoleon I. Other children born of this marriage were Napoléon-Charles, who died in infancy, and Napoléon-Louis.

Youth in exile. Louis-Napoléon's childhood and youth were spent largely in exile. His mother, like all the Bonapartes, was banished from France in 1815 after the fall of Napoleon I. Eventually, she found a new home in Switzerland, where, in 1817, she bought the castle of Arenenberg. Of romantic disposition herself, she inspired young Louis-Napoléon with a longing for his lost fatherland, as well as with enthusiastic admiration of the genius of Napoleon I. After attending a grammar school at Augsburg, Germany (1821–23), her "sweet stubborn boy" was taught by private tutors. During visits to relatives in southern Germany and Italy, he became acquainted not only with other exiled victims of the restoration of the Bourbon monarchy but also with the life of a suppressed people, such as those Italians who were living under Austrian and papal rule. He was, above all, interested in history and inspired by the idea of national liberty. Accordingly, he took part in an unsuccessful plot against the papal government in Rome in 1830 and in the rebellion in central Italy in 1831, in which his beloved brother perished. He himself was saved from the Austrian troops only by his mother's bold intervention.

Claim to the throne. After the death in 1832 of his cousin the Duke of Reichstadt (Napoleon I's only son), Louis-Napoléon considered himself his family's claimant to the French throne. To be better prepared for his task, he completed his military training and pursued his studies of economic and social problems. Soon after, he felt ready to publish his own writings on political and military subjects. In his pamphlet "Rêveries politiques" (1832), he asserted that only an emperor could give France both glory and liberty. He thus wanted to make his name known, propagate his ideas, and recruit adherents. Convinced that as Napoleon's nephew he would be popular with the French army, he vainly tried, on October 30, 1836, to win over the Strasbourg garrison for a coup d'etat. King Louis-Philippe exiled him to the United States, from which he was recalled early in 1837 by his mother's last illness. Expelled from Switzerland in 1838 as a result of French pressure, he settled in England, where he moved in fashionable circles.

Attempted coup d'état and exile

In 1839 he published "Des idées napoléoniennes." So far, Bonapartism had been nothing but a wistful reminiscing of former beneficiaries of the empire or a romantic legend created by those who were dissatisfied with the humdrum present. In his new booklet Louis-Napoléon tried to transform Bonapartism into a political ideology. In doing so, he obeyed mystical inspirations as well as rationalism. To him, ideology and politics were the result of rational reflection as well as of belief. The central exponent in history was, in his opinion, the great personality called by Providence and representing progress. Napoleon I had been such a man, even though he was not allowed to finish his work. But Napoleon, the "Messiah of the new ideas," was survived by the "Napoleonic idea," for the "political creed," like the religious creeds, had its martyrs and apostles. The Napoleonic idea was a "social and industrial one, humanitarian and encouraging trade," that would "reconcile order and freedom, the rights of the people and the principles of authority." Louis-Napoléon saw it as his task to accomplish this mission.

Landing with 56 followers, near Boulogne, on August 6, 1840, he was again unsuccessful. The town's garrison did not join him. He was arrested, brought to trial, and sentenced to "permanent confinement in a fortress." At his "university of Ham" (the castle in which he was held) he spent his time studying to fit himself for his imperial role. He corresponded with members of the French opposition and published articles in some of their newspapers. He also wrote several brochures, among them "Extinction du paupérisme" (1844), which won him some supporters on the left. It was not until May 25, 1846, that he succeeded in escaping and fleeing to Great Britain, where he waited for another chance to seize power.

Presidency. On hearing of the outbreak of the revolution, in February 1848, he travelled to Paris but was sent back by the provisional government. Some of his supporters, however, organized a small Bonapartist party and nominated him as their candidate for the Constituent Assembly. On June 4 he was elected in four *départements* but, awaiting more settled conditions, he refused to take his seat. Running again in September, he was elected in five *départements*, and after his arrival in Paris he lost no time in preparing to run for the presidency. He was now supported by the newly founded Party of Order, which consisted of adherents of the Bourbons, Louis-Philippe, and Catholics. Lacking a suitable candidate, they regarded Louis-Napoléon—not a skilled parliamentarian but a popular figure—as a useful and harmless tool.

He used, now on a large scale, the kind of propaganda that had won him elections before. Because of his name and his descent, the Emperor's nephew captivated the voters. Evoking the Napoleonic legend with its memories of national glory, Louis-Napoléon promised to bring back those days in time of peace. He succeeded also in recommending himself to every group of the population by promising to safeguard their particular interests. He

promised "order" and "prosperity" to the middle class and the farmers and assistance to the poor. In the elections in December 1848, he was the only candidate to obtain votes—totalling 5,434,226—from among all classes of the population.

He assumed office, determined to free himself from any dependence on the Party of Order, which had also won the parliamentary elections of May 1849. The new government sent a military expedition to help the pope reconquer Rome. At home it deprived active Republicans of their government positions and restricted their liberties, but the President could rely on only about a dozen members of the National Assembly who were Bonapartists. Prudently expanding his power by using every right the constitution granted him, Louis-Napoléon soon obtained key positions in the administration and in the army for his adherents. On October 31, he succeeded for the first time in appointing a Cabinet consisting of men depending more on him than on the National Assembly. By travelling through the country he gained wide popularity. Moreover, he used the disfranchisement of 3,000,000 electors of the poorer classes by the National Assembly in 1850 and an economic recession in 1851 as a pretext for agitating against the parties and for advertising himself as the "strong man" against the danger of a nonexistent revolution.

Marshalling his power

The constitution forbade the re-election of the president after expiration of his four-year term, and when Louis-Napoléon realized that he could not obtain the three-fourths majority necessary for a revision of the constitution he carried out a coup d'etat on December 2. Only the Republicans dared to resist him. On December 4 they were defeated in street fighting in Paris, just as they were in other towns and in some regions. Arrests and deportations numbered in the thousands. Louis-Napoleon dissolved the Legislative Assembly and decreed a new constitution, which among other provisions restored universal suffrage. A plebiscite approved the new constitution. Encouraged by his success, he held another plebiscite in November 1852 and was confirmed as emperor after the resolution of the Senate concerning the restitution of the empire. Failing to obtain the hand of a princess of equal birth, Napoleon III married the countess Eugénie de Montijo in January 1853.

Domestic policy as emperor. Napoleon III intended to be always ahead of public opinion so as to be able to understand the requirements of his time and to create laws and institutions accordingly. Hence, he took the greatest pains to study the public opinion and to influence it by means of propaganda. He tried to make his people believe in his mission: "I am not afraid of any attempts on my life. There are men who are instruments of the decisions of Providence. As long as I have not fulfilled my mission, there is no danger for me," he declared publicly on April 29, 1855, after an attempted assassination had failed. But he also made an appeal to his subjects' intelligence by emphasizing the usefulness of measures intended for the "moral and material improvement of man through education and work"—and calling himself "the one elected by the nation, the first citizen of the country and the representative of the interests of all." Although promising "reasonable freedom," for the time being he considered it necessary to use the methods of a police state.

Willing "to take the initiative to do everything useful for the prosperity and the greatness of France," he promoted public works, the construction of railroads, the establishment of institutions of credit, and other means of furthering industry and agriculture. An enthusiastic promoter of great technical projects, he supported inventors and took a personal interest in the rebuilding of modern Paris.

He did not, however, disavow what he called his "love of the diligent and needy." He ensured a lower price for bread, furthered the construction of hygienic dwellings for workers, and established boards of arbitration. In his societies of mutual assistance, employers and employees were to learn to understand each other. He hoped that his social-welfare institutions, to the endowment of which he frequently contributed, would be imitated by the citizens. The middle class, however, looked upon him only as its protector against Socialism and regarded his social ideas as mere utopianism.

Foreign policy as emperor. As in domestic policy, the Emperor immediately took the initiative in foreign affairs. "Louis-Philippe fell because he let France fall into disrepute. I must do something," he declared. He wanted to make France a great power once more by breaking up the European system created by the Congress of Vienna of 1815, which, incidentally, had imposed great humiliations on France. Convinced that in the present "epoch of civilisation the successes of armies were only temporary" and that it was "public opinion which always gained the final victory," he planned "to march at the head of generous and noble ideas," among which the principle of nationality was the most important. In accordance with this principle he wanted an international congress to reconstruct "the European balance of power on more durable and just foundations." And "if other countries gain anything France must gain something also." While advocating international solutions to European problems, he also followed traditions of French foreign policy because they corresponded "to the permanent interests" of France. "He has certain ideas and desires floating in his mind which turn up as circumstances seem favourable," was the comment of the English ambassador about Napoleon III's political activities.

Appeal to nationalism

The Crimean War offered him a chance of realizing one of his favourite ideas: the conclusion of an alliance with Great Britain that would succeed in checking Russian expansion toward the Mediterranean. After the Paris conference, at which the peace terms were settled, Napoleon seemed to become Europe's arbitrator. Ironically, it was an attempt on his life by Felice Orsini, an Italian revolutionary (January 1858), that reminded him of his wish "to do something for Italy." Together with Piedmont-Sardinia, he went to war against Austria in order to expel it from Italy. A promoter of technical warfare, he witnessed the success of his modernized artillery and of the military use of the captive balloon. The fact that at the victorious Battle of Solferino in June 1859 he had been in command convinced him of his military genius. Yet, frightened by the possibility of intervention by the German Confederation, he suddenly made peace. Outmanoeuvred by Count Cavour, who confronted him with a unified Italy instead of the weak federation he had intended, he received Nice and Savoy as a reward. His activities in Italy displeased the British. Despite the conclusion of an Anglo-French commercial treaty in 1860, they remained suspicious, and apprehensively watched his construction of armoured warships and his colonial and oriental policies.

Napoleon III dreamed of "opening new ways to commerce and new outlets to European products overseas," of accelerating "the progress of Christianity and civilization." He was therefore open to a colonial policy bent on furthering commercial interests and the establishment of bases. He intensified the extension of French power in Indochina and West Africa. In the Middle East the Emperor hoped that a better treatment of the Algerians would have a favourable influence on the Arabs from Tunisia to the Euphrates. He supported the construction of the Suez Canal. When the Roman Catholic Maronites who were under French protection in Lebanon were persecuted in 1860, he hoped to profit politically by dispatching an expeditionary force.

Colonial policy

Attempts at reform. In 1860 Napoleon III believed his regime to be sufficiently stable to grant some reasonable freedoms. The commercial treaty with Great Britain was to be the beginning of a new economic policy based on free-trade principles, with the aim of increasing prosperity and decreasing the cost of living. Dissatisfied with the functioning of the legislature, the Emperor decided to give "the great bodies of the state a more direct part in the formation of the general policy of our government."

His hopes were not fulfilled to the extent he had expected. A deterioration in the economy caused dissatisfaction among the middle class and the working people, who joined the Catholics, angered by his antipapal Italian

policy, to become a steadily growing opposition. In the elections of 1857 only five members of the opposition had gained seats in the National Assembly; six years later there were 32, and they demanded more freedom.

At this very time, repeated bladder-stone attacks temporarily incapacitated the Emperor, who had been in poor health since 1856. He had always insisted on exercising control over all decisions of government; in his ministers he had seen nothing but tools. Now, because of his illness he became dependent on persons in his entourage who formed groups and intrigued against each other. In 1863 the authoritarian Eugène Rouher, nicknamed the "Vice Emperor," became prime minister; on the other hand, Napoleon III took the advice of his cousin the Duke of Morny to continue his policy of liberalization. With the help of another cousin, Prince Napoléon, who preached a democratic Bonapartism, he tried to win over the workers. But his concessions (freedom of coalition in 1864, freedom of assembly in 1868, extension of the rights of members of parliament, liberalization of the press laws) were restricted by too many reservations and came too late. He allowed Victor Duruy, his minister of education since 1863, to fight clerical influence in education, yet on the other hand he tried to reconciliate French Catholics by working for a compromise settlement in disputes between the papacy and the new Italian Kingdom.

Foreign adventures

In 1861, by proposing to make the Austrian archduke Maximilian emperor of Mexico and by negotiating with the President of Ecuador about a projected "Kingdom of the Andes," he had begun his Latin American venture. He had expected material rewards for France and also hoped that the planned kingdom would check the growing influence of the United States in Latin America. Yet as soon as the United States had concluded its Civil War it forced the French to withdraw. In Europe when the Polish insurrection broke out in 1863, he did not, in spite of his sympathy, dare to support Poland against Russia. Nevertheless, such sympathies led to an estrangement from Russia. Moreover, the great powers refused Napoleon's proposal of a conference for the reorganization of Europe. Regarding Bismarck's policy with a certain benevolence, he kept aloof when Prussia settled the Schleswig-Holstein question by a war on Denmark. He had always sympathized with Prussia and the idea of the extension of Prussia's power in North Germany. He did not, however, openly tell Bismarck what price he demanded for his help. When in 1866, after routing Denmark, Prussia turned on its former ally, Austria, and defeated it more quickly than Napoleon had expected, he refused any armed intervention in its favour and only acted as mediator. Bismarck, however, was not willing to pay for Napoleon's neutrality by the cession of German territories. The Emperor's plans to obtain compensation in Belgium failed, as did his attempt to gain Luxembourg in 1867. The retreat of Prussian troops from the fortress Luxemburg (in Luxembourg) was no satisfactory reward: "Bismarck tried to cheat me. A French emperor cannot be cheated," grumbled Napoleon III, not willing to allow the Prussians to cross the Main River line and extend their power into southern Germany, which was Bismarck's obvious aim.

Last years. The Emperor's failures in foreign affairs strengthened the opposition. When in the 1869 elections the government received 4,438,000 votes against the opposition's 3,355,000, Napoleon came to recognize that a genuine change of the regime was now inevitable. In January 1870 he appointed Olivier-Émile Ollivier, whom Morny had recommended as the most appropriate prime minister for a liberal empire. The new cabinet informed Great Britain and Prussia that France was ready to disarm, but Bismarck refused to cooperate.

Drift toward war

On July 2 it became known that a Hohenzollern prince, a relative of the King of Prussia, was a candidate for the Spanish throne. In Paris this was regarded as Prussian interference in a French sphere of interest and a threat to its security. Using his favourite means of secret diplomacy, Napoleon played a major part in causing the Hohenzollern prince to renounce his candidature. But then the sick Emperor, influenced by the advocates of a belligerent policy, set out to humiliate Prussia by demanding that the candidature of the Hohenzollern prince would never be revived. "Never was war so popular, and the head of the state could not have stopped the nation's war spirit without risking his crown," he excused himself later. As a result, war broke out on July 19. At the Battle of Sedan the sick Emperor tried in vain to meet his death in the midst of his troops, but on September 2 he surrendered. He was at once deposed, and on September 4 the Third Republic was proclaimed.

Defeat and death

After the conclusion of peace, Napoleon was released by the Germans, and he went to live in England. He studied technical and social problems, defended his politics in various publications, and even thought of landing in France to regain his throne. After undergoing an operation for the removal of bladder stones, he died on January 9, 1873.

His life and work have been variously interpreted. He has been called a revolutionary as well as a reactionary, a precursor of the League of Nations as well as Fascism. He himself believed that he had fulfilled a providential mission. After he was dethroned he confessed: "I may sometimes have been mistaken in my choice of means and men," but he never doubted that his idea of giving France prosperity at home and glory abroad was "an historical necessity."

BIBLIOGRAPHY. *Oeuvres de Napoléon III,* 5 vol. (1854–69). NAPOLEON III, *Histoire de Jules César,* 2 vol. (1865–66). *Biographies:* A.L. GUERARD, *Napoleon III* (1943; Eng. trans., 1955), emphasizes economic and social aspects; F.A. SIMPSON, *The Rise of Louis Napoleon,* 3rd ed. (1950, reprinted 1968); *Louis Napoleon and the recovery of France, 1848–1856,* 3rd ed. (1951), a useful survey; J.M. THOMPSON, *Louis Napoleon and the Second Empire* (1954), an able account; H. BULER, *Napoleon III in seiner Zeit,* vol. 1, *Der Aufstieg* (1961), first volume of a projected three-part biography. See also C.W. HALLBERG, *Franz Joseph and Napoleon III, 1852–1864: A Study of Austro-French Relations* (1955), a fairly good survey; D.I. KULSTEIN, *Napoleon III and the Working Class: A Study of Government Propaganda Under the Second Empire* (1969), an informative study; H.C. PAYNE, *The Police State of Louis Napoleon Bonaparte, 1851–1860* (1966), an impressively documented work; H. PINKNEY, *Napoleon III and the Rebuilding of Paris* (1958), a useful general account; E.A. POTTINGER, *Napoleon III and the German Crisis 1865–1866* (1966), a competent and brilliant study; T. ZELDIN, *The Political System of Napoleon III* (1958), a very interesting analysis; and R.L. WILLIAMS, *The Mortal Napoleon III* (1971), a useful physical test, with regard to the emperor's personality and policy.

(H.G.E.)

Narcotic

Narcotics, which constitute one of the oldest classifications of drugs known to man, produce several effects, primarily analgesia (the relief of pain), but also narcosis (a state of stupor or sleep), addiction (physical dependence), and side effects. The term narcotic is often restricted to the addictive pain-killers.

The narcotics, or narcotic analgesics, as they are more properly called, are the most effective drugs available for the relief of pain during the conscious state. Some types of pain are more amenable to the effects of narcotic analgesics than others. Characteristically, colicky pain originating in hollow viscera, as, for example, the gastrointestinal tract, or ureter, is more affected by narcotic than by non-narcotic analgesics (such as aspirin, acetaminophen, and similar compounds). Narcotic analgesics are useful following major wounds acquired in battles or in civilian accidents. It must be borne in mind, however, that the psychological context in which a wound occurs is of major importance to the perception of pain. For example, four-fifths of a group of male, civilian patients requested pain medication following major surgery, while only one-third of a group of wounded soldiers did so, although the latter had in most instances suffered far greater tissue damage. This difference in response might be explained by the fact that the soldier feels relief at having escaped from the battlefield alive, while the civilian is still concerned with the threat to his well-being posed by major

surgery. Thus there is, generally speaking, no direct relationship found between wound severity and the degree of pain perceived.

General characteristics of narcotics. Narcotic analgesics may be classified as natural, semisynthetic, or synthetic.

Natural narcotics. Opium is the source of natural narcotics. The pharmacologically active constituents of opium are members of a class of chemical compounds known as alkaloids (for a discussion of their biochemistry, see ALKALOIDS); they constitute about 25 percent by weight of opium. Although numerous alkaloids are present in opium, only a few of them—morphine, codeine, papaverine, and noscapine—have wide clinical use. Morphine, the most important alkaloid of opium, is predominantly responsible for its pharmacological properties and remains the standard against which new analgesic narcotics are measured.

Opium

Opium is obtained as a milky juice of unripe seed capsules of a poppy plant (*Papaver somniferum*) native to Asia Minor. The juice, dried in air to form a brownish, gummy mass, is further dried and powdered to provide the final product. Since opium is a source of natural alkaloids, narcotics with actions similar to those of morphine are sometimes called opiates.

Semisynthetic and synthetic narcotics. A number of semisynthetic narcotic analgesics (*e.g.*, heroin), obtained by relatively simple structural modifications of the morphine or codeine molecules, find use; these compounds are listed in Table 1. Fully synthetic compounds, sometimes called opioids, with structural resemblance to the whole, or to a part, of the morphine molecule also are employed as narcotic analgesics. The chemical and generic names of these compounds appear in Table 2.

Table 1: Semisynthetic Narcotic Analgesics

generic name	relative analgesic effect*	relative effect in addicts*	duration of action of average dose (hours)
Morphine†	100	100	3–4
Codeine	8–15	20	3–4
Ethylmorphine	10–20	—	3–4
Heterocodeine	100–200	100–200	2–3
Heroin	200–330	280	2–4
Dihydromorphine	120	333	2–4
Dihydrocodeine	30	30	2–3
Desomorphine	500–1,000	500	2–3
Hydromorphone	200–500	700	2–4
Metopon	200–300	700	3–5
Hydrocodone	70	100	2–3
Oxymorphone	500–1,000	800–1,000	4–6
Oxycodone	100	76	2–3

*The analgesic effect of morphine sulfate and its ability to suppress abstinence syndrome in addicts is rated 100.
†Morphine included for the sake of comparison.

Table 2: Synthetic Narcotic Analgesics

generic name	relative analgesic effect*	relative effect in addicts*	duration of action of average dose (hours)
Morphinan derivatives			
Morphinan	negligible	negligible	—
Levorphan (Levorphanol)	330–500	200–500	4–5
Pentazocine	500	200–300	—
Methadone derivatives			
Methadone	100–130	420	3–4
Isomethadone	40–80	140	3–4
Dipipanone	40	100	2–3
Phenadoxone	15	80	2–4
Dextromoramide	200	300–400	3
Meperidine derivatives			
Meperidine (Pethidine)	10	40	1–2
Alphaprodine	25	60	1–2
Ketobemidone	70–100	100	2–4
Anileridine	25	35	2–3
Phenoperidine	700–1,500	—	1–2
Fentanyl	5,000–10,000	—	1–5

*The analgesic effect of morphine sulfate and its ability to suppress abstinence syndrome in addicts is rated 100.

Relationship between chemical structure and narcotic action. The mechanism and the site of action of narcotic analgesics have not been clearly delineated. In all potent narcotics, it is possible to recognize three molecular characteristics that are believed to be related to the mode of action of the drugs. These consist of (1) a tertiary basic nitrogen; (2) a quaternary carbon separated from the basic nitrogen by an ethylene ($-CH_2\ CH_2-$) chain; and (3) a phenolic hydroxyl or a ketone. A phenyl group, or a group with a similar shape, attached to the quaternary carbon atom is also thought to be an important constituent of the narcotic analgesic molecule.

The primary sites of action of morphine in the body probably consist of the sensory cortex of the frontal lobes and the diencephalon (the posterior portion of the forebrain). Morphine has been shown to depress the pain responses travelling from the brain centres that initially perceive them to those responsible for the subjective reaction of pain. It may be said, therefore, that the analgesic effects of narcotics resemble the effects of transecting the thalamocortical tracts, a surgical procedure known as prefrontal lobotomy. While the precise mechanism of action of the narcotics has not been elucidated, there is general agreement the drugs act on the central nervous system rather than upon peripheral pain receptors.

Biological disposition of narcotics. *Absorption.* Narcotics are absorbed well by all species following hypodermic or intramuscular administration. Absorption from the gastrointestinal tract and other mucous membranes is good but less reliable; codeine is absorbed in this way better than is morphine. Following hypodermic or intramuscular administration of narcotics, results appear within 15 to 30 minutes, with peak effect occurring in 45 to 90 minutes. After intravenous administration, the onset of action is noticeable within one minute and, depending upon the type of narcotic used, reaches its peak in three to six minutes.

Distribution. Free morphine leaves the blood rapidly and, as is the case with other narcotics, accumulates in the parenchymatous organs (the kidney, lung, liver, and spleen), with only a small fraction of the total amount being found in the central nervous system. Narcotics readily cross the placental barrier; concentrations of meperidine, for example, in fetal blood are comparable to those of maternal blood.

Metabolism. The liver is primarily responsible for the metabolic transformation of narcotics. Known and postulated metabolic pathways for morphine include conjugation, dimerization, removal of the methyl group from nitrogen, and addition of a methyl to the 3-oxygen. Narcotics of the ester type, such as meperidine, anileridine, and heroin, are hydrolyzed in the body and excreted as acids. Compounds such as codeine, with an ether linkage in the 3-position, undergo dealkylation. Only the first of these—conjugation with glucuronic acid—has been established for morphine *in vivo*.

Physical dependence. The condition popularly known as drug addiction is more appropriately described as physical dependence upon a narcotic drug. Physical dependence is closely related to drug tolerance, a condition in which standard doses of the drug produce progressively less effect. Dependence has been attributed to various factors, including delayed or incomplete absorption, altered tissue distribution in the central nervous system, or increased rate of metabolism or excretion; it is generally agreed that physiologic dependence develops initially at the cellular level. This has been demonstrated for morphine in cell cultures of several types. Cells of the central nervous system that initially acquire the ability to function in the presence of relatively high levels of narcotics soon become unable to function normally in their absence. (The plasma levels of narcotics that prevent the onset of withdrawal symptoms, however, may be substantially below the peak levels that follow intravenous injection.) Because of tolerance, the euphoria initially produced by the drug diminishes with chronic exposure and progressive physical dependence. The resulting paradox in established addiction is the absence of euphoria and the need for increasing dosage.

Withdrawal symptoms

Upon withdrawal, the following signs and symptoms of physiological dependence in man appear: yawning, tearing, perspiration, running of the nose, dilation of the pupils, tremors, restlessness, abdominal cramps, gooseflesh, defecation, vomiting, and increases in systolic pressure, rectal temperature, and respiratory rate. Subjects who are excessively dependent may experience convulsions, respiratory failure, and death. Dependence develops more rapidly to heroin than to morphine, and more rapidly to morphine than to anileridine, oxymorphone, or meperidine. The administration of a narcotic antagonist can precipitate signs of physiological dependence (see below).

A new approach to the rehabilitation of heroin and other opiate addicts is termed "methadone therapy." Methadone is a narcotic with analgesic and other effects similar to morphine. Unlike most other narcotics, it is effective upon oral administration. It is used in the management of heroin addicts because of its ability to block the euphorogenic effects and cravings for heroin without, apparently, duplicating the physical and mental debilitation produced by that drug. Because of its relatively long duration of effect, a single, daily oral dose of methadone may suffice in most cases of heroin dependence. Methadone itself is addicting, though less so than heroin. In methadone therapy, methadone addiction is substituted for heroin addiction because it is believed that the former allows the patient to function more normally in a social matrix (see also DRUG PROBLEMS).

The clinical use of narcotic analgesics. *General considerations.* It is generally accepted that there is an optimum dose of narcotic analgesics, below which there is insufficient effect and above which there is little gain in analgesia but an exponential increase in undesirable side effects. The side effects include: respiratory and circulatory depression, nausea, vomiting, pruritis, and dysphoria. The optimal therapeutic dose of morphine in a 150-pound man is 15 milligrams. Although all narcotic analgesics share certain common chemical characteristics, there are marked chemical differences between the naturally occurring compounds, typified by morphine, and such synthetic analgesics as meperidine, alphaprodine, or fentanyl. Nevertheless, the pharmacological characteristics of the narcotics as a general class are so similar that specific drugs are selected primarily for differences in duration of effect, degree of sedation, and absorbability from the gastrointestinal tract. The few pharmacological differences creating relative indications for specific drugs are mentioned below, as appropriate.

Circulatory effects

The circulatory effects of narcotics are probably a result of both narcotic influence in the brain stem, where regulatory centres for circulation exist, and a direct action upon the peripheral blood vessels themselves. Since narcotics can cause the liberation from tissues of histamine, a known blood-vessel dilatator, it is likely that dilatation in the peripheral blood vessels following the administration of a narcotic is secondary to the liberation of histamine. (Morphine and most other narcotics cause histamine release although, as yet, this property has not been demonstrated with fentanyl.) Narcotics in general, and morphine in particular, are drugs used in the treatment of pulmonary edema. This condition may occur during congestive heart failure when the flow of blood returning to the heart from the lungs cannot be adequately handled by the left atrium and ventricle. The resulting decrease in available oxygen produces restlessness and apprehension. Morphine, in addition to tranquillizing such a patient, may dilate the peripheral vascular bed and cause the sequestration of sufficient blood to reduce the work load upon the heart, thus improving the clinical situation. Narcotics such as meperidine, which cause an increase in heart rate, would be avoided in such instances.

The crushing chest pain that may accompany coronary occlusion is a prime indication for narcotic analgesics. A given dose of narcotic produces less respiratory depression in a patient with pain than in a patient without pain.

Preoperative. Claude Bernard, the renowned physiologist of the late 19th century, was the first to administer morphine to animals prior to the induction of ether anesthesia. Narcotics are commonly included in preanesthetic medications today; they enhance sedation and amnesia and reduce apprehension and anxiety. In addition, the total dose of anesthetics required is reduced. Narcotics may be especially useful when the patient is in pain prior to surgery, when he suffers from chronic alcoholism or barbiturate or narcotic addiction, and when he should not be exposed to postoperative restlessness and excitement.

Postoperative. The relief of postoperative pain is one of the most frequent indications for administering narcotic analgesics. Sex and nationality have been listed among the factors that may alter the postoperative response to pain (and therefore the need for narcotics), but little documentation exists to substantiate their inclusion. On the other hand, fear, anxiety, and concern have definitely been shown to increase the postoperative need for narcotics. Additionally, such sociological factors as income bracket seem to have an effect: the average number of doses of morphine were tabulated following gallbladder removal in ward patients, semiprivate patients (two beds to a room), private patients, and patients in the most luxurious rooms. The respective numbers of doses of morphine required by these four groups were 3, 5, 9 and 12. Another, more obvious, factor is the site of operation. The most severe pain is likely to occur when the wound is in an area where unavoidable movements (such as breathing) occur. Generally, operations involving the upper abdomen are associated with moderate to severe pain and require narcotic analgesics. When a patient is educated before surgery about the pain he is likely to encounter, the reasons for it, and the means by which he may avoid the movements that will enhance it, there is usually a marked reduction in the need for postoperative narcotics.

During childbirth. Morphine has been used for many years to relieve the pain of childbirth. R. von Steinbüchel was the first to describe the combination of morphine and scopolamine for the relief of labour pain. The combined effect was referred to as "twilight sleep," a translation of the German word *Dämmerschlaf*. While no increase in maternal mortality could be attributed to this combination of drugs, the frequent occurrence of the depressed "blue baby" following its use led many obstetricians to condemn it.

Twilight sleep

Fetal depression attributed to the use of morphine in childbirth has led to many attempts to replace the drug with less depressant narcotic analgesics. Most of the narcotics in Tables 1 and 2 have been used in attempts to combine pain relief for the mother with less respiratory depression for the newborn. It has become almost axiomatic, however, that the degree of respiratory depression parallels the degree of pain relief. There is some evidence that meperidine and alphaprodine are exceptions to this rule; this is probably due to the fact that the immature liver of the newborn is better able to detoxify these drugs than it is morphine and other opioids.

In general anesthesia. Narcotics also are used during general anesthesia. After the production of sleep with a short-acting barbiturate, such as thiopental, one common technique of inducing general anesthesia consists of the administration of nitrous oxide plus oxygen. In many instances, however, nitrous oxide is not sufficiently potent to provide the depth of anesthesia required; intermittent increments of narcotic are then given intravenously for reinforcement. In the recent past, morphine has been used in this fashion with great success, especially in extreme poor-risk patients, such as those undergoing cardiac surgery. The exceedingly high doses of morphine used successfully in such cases attests to the general safety of narcotics when respiratory support is maintained.

The most potent narcotic in current clinical use is fentanyl. This drug is approximately 50 to 100 times more potent than morphine and has recently been employed along with a potent tranquilizer, droperidol. The administration of this combination of drugs produces a state that has been called neuroleptanalgesia (indicating that both the autonomic and the central nervous systems are affected). When nitrous oxide and oxygen are also administered, the state induced is referred to as neuroleptanesthesia.

Narcotic antagonists. Chemically, the narcotic antagonists are usually N-allyl derivatives of narcotic analgesics.

There is a close correlation between the analgesic and respiratory effects of the narcotics themselves and the pharmacological effects of the antagonists derived from them. The mechanism of action of the antagonists is most probably a competition between the narcotic and the antagonist for a single receptor site. When they are administered to human subjects, the narcotic antagonists show some pharmacological effects similar to those of the narcotics. For example, the analgesic activity of nalorphine is as great, milligram per milligram, as that of morphine. On the other hand, the antagonists levallorphan and naloxone have insignificant analgesic effects in doses capable of counteracting narcotic-induced respiratory depression. A significant characteristic of a narcotic antagonist is its inability to produce physiological dependence; this is probably related to the fact that these drugs do not produce the euphoria that results in addiction. Pentazocine, a synthetic agent, was discovered as the result of a search for a narcotic antagonist that could be employed primarily as an analgesic.

Treatment of overdosage

The primary use of narcotic antagonists is in the treatment of accidental or purposeful overdosage with narcotics. They also may be administered to newborns who manifest respiratory depression attributed to the narcotic administered to the mother for purpose of pain relief during labour and delivery. Physical dependence on narcotics can be demonstrated by administering a narcotic antagonist; this so-called "allyl test" precipitates the symtoms of withdrawal in addicts.

BIBLIOGRAPHY. H.K. BEECHER, *Measurement of Subjective Responses* (1959), a critical review of the methods used to assess the effect of analgesics; F.F. FOLDES, M. SWERDLOW, and E.S. SIKER, *Narcotics and Narcotic Antagonists: Chemistry, Pharmacology and Applications in Anesthesiology and Obstetrics* (1964), a reference textbook for physicians who use narcotic analgesics; R.S. KNIGHTON and P.R. DUMKE (eds.), *Pain* (1966), a multidisciplinary approach to the problem of pain, including many references to current trends in the use of narcotic analgesics; A.K. REYNOLDS and L.O. RANDALL, *Morphine and Allied Drugs* (1957), a review primarily intended for the pharmacologist and researcher.

(E.S.S.)

Nasser, Gamal Abdel

No other Arab leader in modern times has succeeded in winning the sometimes hysterical support of Arab masses throughout the Middle East as did Gamal Abdel Nasser during the last 15 years of his life. Even the loss of two wars, with disastrous results for Egypt, did not dim the popularity of this charismatic, almost mythogenic, army officer who became the first true Egyptian to rule the country in several millennia, giving his people the dignity denied them under foreign rule. Yet, he failed in his ambition to create a unified Arab world, and before his death he was forced to sacrifice some of Egypt's political independence for the military support of the Soviet Union.

© Karsh—Rapho Guillumette

Nasser.

Although complex and revolutionary in his public life, privately Nasser was conservative and simple. He knew the loneliness of power. He was more than six feet tall, weighed 200 pounds, had close-cropped graying hair, a swarthy complexion, the physique of an athlete, and a warm, pleasant personality.

Early life

He was born on January 15, 1918, in a mud-brick house on an unpaved street in the Bacos section of Alexandria, Egypt, where his father was in charge of the local post office. In an effort to identify him with the fellahin (peasants), Egyptian government publications for years gave his birthplace as Banī Murr, the primitive Upper Egypt village of his ancestors. From Alexandria, Nasser's father was transferred to al-Khaṭāṭibah, a squalid delta village, where the boy got his first schooling. Then he went to live in Cairo with an uncle who had just been released from a British prison and had rooms in a building occupied by nine Jewish families.

Constantly in trouble with schoolteachers, some of them British, Nasser took part in many anti-British street demonstrations. In one he received a blow on the forehead that left a lifetime scar. After secondary school he went to a law college for several months, then entered the Royal Military Academy, graduating as a second lieutenant.

While serving in the Egyptian Army in The Sudan, he met three fellow officers—Zakaria Mohieddine, later vice president of the United Arab Republic; Abdel Hakim Amer, later field marshal; and Anwar Sadat, who succeeded Nasser as president. Together, they planned a secret revolutionary organization, the Free Officers, whose composition would be known only to Nasser; their aim was to oust the British and the Egyptian royal family.

In the 1948 Arab war against the newly created state of Israel, Nasser was an officer in one of three battalions surrounded for weeks by the Israelis in a group of Arab villages called the Faluja Pocket.

On July 23, 1952, Nasser and 89 other Free Officers staged an almost bloodless coup d'état, ousting the monarchy. Sadat favoured the immediate public execution of King Farouk and some members of the establishment, but Nasser vetoed the idea and permitted Farouk and others to go into exile. The country was taken over by a Revolutionary Command Council of 11 officers controlled by Nasser, with Maj. Gen. Mohammad Naguib as the puppet head of state. For more than a year Nasser kept his real role so well hidden that astute foreign correspondents were unaware of his existence; but in the spring of 1954, in a complicated series of intrigues, Naguib was deposed and placed under house arrest; Nasser emerged from the shadows and named himself prime minister.

Attainment of power

That same year an Egyptian fanatic tried to assassinate Nasser at a mass meeting in Alexandria. When he confessed that he had been given the assignment by the Muslim Brotherhood, Nasser cracked down on this extremist Arab religious organization.

In January 1956 Nasser announced the promulgation of a constitution under which Egypt became an Islāmic, Arab, welfare state with a one-party political system. In June, 99.948 percent of the 5,000,000 Egyptians voting marked their ballots for Nasser, the only candidate, for president. The constitution was approved by 99.8 percent.

As Nasser took titular as well as actual control, Egypt's prospects looked bright. A secret contract had been signed with Czechoslovakia for war material, and Great Britain and the United States had agreed to put up $270,000,000 to finance the first stage of the Aswān High Dam project. But on July 20, 1956, the U.S. secretary of state John Foster Dulles cancelled America's offer; the next day Britain followed suit. Five days later, addressing a mass meeting in Alexandria, Nasser announced nationalization of the Suez Canal, promising that the tolls Egypt collected in five years would build the dam. On October 29, while attending a birthday party for one of his children, Nasser received news that the Israeli Army had invaded the Sinai Peninsula. Two days later, French and British planes attacked Egyptian airfields. Although the Israelis occupied the Sinai Peninsula to Sharm ash-Shaykh and the Egyptian Air Force was virtually de-

stroyed, Nasser emerged from the brief war with undiminished prestige throughout the Arab world.

In *Philosophy of the Revolution*, which he wrote in 1954, Nasser told of "heroic and glorious roles which never found heroes to perform them" and outlined his aspiration to be the leader of the 55,000,000 Arabs, then of the 224,000,000 Africans, then of the 420,000,000 followers of Islām. In 1958 Syria and Egypt formed the United Arab Republic, which Nasser hoped would someday include the entire Arab world. Syria withdrew in 1961, but Egypt continued to be known as the United Arab Republic. That was as close as Nasser ever came to realizing his tripartite dream.

Nasser's accomplishments

But there were other accomplishments. The Aswān High Dam, built with the help of the Soviet Union, began operating in 1968; 20th-century life was introduced into many villages; industrialization was accelerated; land reforms forbade any individual to own more than 100 feddans (104 acres); a partially successful campaign was conducted against corruption; and women were accorded more rights than they had ever had. A new middle class began to occupy the political and economic positions once held in Egypt by Italians, Greeks, Frenchmen, Britons, and other foreigners, whom Nasser now encouraged —sometimes not gently—to leave the country. Nasser's outstanding accomplishment was his survival for 18 years as Egypt's political leader, despite the strength of his opponents: Communists, Copts, Jews, old political parties, rival military cliques, dispossessed landowners, supporters of Naguib, and what was left of the foreign colony.

On the negative side, Nasser made Egypt a police state, in which mail was opened, the communications media were strictly censored, the chief newspapers nationalized, telephones tapped, and visitors' rooms searched. Political democracy in the Western sense was nonexistent. One-party candidates for office were handpicked by Nasser and his close associates. Political enemies were herded into concentration camps in the desert. Life was little changed for most fellahin. The birth rate remained so high as to defeat attempts to increase the living standard.

In foreign affairs Nasser joined Tito of Yugoslavia and Jawaharlal Nehru of India as an advocate of nonalignment, or "positive neutrality." At the Bandung conferences of Asian and African nations in 1955, he emerged as a world figure. His refusal to recognize Israel and Egypt's defeat by Israel in 1956 forced him to divert vast sums into military channels that might have gone to implement his social revolution.

Egyptian troops supported the Republican Army in Yemen's civil war starting in 1962. But they were withdrawn in 1967, when war broke out again between Egypt and Israel after Nasser had requested the United Nations to remove its peace-keeping troops from the Gaza Strip and Sharm ash-Shaykh, then closed the Gulf of Aqaba to Israeli shipping. After the Egyptian Air Force was destroyed on the ground and the Egyptian Army forced to retreat across the Suez, Nasser attempted to resign, but massive street demonstrations and a vote of confidence by the National Assembly induced him to remain in office. The Soviet Union immediately began replacing all the destroyed war equipment and installed surface-to-air missiles along the Suez as a cover for Egypt's artillery emplacements. Nasser had tentatively accepted an American plan leading to peace negotiations with Israel when he died on September 28, 1970, from a heart attack.

BIBLIOGRAPHY. ROBERT ST. JOHN, *The Boss* (1961), the definitive biography; WILTON WYNN, *Nasser of Egypt: The Search for Dignity* (1959), a brief but extremely sympathetic biography by a former Associated Press correspondent in Cairo; KEITH WHEELOCK, *Nasser's New Egypt* (1960), one of the best books about Egypt during Nasser's time; TOM LITTLE, *Egypt* (1958), on Nasser and Egypt by a long-time British expert; UNITED ARAB REPUBLIC, MASLAHAT AL-INSTI'LAMAT, *President Gamal Abdel Nasser* (1965), the official biography; *The Philosophy of the Revolution* (1959), Nasser's own book on his revolutionary concepts and ambitions; MILES COPELAND, *The Game of Nations* (1969), by an ex-CIA official on Middle Eastern intrigues, especially in the United Arab Republic.

(R.St.J.)

Natal

Natal, the smallest of the four provinces of the Republic of South Africa, is bounded on the east by the Indian Ocean, on the south by Cape Province, on the west by Lesotho, the Orange Free State, and Transvaal, and on the north by Swaziland and Mozambique. Its area, which includes Zululand and Tongaland, is 33,578 square miles (86,967 square kilometres), and its population, in 1970, was about 4,250,000. It was given its name by Vasco da Gama, who, on Christmas Day 1497, sighted the entrance to the harbour of Port Natal, today called Durban. Its capital is Pietermaritzburg. The province attained its present extent by 1902.

History. The area was occupied centuries ago by the Nguni branch of the Bantu, who entered in strength from the north. The first European settlers established the trading post, called Port Natal, in 1824, but little attempt was made to develop the interior, whose inhabitants had been decimated by the Zulu chief Shaka by 1820. This apparently empty land was entered by the Voortrekkers, Afrikaners who left Cape Colony, via the Drakensberg (see below) passes in 1837. After the defeat of the Zulus by the Afrikaners at Blood River in 1838, the Afrikaners established the Republic of Natal. When Natal was annexed by the British in 1843, many of the Afrikaner inhabitants left for the interior and were replaced by immigrants, mainly from Britain. From 1860 onward, increasing numbers of Indians entered to work as labourers in the sugar plantations on the coast. The colony was extended by successive acquisitions—notably that of Zululand (the area north of the Tugela River), which was added after the Zulu War of 1879, fought between the Zulu and the British. The lands north of the Buffalo River were added in 1902. During the South African War (1899 to 1902), when the British fought the Boers, the Afrikaner descendants of Dutch settlers, Natal was invaded by the Boer forces who were checked by the British defense at Ladysmith. In 1910 the colony became a province of the Union of South Africa and in 1961 of the republic.

European settlement

The natural landscape. Natal is generally hilly or mountainous, especially along its western border, and rises from the coast to over 11,000 feet at the Drakensberg Escarpment (the edge of the interior plateau of South Africa and of Lesotho). The slope, however, is not gradual and is converted by various outcrops into steps of undulating land separated by escarpments. The lowest step is the coastal plain, less than 500 feet above sea level, widening to the north; then follow surfaces at 2,000 to 2,500 feet and at 3,500 to 4,500 feet (parts of which are referred to as the Midlands). Beyond the province, to the west, lies the Highveld, or high plateau. The continuity of these steps is often interrupted by spurs extending from the Drakensberg and by rivers—notably the Tugela—that cross the province in troughs hundreds of feet below the surrounding countryside.

Relief and drainage

Climate. The climate is generally related to major relief features. Rainfall decreases from more than 50 inches annually along parts of the coast to 30 to 40 inches inland, and then increases to well over 60 inches in the Drakensberg. About 60 percent of the coastal rainfall and 85 percent in the west falls in the summer months, October to March, often as heavy afternoon showers associated with thunderstorms. Temperatures decrease with altitude from the frost-free coast (January 70° to 75° F [21° to 24° C], July 60° to 65° F [16° to 18° C]) to the west (January 60° to 65° F [16° to 18° C], July 40° to 45° F [4° to 7° C]). There are significant day-to-day changes in weather and innumerable local climates due to minor differences in relief. In sheltered valleys or basins annual rainfall is often below 25 inches, and temperature inversions occur in winter (April to September); sea-facing slopes receive higher rainfall, while in the Midlands, mist belts often form in summer along escarpments and hot mountain winds blow from the interior occasionally. In general, summers are hot with occasional rain, and the humidity is high near the coast; hence, the Drakensberg mountain resorts are well patronized at this season. The warm, dry, and sunny winters have made the coast the principal holiday playground of southern Africa.

Vegetation and animal life. The natural vegetation, where it has not been modified or destroyed, falls into several belts according to altitude and climate. Tropical forest or bush, mangrove-fringed in the north, occurs in patches on the more open land along the coast. Farther inland in the north is bushveld, or savanna (grassy parkland), while farther south, in the river basins and valleys, cactus-like euphorbias (spurges) are characteristic. The Midlands and southwest sustain temperate forests, but these now exist only in relict patches. Open country, studded with thorn trees, covers the western half of the province; with true grassland occurring only at the high altitudes.

The province, especially its coastal margin, is regarded as a transitional corridor along which tropical and more temperate animal life and vegetation have spread and intermingled. Measures for conservation are taken. The Natal Parks Board controls 34 parks or game and nature reserves, with a total extent of 527,000 acres, which provide facilities for wilderness trail hiking. The larger reserves are situated in two different environments—along and behind the Zululand coast and in the Drakensberg. In Zululand, rhinoceros, buffalo, hippopotamus, and crocodile are characteristic varieties of animals. Fishing owls and fish eagles are also found, while vegetation includes the fever tree and the crane flower. In the Drakensberg are found the eland (a large African antelope) and the baboon, the martial eagle, and the lammergeier (bearded vulture). Vegetation includes tree ferns and ground orchids. Trout are found in the rivers of the interior, while along the coast, fish catches range from shad to sharks; in winter, sardine or pilchard shoals attract many game fish.

Game and nature reserves

Population. The people of Natal form a mosaic of different ethnic groups living together within a complex social structure. As elsewhere in South Africa, the main distinction is between two composite groups, white and black, who, in accordance with the government's policy of apartheid (separate development), live in different but coexistent worlds. The whites are mainly of British or Dutch descent, while the nonwhites include a Bantu majority (*i.e.*, Africans speaking Bantu languages), a number of Asians (Indians), and a group of Coloured people who are the product of interbreeding. In the early 1970s the Bantu people had the highest rate of increase, followed by the Coloureds, Indians, and Europeans. In 1970, out of a total of about 4,250,000, there were 3,210,000 Bantu, 500,000 Indians, 450,000 Europeans, and 70,000 Coloureds. Most of Natal's people live along and behind the coast (except in the north) or in the centre of the province; the extreme west and northeast is lightly populated. The Bantu are especially concentrated in many fragmented areas (now referred to as homelands), which, in 1970, were designated as the Zulu Territorial Authority and granted some autonomy. These homelands cover about a third of the province, and—except in the northeast—consist of broken rugged country. About half the Bantu live in them; the remainder live in the European rural areas or towns. Two-thirds of Natal is European-owned, and most Europeans live in or near Durban, in the towns along the route running west to the Witwatersrand ridge on South Africa's central plateau, or along the coast. More than 80 percent of South Africa's Indians live in Natal, and these and the Coloured people have a distribution similar to the Europeans.

Bantu homelands

Urban–rural distribution. Urbanization is proceeding apace. In 1970, out of the total population, a little over one-third (38 percent) was urbanized. About 90 percent of the Europeans, 89 percent of the Coloureds, 85 percent of the Indians, and 22 percent of the Bantu are urbanized. The towns are a European innovation, superimposed upon a previous Bantu rural settlement pattern, but they have proved increasingly attractive to the other population groups who live in separate residential areas within them, and few towns now have a majority of Europeans. Durban (population 1,000,000) and to a lesser extent Pietermaritzburg (170,000) dominate the urban scene. The other towns are few in number and are considered small in size.

Customs, language, and religion. Each population group has a cultural identity reflected in its religion, language, and customs. Some 70 percent of the Europeans have English as their basic language, most of the remainder speak Afrikaans, and many Europeans speak both languages. The Europeans are mainly Christians; the Coloured people have a close affinity to the Europeans and have a similar culture. The Indians, on the other hand, have adopted many elements of the Western culture, but they retain an identity and a diversity by maintaining much of their own cultures. Some 70 percent profess Hinduism, 20 percent are Muslims, and the remainder are largely Christians. Most are fluent in English, but about 85 percent speak various Indian languages at home. The Bantu have not been culturally assimilated to the same extent as the other non-Europeans. They have retained their identity through their use of Zulu, which is the language of some 90 percent of the Bantu, and through a rich heritage of folklore, ceremony, and customs that reflect the diversity of tribal allegiances. The rest of the Bantu speak cognate languages.

Administration. The constitution of the republic provides for provincial administration from Pietermaritzburg, through an administrator (chief executive officer), a provincial council, and its executive committee. The province has powers in such matters as taxation, education, hospitals, charitable institutions, and local works. The local administrative unit is the magisterial district. The province, under the aegis of the central government, imposes duties and grants certain rights to local government authorities functioning through municipal and borough councils or town boards—normally under the chairmanship of mayors.

Social conditions. *Public health.* The provincial administration, local authorities, voluntary bodies, and a range of welfare authorities cooperate with the South African Department of Health in caring for health and welfare. The province controls hospital services, which, in 1965, included some 19,000 beds in 20 provincial and many private institutions, some 30 of which were in the Bantu homelands. Medical facilities are largely concentrated in the towns but are supplemented by many outlying clinics and health centres. Medical training for doctors is available for nonwhites at Natal University, Durban. There are about 900 doctors in the province. Diseases that especially affect the Bantu include tuberculosis, bilharzia (a parasitic disease of the blood), and kwashiorkor (malnutrition resulting from severe protein deficiency).

Education. Education below the university level is provided for whites at some 360 schools, two technical colleges, and three teacher training colleges; and for nonwhites at about 1,200 schools, nine technical colleges, and ten teacher training colleges. The University of Natal, which was founded in 1909 as a constituent college of the University of South Africa and assumed university status in 1949, is the chief institution of higher education. It is located partly in Pietermaritzburg and partly in Durban. Higher education for Indians is available at the University of Durban-Westville—established in 1960—and for Bantu at the University of Zululand, located at Kwa Dlangezwa, near Empangeni, established in 1957.

Universities

Economy. Natal has two coexistent economies. One is the subsistence economy of the Bantu homelands based upon pastoralism (largely cattle raising) and cultivation (mainly of corn [maize] grown in small plots) and supplemented by migrant labour. The homelands have poor resources, although considerable effort is being directed toward their development; their economy is supplemented by the earnings of Bantu who work elsewhere in South Africa. The other is the advanced commercial economy pursued by the whites, with the assistance of nonwhites, elsewhere in the province. It is through this economy that most of the considerable and varied resources of the province are being developed.

Minerals. The mineral wealth of Natal consists essentially of coal, which is mined in the Newcastle-Vryheid-Dundee-Utrecht area of the northwest; 85 percent of the output is high-grade bituminous coal, much of it of cok-

ing grade. This coalfield is South Africa's chief source of coking and semi-anthracite coal. Annual production is about 10,000,000 tons.

Agriculture and forestry. The most important agricultural area lies along the coast, where sugarcane is the major crop. Production amounts to some 1,500,000 tons of sugar, grown on 700,000 acres and processed by 18 mills. Sugar growing is now of increasing importance in the Midlands. Refining for domestic and export requirements is mainly carried out at Durban. Plantations of pine and eucalyptus provide raw materials for sawmills and for paper, hard and soft board, and rayon pulp mills. Subtropical fruits, especially pineapples and bananas, are also grown. Durban is also the base for an important whaling industry. In the Midlands, plantations of pine, eucalyptus, poplar, and especially black wattle cover large areas. A tanning agent, of which South Africa is the largest world producer, is made from the wattle bark.

Dairying is important. Milk is produced near Durban and Pietermaritzburg; butter and cheese are produced inland. The northwest is largely devoted to cattle raising and maize cultivation, although the Drakensberg Escarpment cannot be used for agriculture because of its rough terrain. The northeast coastal plain is also unproductive, but projected irrigation schemes may provide the impetus for a final movement of the economic frontier of the province to the Mozambique border.

Industry. The products of mining, farming, forestry, and fishing, together with imported materials, support important manufacturing industries. Most are established in the towns, with the number of factories and output values roughly proportional to urban populations. Rural industries are generally restricted to those where processing takes place on the spot, although since 1960 decentralization in the form of the establishment of border industry (industry located just outside the Bantu homelands) has been in progress, notably at Elangeni (Hammarsdale), between Durban and Pietermaritzburg.

Rural and border industries

Durban, with nearby Pinetown, is the economic heart of the province and its industrial centre. It possesses over 80 percent of Natal's factories, and—with more than 10 percent of those of the republic—is South Africa's second or third most important industrial region. Its factories are associated primarily with shipping, food processing, chemicals, sugar, and oil refining. An oil pipeline connects Durban with the southern Transvaal. Pietermaritzburg has a number of industries, including an aluminum plant, several footwear factories, and food processing establishments. The northern coalfield has not yet developed industrially, but plans for the development of the Tugela Basin and a recent decision to establish a large steel plant at Newcastle may engender considerable industrial growth. Production was due to begin on a reduced scale in 1974.

Transportation and communications. A railway network consists of a main electrified axis from Durban, northwest through Pietermaritzburg to the Witwatersrand, a line along the coast, and various branch lines. The roads parallel the railways and also serve more remote areas; main roads are tarred, country roads have gravel or earth surfaces. Aircraft fly from Durban's airport to South African cities, and to Mozambique and Rhodesia, and—via Jan Smuts Airport (Johannesburg/Pretoria)—to world centres; other towns have smaller airfields.

A completely new port and industrial growth point is being developed at Richard's Bay, 100 miles north of Durban; dredging began in 1972, and the port was expected to open in 1976. An aluminum smelter was already in production in the early 1970s. In the meantime, however, Durban remains Natal's only port. It serves much of the interior of southern Africa and is South Africa's main cargo port. It has shipping links with the world, and its importance was considerably increased by the closure of the Suez Canal.

Cultural life. The people of Natal have a rich cultural heritage drawn from their differing backgrounds. Generally speaking, that of the Europeans and Coloureds is of the West, the Indians of the East, and the Bantu of Africa, although non-Europeans are becoming increasingly westernized. Cultural institutions include botanic gardens, parks, museums, libraries, art galleries, cinemas, theatres, music societies, and agricultural and industrial exhibitions in many centres, notably Durban and Pietermaritzburg. Of particular importance is the recently established Natal Performing Arts Council.

BIBLIOGRAPHY. Authoritative studies include: *Natal Regional Survey*, vol. 1–15 (1951–); *Reports* of the Department of Economics, University of Natal (1951–); and *Reports* of the Natal Town and Regional Planning Commission, Pietermaritzburg (1952–). See also E.H. BROOKES and C.B. WEBB, *A History of Natal* (1965); E.J. KRIGE, *The Social System of the Zulus*, 2nd ed. (1951); F. MEER, *Portrait of Indian South Africans* (1969); and various issues of the *South African Geographical Journal.*

(O.W.)

National Income Accounting

The purpose of national income accounting, often called simply national accounting, is to furnish measures of the income and production of a nation.

National production, or output, is the total of final, or end, products produced by the nation during a specified period. That is to say, it is the sum of products available for consumption or for additions to the stock of capital. It does not include raw materials and semifinished products used up in production. For instance, if flour is baked into bread, only the bread is counted. The totals are expressed in monetary values, since these provide a common unit for summing the wide variety of goods and services produced. In Communist countries, it should be pointed out, the concept of national product includes only material goods and "productive" services; other services, such as those of government workers, soldiers, musicians, and household help are not counted.

Measures of national output. One commonly used measure of output is gross national product (GNP), which is the country's total output of goods and services valued at market prices. It includes also output produced abroad to which residents of the country have a claim. For some purposes it is desirable to count only the output produced within the country; the resulting total is called gross domestic product. Both of these measures of output are referred to as "gross" because they make no allowance for the amount of fixed capital (buildings, equipment, and inventories) used up in the process of production. It is logically inconsistent not to allow for the using up of fixed capital, since it cannot be considered output, and therefore statisticians try to estimate net national product by deducting an allowance for capital depreciation.

The three foregoing concepts of national output measure it in terms of market prices. The prices at which goods and services are sold include a substantial amount of indirect taxes (sales and turnover taxes, payroll taxes, and some property taxes). These may be subtracted from net national product to arrive at a figure that reflects the actual costs of production, called net national product at factor prices (or factor costs). This is generally known as national income, since it is the sum of wages, salaries, profits, and other incomes received by the population of the country.

Relationship between national product and national income

The relationship between national product and national income is basic to the understanding of national accounting statistics. It resembles on a larger scale the operations of double-entry bookkeeping in a business enterprise. It may in fact be seen as the summation of the operations of all the business enterprises in the country, if one also includes the government and the activities of paid household help. The typical business enterprise produces goods that it either sells or adds to its inventories. Chargeable to this production are purchases of raw materials and other intermediate products from other enterprises (but not purchases of fixed plant and equipment); wages, salaries, and other incomes paid out in production; indirect business taxes, such as sales, excise, and property taxes; and depreciation and kindred allowances for fixed capital used up. A residual item of profit or loss equalizes the expenses chargeable to production with its value.

If a balance sheet is prepared for the business system as

Table 1: Expenditure on Gross National Product at Market Prices* (1968)

	Argentina	Japan	United Kingdom	United States	Federal Republic of Germany
Private consumption	4,208.7	26,682	27,020	538.9	297.3
General government consumption	657.9	4,288	7,747	184.6	83.5
Gross domestic fixed capital formation	1,245.5	17,242	7,798	146.2	121.3
Increase in stocks	3.1	2,425	204	8.5	10.2
Exports of goods and services	547.6	5,353	8,610	43.6	121.7
Less imports of goods and services	501.9	4,763	9,038	45.8	103.4
Gross domestic product	6,161.0	51,223	42,341	876.0	530.7
Net factor income from the rest of the world	−50.8	−150	558	4.7	0.1
Gross national product	6,110.2	51,077	42,899	880.8	530.8
Less indirect taxes	—	3,833	6,960	80.2	72.4
Plus subsidies	—	551	886	0.8	5.8
Less provisions for the consumption of fixed capital	—	6,574	3,375	84.1	59.3
Statistical discrepancy	—	−405	—	2.5	—
National income (net national product at factor cost)	—	40,817	33,450	719.8	404.9

*In 000,000,000 Argentine pesos, 000,000,000 Japanese yen, 000,000 British pounds, 000,000,000 U.S. dollars, and 000,000,000 deutsche marks.
Source: United Nations Statistical Office, *Yearbook of National Accounts Statistics 1969.*

a whole, those sales of enterprises that represent purchases of intermediate products by other enterprises cancel one another. Thus, purchases of intermediate products disappear from the left-hand side of the statement, and sales on the right-hand side are correspondingly reduced. Matching the noncancelling final products there remains a numerically equivalent total consisting of income streams (including profits), indirect taxes, and capital-consumption allowances. (Inasmuch as the income aggregate reflects current production, capital gains and losses are not included.)

Non-market activities

National product as defined for statistical purposes does not include all the goods and services produced in the country. The bulk of the production covered by most measures of national output consists of items that are intended for sale in the market. Goods provided outside the marketplace are usually excluded, even though they may resemble items that are included. For instance, the services of housewives are usually excluded, while services rendered by domestics are included. On the other hand, some goods and services that do not enter the market are included in the statistics. Monetary values are generally imputed to wages paid in kind, to food and fuel produced on farms for home consumption, to the services owner-occupants derive from their homes, and sometimes also to certain services, such as the handling of bank deposits, which financial institutions render free of charge to individuals. The treatment of nonmonetary items raises difficult problems in the measurement of the output of the less developed countries, where many types of productive activity that are channelled through the market in Western countries are performed within the confines of the individual household.

For some purposes it is useful to have a measure of the actual incomes received by individuals. One such measure is personal income, or the sum of the incomes received by individuals. It is derived by subtracting from national income all earnings of corporations not paid in dividends and adding to it government payments such as welfare, social security, and interest on the national debt. A related concept, disposable income, subtracts all taxes in order to arrive at an estimate of the income available to private individuals for spending.

Methods of presentation. Statistics of national output can be broken down in a number of ways according to the purposes for which they are to be used. One way, as shown in Table 1, is according to the character of the products—that is, whether goods are used for consumption (both private and governmental), investment, or export. Each category can, of course, be broken down into further detail to show sales of products to various groups of consumers, businesses, and foreign purchasers. Table 1 also shows the relationship between national income and gross national product: the former is equal to the latter, minus indirect taxes on business, plus subsidies to business, and minus an allowance for the depreciation of fixed capital.

The national output may also be analyzed according to the industries in which it is created. Table 2 does this for five countries in terms of broad categories such as manufacturing, construction, and wholesale and retail trade. As one would expect, the proportion of the national product originating in the manufacturing sector is relatively large in countries such as the United States and the United Kingdom, whereas the importance of the agricultural sector is relatively greater in less industrialized countries such as Turkey.

Table 2: Industrial Origin of Gross Domestic Product at Factor Cost* (1968)

	Argentina	Turkey	United Kingdom	Italy	United States
Agriculture, forestry, hunting, and fishing	782.1	33.7	1,127	4,591	25.3
Mining and quarrying	97.8	1.7	687	309	14.2
Manufacturing	1,683.8	16.3	12,527	11,290	246.4
Construction	305.5	7.5	2,455	3,341	39.5
Electricity, gas, and water	161.5	1.0	1,288	1,046	20.0
Transport, storage, and communication	532.1	8.5	3,065	3,017	53.6
Wholesale and retail trade	721.3	8.2	4,082	5,033	142.2
Banking, insurance, and real estate	245.5	3.8	882	2,091	55.1
Ownership of dwellings		6.0	1,801	2,310	64.1
Public administration and defense	545.6	10.1	2,582	4,924	123.1
Services	356.6	5.3	6,549	3,485	95.3
Gross domestic product at factor cost	5,431.8	102.1	36,267†	41,437	876.0‡

*In 000,000,000 Argentine pesos, 000,000,000 Turkish liras, 000,000 British pounds, 000,000,000 Italian lire, and 000,000,000 U.S. dollars. †Includes adjustments for inventory valuation and for statistical discrepancy. ‡Includes adjustment for statistical discrepancy.
Source: United Nations Statistical Office, *Yearbook of National Accounts Statistics 1969.*

The value of national output may be broken down according to the various types of incomes—wages and salaries, interest, rents, profits, and others—originating in production. Table 3 shows how the importance of the different categories varies among five countries.

Another useful form of national income account shows the financial relationships among various groups in the economy. This makes it possible to analyze the sources of savings and the channels by which they find their way into investment. Table 4 is a simplified presentation of the financing of capital investment. It shows that the largest single source of investment is saving (in the customary meaning of the term) by government and business as well as by private households. A very large part of investment, however, shows up in the accounts as the using up of

Table 3: Distribution of the National Income*
(1968)

	Canada	Ceylon	France	United Kingdom	United States
Compensation of employees	35,921	4,331	296.0	25,267	515.5
Income from unincorporated enterprises	5,444	4,566	110.6	2,391	63.8
Income from property	5,645	297	40.2	4,149	98.4
Net rent	4,767	259	16.5	989	54.1
Interest			15.1	3,160	21.2
Dividends	878	38	8.6		23.1
Corporate transfer payments	87	—	0.5	†	3.4
Saving of corporations	1,503	113	21.2	1,292	23.4
Direct taxes on corporations	2,619	142	11.5	1,068	41.3
General government income from property and entrepreneurship	2,136	93	3.8	1,204	—
Less interest on the public debt	2,262	166	8.2	1,792	11.9
Less interest on consumers' debt	330	13	0.8	—	14.2
National income	50,458‡	9,363	475.0§	33,450‖	719.8

*In 000,000 Canadian dollars, 000,000 Ceylonese rupees, 000,000,000 French francs, 000,000 British pounds, and 000,000,000 U.S. dollars. †Included in income from property. ‡Includes adjustment for stock valuation. §Includes a small amount of un-itemized income. ‖Includes adjustment for statistical discrepancy.
Source: United Nations Statistical Office, *Yearbook of Accounts Statistics 1969.*

Table 4: Finance of Domestic Investment*
(1968)

	Australia	Taiwan	South Africa	United Kingdom	United States
Capital consumption (depreciation allowances)	2,356	10.3	955	3,375	84.1
Noncorporate private sector	—	7.3	674	615	27.4
Private corporations	—			1,300	45.9
Public corporations	—	3.0	105	999	
Government enterprises	—		131	240	10.8
General government	—	—	45	221	
Saving	4,698	28.6	1,487	4,491	72.8
General government	1,557	5.2	508	1,735	11.0
Public corporations		4.1	21	−190	
Private corporations	1,209	19.3	386	1,482	23.4
Households and private nonprofit institutions	1,932		572	1,464	38.4
Deficit of the nation on current account	961	4.2	−90	265	0.3
Statistical discrepancy	52	—	—	−129	−2.5
Domestic investment	8,067	43.1	2,352	8,002	154.7

*In 000,000 Australian dollars, 000,000,000 New Taiwan dollars, 000,000 South African rand, 000,000 British pounds, and 000,000,000 U.S. dollars.
Source: United Nations Statistical Office, *Yearbook of Accounts Statistics 1969.*

capital that is already in existence; it is written off the books in annual depreciation allowances.

More elaborate breakdowns of the financial relationships in the national economy, known as flow-of-funds accounts, are made in some countries. These show in considerable detail the amount of borrowing and lending in various sectors of the economy by various kinds of institutions.

Input–output statistics

Two other ways of presenting statistical information on an aggregate basis may also be mentioned. Input–output statistics show the flows of products between industries —for example, the quantities of coal and iron ore used by the steel industry or the amounts of steel, rubber, glass, and other materials needed in the production of automobiles. Input–output statistics are used for such purposes as estimating the total effect on the economy of changes occurring in a single industry or the effects on different industries of a change in the supply of a particular material.

Statistics of national wealth have been developed in close relation to national income statistics. The difference between income and wealth is that the latter consists of stocks, or assets, whereas the former consists of flows. National wealth is the sum of domestic tangible assets and the net foreign balance. Changes in tangible assets can be estimated on the basis of changes in flows of national income, primarily saving and investment.

Uses of national income statistics. The concepts of national product and national income provide a means of appraising the performance of the economy as a whole. By reducing the great complexity of economic activity to a single unit of measurement, they make it possible to compare one year's level of output with another's or to measure changes in the distribution of income and product among different groups or sectors. To give only a few examples, the gross domestic product of Japan grew by more than 10 percent annually during the 1960s, as compared with less than 5 percent annually in the case of West Germany and the United States and around 3 percent for Great Britain—a comparison that sheds some light on the economic problems of those countries. The relative importance of agriculture, hunting, forestry, and fishing to total production has declined in recent decades in the industrialized countries: in the United States it declined from about 5 percent of the domestic product in 1953 to less than 3 percent in 1968; in Japan, from about 21 percent in 1953 to about 10 percent in 1968; and, in West Germany, from just over 9 percent in 1953 to under 4 percent in 1968. (Agricultural production, it should be noted, did not decline but increased; the production of other sectors, however, particularly of manufacturing, increased much more rapidly.)

The best known uses of national income and product estimates are for public-policy purposes. Economic planning in the non-Communist industrial countries is nowadays based upon projections of national income. The effects of an increase in business investment or in government spending or of a tax reduction upon national income are widely discussed, and these statistics have even crept into the speeches of presidents and prime ministers (see FISCAL AND MONETARY POLICY). Elaborate forecasting techniques have been developed for predicting the consequences of such changes upon the GNP and upon its various components (see ECONOMIC FORECASTING and ECONOMETRICS).

Limitations of national income statistics

The wide use of national income statistics does not mean that those who employ them are able to predict the future. Such estimates provide only an indication of what has already happened, and there is no guarantee that the regularities disclosed in them will persist. The behaviour of savers and investors is notably irregular, even when taken in the aggregate; no one can be certain how the general public will respond to a change in the income tax or whether the stated plans of businessmen will bear fruit in the capital market.

Sources and methods. The creation of reliable statistics of national income and product require large staffs of skilled researchers. The components of national output must be estimated on the basis of data that have been prepared largely for other purposes. These data include census enumerations, sample surveys, and statistics that are by-products of various governmental activities such as the administration of social security systems, tax laws, and expenditure programs. Many other sources, too varied to summarize, are also used. The basic data often depart definitionally and in coverage from the items in the national income accounts. Estimation of these items involves, accordingly, the integration of all the available —and sometimes conflicting—information and the filling of gaps by resorting to partial and indirect evidence.

National income figures are thus subject to error not only because of inaccuracies in the basic data but also because of imperfections in the estimating methods. Moreover, their margin of error cannot be precisely quantified because the usual mathematical techniques for measuring error are not applicable to this complex case. Estimates of the margin of error that are occasionally given with national income figures fall short of definitional precision and incorporate a great deal of subjective judgment.

The reliability of national income statistics varies greatly from country to country and within each country among the several entries in the accounts and according to the time period covered. Reliability is likely to be vastly higher for advanced industrial countries than for countries in which industrial processes have not de-

veloped. This is so mainly because industrial economies are characterized by transactions that lend themselves to registration in statistical form and because the problems confronting such economies lead to extensive data collection.

Relative errors are likely to be smaller in broad aggregates than in separately estimated components because of offsetting errors. For any given series, estimates referring to the remote past and to very recent periods are likely to be less accurate than estimates for intervening dates. This is because the quantity of primary data diminishes rapidly as one moves into the past and because estimates for recent periods are usually made before all the source material is at hand. Quarterly and monthly series are likely to be less accurate than annual estimates for similar reasons.

Skill is required not only in the preparation of national income statistics but also in their use. For some purposes it may be necessary to go beyond the published estimates and examine the methodology underlying them in order to form a judgment as to their reliability for a particular purpose.

National income estimates were prepared in England as early as the 17th century. But workmanlike estimates covering a span of years and built up from detailed components are largely a product of the 20th century. The economic problems associated with the depression of the 1930s and World War II gave impetus to the development of the statistics, largely under governmental aegis. National income estimates are now available for a large number of countries, but they reflect the wide differences among them in technologies, institutions, and statistical methods. The United Nations Statistical Office has been active in promoting international comparability, both by developing standard accounting systems and by preparing compilations in which reported series are adjusted to uniform definitions.

Difficulties of international comparisons

Comparability. Because countries differ greatly in their economies and patterns of life, comparisons of their GNP or income are apt to be misleading. If, for instance, the output of radios in one country is double that in another while the output of television sets is one-half, it is not possible to make an unambiguous comparison of the outputs of radio and TV sets combined. The result will vary according to the relative value attached to each. Thus if one country values its TV sets at three times its radio sets, while the other country values its TV sets at two times its radio sets, their statistics of total output may show one country's national product as being 40 percent greater than the other's or only 25 percent greater, depending upon which country's values are used. In reality the problem is much more complex because of the multitude of items making up each country's national output. For this reason, tables showing national income per capita for a number of different countries expressed in U.S. dollars cannot be considered very precise, even when there is no question as to the relative values of currencies and the accuracy of population estimates.

A similar difficulty arises in making comparisons over time. If in one time period the output of transportation equipment consists of horse-drawn carriages and in another of automobiles, it is not really possible to make a quantitative comparison between the two.

Another kind of difficulty arises when countries use different definitions of their national product. The national output statistics of the Soviet Union and other Communist countries cannot be directly compared with those of the United States and other Western industrial countries because of the broad difference in what is included.

Problems of design. Several unsolved problems have to do with the design of national income statistics.

As noted earlier, measures of national output are confined largely to output intended for sale, but a few departures from this rule are made by imputing values to certain nonmarket items. These imputations are a source of irritation logically because they do not flow from any general definition of output that can be specified. Moreover, the treatment of nonmarket production constitutes an acute practical problem in the analysis of economic conditions in countries whose market economies are undeveloped and also in international comparisons. Accordingly, much thought has been devoted to nonmarket output. Some progress has been made through the proposal of international conventions for its treatment, and ad hoc solutions may be found to serve the needs of specific analyses. But it seems doubtful whether an operational definition of national output not anchored to the market criterion can ever be devised.

Intermediate and final products

Questions have been raised as to the proper distinction between final output and intermediate products. Specifically, it has been urged that not all of the output purchased by government be considered final, as is commonly done. Some of it should be excluded from the national output total, it is said, as are intermediate products. Somewhat more tentatively it has been suggested that not all output bought by consumers be regarded as final either. No clear-cut proposals have been made, but the general idea appears to be to distinguish purchases that are direct causes of final human satisfaction from those that are only indirect conditions of it and to exclude the latter. Opponents of the idea have argued that the proposed distinction cannot in fact be drawn; and they have not been refuted, either in logic or through the presentation of a viable alternative to the concept in use.

For some purposes it would be desirable to extend the definition of capital formation (investment) beyond that of tangible capital formation by business to intangible items, such as expenditures for industrial research, and to capital formation by government and consumers. Little has been done to recognize intangible capital, but capital expenditures by government have increasingly been taken into account instead of being treated as current consumption.

The acquisition of residences is the only consumer purchase that is generally recognized as capital formation. There are sound reasons for extending this treatment to other consumer durables, inasmuch as these also outlast the normal accounting period and thus constitute additions to wealth.

Measurement of capital formation on a gross basis (*i.e.*, before deduction of capital consumption allowances) raises problems because it is difficult to frame a uniform definition. It is hard to draw a systematic line, for instance, that will either exclude or include repair and maintenance expenditures. In addition, the gross measure is insufficient because it does not show net additions to the capital stock. But it appears difficult also to obtain a net capital formation measure that is meaningful for output measurement.

The conventional capital consumption allowances made by business are inadequate. To an increasing extent they are made noncomparable over time by changes in tax laws. In addition, they are usually stated in terms of the original cost of the capital goods and hence cannot be deducted from gross capital formation, which is expressed in current prices, to obtain a meaningful indication of the net change in capital stock. The realism of the methods used to allocate capital consumption charges over the lifetime of the assets is also often in doubt. It is not easy, however, to devise alternative, more meaningful measures of capital consumption.

BIBLIOGRAPHY. Most contemporary textbooks of economics contain discussions of the methods and purposes of national income accounting. Basic introductory works are J.R. HICKS, *The Social Framework of the American Economy*, 2nd ed. (1955); JAMES E. MEADE and RICHARD STONE, *National Income and Expenditure*, 8th ed. (1966); RICHARD and NANCY RUGGLES, *National Income Accounts and Income Analysis*, 2nd ed. (1956); and PAUL STUDENSKI, *The Income of Nations (with Corrections and Emendations)*, 2 vol. (1961). The UNITED NATIONS STATISTICAL OFFICE publishes uniform national income data for many of the member countries of the UN in its *Yearbook of National Accounts Statistics*. It has also published a survey of the methods used in various countries in *National Accounting Practices in Sixty Countries* (1964). The UNITED STATES DEPARTMENT OF COMMERCE, OFFICE OF BUSINESS ECONOMICS, publishes regular accounts for the United States in its monthly *Survey of Current Business*. See also its *National Income and Product Accounts of the United States, 1929–1965: Statistical Tables* (1966).

Nationalism

Nationalism may be defined as a state of mind in which the individual feels that everyone owes his supreme secular loyalty to the nation-state. Nationalism is a modern movement. Throughout history men have been attached to their native soil, to the traditions of their parents, and to established territorial authorities; but it was not until the end of the 18th century that nationalism began to be a generally recognized sentiment molding public and private life and one of the great, if not the greatest, single determining factors of modern history. Because of its dynamic vitality and its all-pervading character, nationalism is often thought to be very old; sometimes it is mistakenly regarded as a permanent factor in political behaviour. Actually, the American and French revolutions may be regarded as its first powerful manifestations. After penetrating the new countries of Latin America, it spread in the early 19th century to central Europe and from there, toward the middle of the century, to eastern and southeastern Europe. At the beginning of the 20th century, nationalism flowered in the ancient lands of Asia and Africa. Thus the 19th century has been called the age of nationalism in Europe, while the 20th century has witnessed the rise and struggle of powerful national movements throughout Asia and Africa.

Nationalism, translated into world politics, implies the identification of the state or nation with the people—or at least the desirability of determining the extent of the state according to ethnographic principles. In the age of nationalism, but only in the age of nationalism, the principle was generally recognized that each nationality should form a state—its state—and that the state should include all members of that nationality. Formerly states, or territories under one administration, were not delineated by nationality. Men did not give their loyalty to the nation-state, but to other, different forms of political organization: the city-state, the feudal fief and its lord, the dynastic state, the religious group, or the sect. The nation-state was nonexistent during the greater part of history, and for a very long time it was not even regarded as an ideal. In the first 15 centuries of the Christian Era, the ideal was the universal world-state, not loyalty to any separate political entity. The Roman Empire had set the great example, which survived not only in the Holy Roman Empire of the Middle Ages, but also in the concept of the *res publica christiana* ("Christian republic" or community) and in its later secularized form of a united world civilization and in world policy as it appeared in the writings of the 17th century.

As political allegiance, before the age of nationalism, was not determined by nationality, so civilization was not thought of as nationally determined. During the Middle Ages, civilization was looked upon as determined religiously; for all the different nationalities of Christendom as well as for those of Islam there was but one civilization —Christian or Muslim—and but one language of culture —Latin (or Greek) or Arabic (or Persian). Later, in the periods of the Renaissance and of Classicism, it was the ancient Greek and Roman civilizations that became a universal norm, valid for all peoples and all times. Still later, French civilization was accepted throughout Europe as the valid civilization for educated people of all nationalities. It was only at the end of the 18th century that, for the first time, civilization was considered to be determined by nationality. It was then that the principle was put forward that a man could be educated only in his own mother tongue, not in languages of other civilizations and other times, whether they were classical languages or the literary creations of other peoples who had reached a high degree of civilization.

From the end of the 18th century on, the nationalization of education and public life went hand in hand with the nationalization of states and political loyalties. Poets and scholars began to emphasize cultural nationalism first. They reformed the mother tongue, elevated it to the rank of a literary language, and delved deep into the national past. Thus they prepared the foundations for the political claims for national statehood soon to be raised by the people in whom they had kindled the spirit.

Cultural nationalism

Before the 18th century there had been evidences of national feeling among certain groups at certain periods, especially in times of stress and conflict. The rise of national feeling to major political importance was encouraged by a number of complex developments: the creation of large, centralized states ruled by absolute monarchs who destroyed the old feudal allegiances; the secularization of life and of education, which fostered the vernacular languages and weakened the ties of church and sect; the growth of commerce, which demanded larger territorial units to allow scope for the dynamic spirit of the rising middle classes and their capitalistic enterprise. This large, 18th-century, unified territorial state with its political and economic centralization became imbued with a new spirit—an emotional fervour similar to that of religious movements in earlier periods. Under the influence of the new theories of the sovereignty of the people and the rights of man, the people replaced the king as the centre of the nation. No longer was the king the nation or the state; the state had become the people's state, a national state, a fatherland. State became identified with nation, as civilization became identified with national civilization.

The roots of national feeling

That development ran counter to the conceptions that had dominated political thought for the preceding 2,000 years. Hitherto man had commonly stressed the general and the universal and had regarded unity as the desirable goal. Nationalism stressed the particular and parochial, the differences, and the national individualities. Those tendencies became more pronounced as nationalism developed. Its less attractive characteristics were not at first apparent. In the 17th and 18th centuries, the common standards of Western civilization, the regard for the universally human, the faith in reason (one and the same everywhere) as well as in common sense, the survival of Christian and Stoic traditions—all of these were still too strong to allow nationalism to develop fully and to disrupt society. Thus nationalism in its beginning was thought to be compatible with cosmopolitan convictions and with a general love of mankind, especially in western Europe and North America.

EUROPEAN NATIONALISM

The first full manifestation of modern nationalism occurred in 17th-century England, in the Puritan revolution. England had become the leading nation in scientific spirit, in commercial enterprise, in political thought and activity. Swelled by an immense confidence in the new age, the English people felt upon their shoulders the mission of history, a sense that they were at a great turning point from which a new true reformation and a new liberty would start. In the English revolution an optimistic humanism merged with Calvinist ethics; the influence of the Old Testament gave form to the new nationalism by identifying the English people with ancient Israel.

The new message, carried by the new people not only for England but for all mankind, was expressed in the writings of John Milton in whose famous vision the idea of liberty was seen spreading from Britain, "celebrated for endless ages as a soil most genial to the growth of liberty" to all the corners of the earth.

> Surrounded by congregated multitudes, I now imagine that . . . I behold the nations of the earth recovering that liberty which they so long had lost; and that the people of this island are . . . disseminating the blessings of civilization and freedom among cities, kingdoms and nations.

English nationalism then was thus much nearer to its religious matrix than later nationalisms that rose after secularization had made greater progress. The nationalism of the 18th century shared with it, however, its enthusiasm for liberty, its humanitarian character, its emphasis upon the individual and his rights and upon the human community as above all national divisions. The rise of English nationalism coincided with the rise of the English trading middle classes. It found its final expression in John Locke's political philosophy, and it was in that form that it influenced American and French nationalism in the following century.

American nationalism was a typical product of the 18th

century. British settlers in North America were influenced partly by the traditions of the Puritan revolution and the ideas of Locke and partly by the new rational interpretation given to English liberty by contemporary French philosophers. American settlers became a nation engaged in a fight for liberty and individual rights. They based that fight on current political thought, especially as expressed by Thomas Jefferson and Thomas Paine. It was a liberal and humanitarian nationalism that regarded America as in the vanguard of mankind on its march to greater liberty, equality, and happiness for all. The ideas of the 18th century found their first political realization in the Declaration of Independence and in the birth of the American nation. Their deep influence was felt in the French Revolution.

Jean-Jacques Rousseau had prepared the soil for the growth of French nationalism by his stress on popular sovereignty and the general cooperation of all in forming the national will, and also by his regard for the common people as the true depository of civilization.

The nationalism of the French Revolution was more than that: it was the triumphant expression of a rational faith in common humanity and liberal progress. The famous slogan "liberty, equality, fraternity" and the Declaration of the Rights of Man and of the Citizen were thought valid not only for the French people but for all peoples. Individual liberty, human equality, fraternity of all peoples: these were the common cornerstones of all liberal and democratic nationalism. Under their inspiration new rituals were developed that partly took the place of the old religious feast days, rites and ceremonies: festivals and flags, music and poetry, national holidays and patriotic sermons. In the most varied forms, nationalism permeated all manifestations of life. Like the rise of American nationalism, the rise of the French produced a new phenomenon in the art of warfare: the nation in arms. In America and in France, citizen armies, untrained but filled with a new fervour, proved superior to highly trained professional armies that fought without the incentive of nationalism. The revolutionary French nationalism stressed free individual decision—in the formation of nations. Nations were constituted by an act of self-determination of their members. The plebiscite became the instrument whereby the will of the nation was expressed. In America as well as in revolutionary France, nationalism meant the adherence to a universal progressive idea, looking toward a common future of freedom and equality, not toward a past characterized by authoritarianism and inequality.

Napoleon's armies spread the spirit of nationalism throughout Europe and even into the Near East; while at the same time, across the Atlantic, it aroused the Latin Americans. But Napoleon's yoke of conquest turned the nationalism of the Europeans against France. In Germany the struggle was led by writers and intellectuals, who rejected all the principles upon which the American and the French revolutions had been based as well as the liberal and humanitarian aspects of nationalism.

German nationalism began to stress instinct against reason; the power of historical tradition against rational attempts at progress and a more just order; the historical differences between nations rather than their common aspirations. The French Revolution, liberalism, and equality were regarded as a brief aberration, against which the eternal foundations of societal order would prevail.

That German interpretation was shown to be false by the developments of the 19th century. Liberal nationalism reasserted itself; it permeated more and more peoples: the rising middle class and the new proletariat. The revolutionary wave of 1848, the year of "the spring of the peoples," seemed to realize the hopes of nationalists such as Giuseppe Mazzini, who had devoted his life to the unification of the Italian nation by democratic means and to the brotherhood of all free nations. Though his immediate hopes were disappointed, the 12 years from 1859 to 1871 brought the unification of Italy and Romania, both with the help of Napoleon III, and of Germany; at the same time the 1860s saw great progress in liberalism, even in Russia and Spain. The victorious trend of liberal nationalism, however, was reversed in Germany by Bismarck. He unified Germany on a conservative and authoritarian basis and defeated German liberalism. The German annexation of Alsace-Lorraine against the will of the inhabitants was contrary to the idea of nationalism as based upon the free will of man. The people of Alsace-Lorraine were held to be German by objective factors, by race, independent of their will or of their allegiance to any nationality of their choice.

The 1848 revolutionary wave

In the second half of the 19th century, nationalism disintegrated the supranational states of the Habsburgs and the Ottoman sultans, both of which were based upon prenational loyalties. In Russia, the penetration of nationalism produced two opposing schools of thought. Some nationalists proposed a westernized Russia, associated with the progressive, liberal forces of the rest of Europe. Others stressed the distinctive character of Russia and Russianism, its independent and different destiny based upon its autocratic and orthodox past. These Slavophiles, similar to and influenced by German romantic thinkers, saw Russia as a future saviour of a West undermined by liberalism and the heritage of the American and French revolutions.

Central and eastern Europe

One of the consequences of World War I was the triumph of nationalism in central and eastern Europe. From the ruins of the Habsburg and Romanov empires, emerged the new nation-states of Austria, Hungary, Czechoslovakia, Poland, Yugoslavia and Romania. Those states in turn, however, were to be strained and ravaged by their own internal nationality conflicts and by nationalistic disputes over territory with their neighbours.

Russian nationalism was in part suppressed after Lenin's victory in 1917, when the Bolsheviks took over the old empire of the Tsars. But the Bolsheviks also claimed the leadership of the world Communist movement, which was to become an instrument of the national policies of the Russians. During World War II Stalin appealed to nationalism and patriotism in rallying the Russians against foreign invaders. After the war he found nationalism one of the strongest obstacles to the expansion of Soviet power in eastern Europe. National communism, as it was called, became a divisive force in the Soviet bloc. In 1948 Tito, the Communist leader of Yugoslavia, was denounced by Moscow as a nationalist and a renegade; nationalism was a strong factor in the rebellious movements in Poland and Hungary in the fall of 1956; and subsequently its influence was also felt in Romania and Czechoslovakia.

ASIAN AND AFRICAN NATIONALISM

Nationalism began to appear in Asia and Africa after World War I. It produced such leaders as Kemal Atatürk in Turkey, Sa'd Pasha Zaghūl in Egypt, Ibn Sa'ūd in the Arabian peninsula, Mahatma Gandhi in India, and Sun Yat-sen in China. Atatürk succeeded in replacing the medieval structure of the Islāmic monarchy with a revitalized and modernized secular republic in 1923. Demands for Arab unity were frustrated in Africa and Asia by British imperialism and in Africa by French imperialism. Yet Britain may have showed a gift for accommodation with the new forces by helping to create an independent Egypt (1922; completely, 1936) and Iraq (1932), and displayed a similar spirit in India, where the Indian National Congress, founded in 1885 to promote a liberal nationalism inspired by the British model, became more radical after 1918. Japan, influenced by Germany, used modern industrial techniques in the service of a more authoritarian nationalism.

The new nations

The progress of nationalism in Asia and Africa is reflected in the histories of the League of Nations after World War I and of the United Nations after World War II. The Treaty of Versailles, which provided for the constitution of the League of Nations, also reduced the empires of the defeated Central Powers, mainly Germany and Turkey. The league distributed Germany's African colonies as mandates to Great Britain, France, Belgium, and South Africa, and its Pacific possessions to Japan, Australia, and New Zealand under various classifications according to their expectations of achieving independence. Among the League's original members, there were only five Asian countries (China, India, Japan, Thailand,

and Iran) and two African countries (Liberia and the Union of South Africa), and it added only three Asian countries (Afghanistan, Iraq, and Turkey) and two African countries (Egypt and Ethiopia) before it was dissolved in 1946. Of the mandated territories under the League's control, only Iraq, Lebanon, and Syria achieved independence during its lifetime.

Of the original 51 members of the United Nations in 1945, eight were Asian (China, India, Iraq, Iran, Lebanon, Saudi Arabia, Syria, and Turkey) and four were African (the same as in the League). But less than a quarter of a century after its founding, the United Nations had added 66 member nations, 55 of them Asian and African. Whereas Asian and African nations had never totaled one-third of the membership in the League, they came to represent nearly half of the United Nations. Of these new Asian and African nations, four in Asia and seven in Africa were in part or in whole created from mandated territories.

After World War II, India, Pakistan, Ceylon, Burma, and Malaya in Asia, and Ghana in Africa achieved independence peacefully from the British Commonwealth, as did the Philippines from the United States. Others had to fight hard for their independence in bitter colonial wars, as in French Indochina (Vietnam, Laos, Cambodia) and French North Africa (Tunisia, Algeria). Communism had sought to recruit supporters from within the ranks of the new nationalist movements in Asia and Africa, first, by helping them in their struggle against Western capitalist powers, and later after independence was achieved, by competing with Western capitalism in extending financial and technical aid. Chinese nationalism under Chiang Kai-shek during World War II was diminished with the take over of the Chinese Communists. But Chinese communism soon began to drift away from supranational communism, as the European Communist countries had earlier. By the late 1960s, Russian and Chinese mutual recriminations revealed a Chinese nationalism in which Mao Tse-tung had risen to share the place of honour with Lenin. As Chinese communism turned further and further inward, its influence on newborn Asian and African nations apparently waned.

Political and religious differences

Ambitions among new Asian and African nations clashed. The complex politics of the United Nations illustrated the problems of the new nationalism. The struggle with Dutch colonialism that brought the establishment of Indonesia continued with the UN mediating the dispute over West Irian (West New Guinea) well into the late 1960s. In the Suez crisis of 1956, UN forces intervened between those of Egypt and Israel. Continuing troubles in the Middle East beginning with the establishment of Israel and including inter-Arab state disputes brought on by the establishment of the United Arab Republic, concerned the UN. Other crises involving the UN included: the India-Pakistan dispute over Jammu and Kashmir; the Korean partition and subsequent war; the four-year intervention in the Congo; the struggle of Greece and Turkey over newly independent Cyprus; Indonesian and Philippine objection to the inclusion of Sarawak and Sabah (North Borneo) in the newly formed Malaysia.

Many new nations, all sharing the same pride in independence, faced difficulties. As a result of inadequate preparation for self-rule, the first five years of independence in the Peoples' Republic of the Congo passed with no semblance of a stable government. The problem of varying peoples and languages was exemplified in the Federal Republic of Nigeria, where an uncounted population including an uncounted number of tribes (at least 150, with three major divisions) used an uncounted number of languages that included at least a hundred language and dialect clusters. Religious differences kept alive the question of whether the predominantly Muslim state of Jammu and Kashmir should go with Muslim Pakistan or Hindu India for more than 20 years after the India Independence Act became effective in 1949. Desperate economic competition caused trouble, as in Israel where the needed life-giving water of the Jordan River kept it in constant dispute with its water hungry Arab neighbours.

In Europe the spirit of nationalism seemed to have waned after World War II with the establishment of international economic, military and political organizations such as NATO, the European Coal and Steel Community, Euratom, the Common Market, and others. But the policies pursued by France under Pres. Charles de Gaulle, and the problem of a divided Germany, showed that the appeal of the nation-state was still very much alive.

BIBLIOGRAPHY. The origins and development of nationalism as a political idea are discussed in detail in HANS KOHN, *The Idea of Nationalism* (1944, reprinted 1967). The various interpretations of nationalism in modern times will be found analyzed in LOUIS L. SNYDER, *The Meaning of Nationalism* (1954). The most important attempts at solving the nationality problem within a supranational framework were made in the 19th century in Austria; on this topic, see ROBERT A. KANN, *The Multinational Empire: Nationalism and National Reform in the Habsburg Monarchy, 1848–1918*, 2 vol. (1950). A new approach to the problem of nationalism was opened up by KARL W. DEUTSCH, *Nationalism and Social Communication* (1953). On nationalism in the Third World, see RUPERT EMERSON, *From Empire to Nation: The Rise to Self-Assertion of Asian and African Peoples* (1960). There are two valuable bibliographies: KOPPEL S. PINSON, *A Bibliographical Introduction to Nationalism* (1935); and KARL W. DEUTSCH, *Interdisciplinary Bibliography on Nationalism* (1956).

(H.K./Ed.)

Native Elements

Excluding the free gases of the atmosphere, there are 20 elements that occur in nature in pure or nearly pure form; these are called the native elements. Of the 20 native elements, only eight have historic or major commercial significance. The remaining elements are not abundant, and a few are considered to be mineralogical curiosities.

Although most elements in native form are rare, from ancient times several have been known, used, and cherished by man. It is generally agreed that the first metals known to man were those found in the native state. The earliest of these probably were nuggets of gold from alluvial sand and gravel deposits. Too soft for tools or weapons, this gold was used in jewelry or other decorative objects; native silver was probably used in the same way. Native copper, though soft, hardens when it is hammered and could have been used for sickles and daggers capable of cutting corn or flesh. Iron and nickel-iron, although less widely distributed, were known and were fashioned into implements by early man. The nonmetals, carbon (both as graphite and diamond) and sulfur also have been known throughout historic time.

Native gold and silver are important commercially as principal ores of these metals. Native copper is used as an ore but is not of great commercial importance. Native iron and nickel-iron are relatively rare and are of no commercial importance. Platinum in the native state, on the other hand, is practically the exclusive source of this precious metal. Platinum was unknown to the ancients, and man's first knowledge of it dates from its discovery in 1735, in the form of nuggets found in a Colombia stream bed. Early in the 20th century graphite was synthesized, but the natural product remains an important source of the substance. Diamonds have been man-made, but synthetic stones are small and are used only for industrial purposes. Not until the mid-20th century have other sources of sulfur challenged the dominant position of that native element as a basic raw material of the chemical industry.

CLASSIFICATION AND PROPERTIES

Native elements are classified as metals, semimetals, or nonmetals. In the mineralogical classification below, naturally occurring solid solutions or compounds of two or more metals are included with the metals.

Metals. The native metals of mineralogical significance are in the gold group, the platinum group, and the iron group; the others are of only minor importance.

Gold, silver, and copper

The gold group comprises three important metals; namely, gold, silver, and copper. These elements fall in the same family in the periodic classification (classification by atomic weight and number) of elements and, hence, have somewhat similar crystallographic, chemical, and physi-

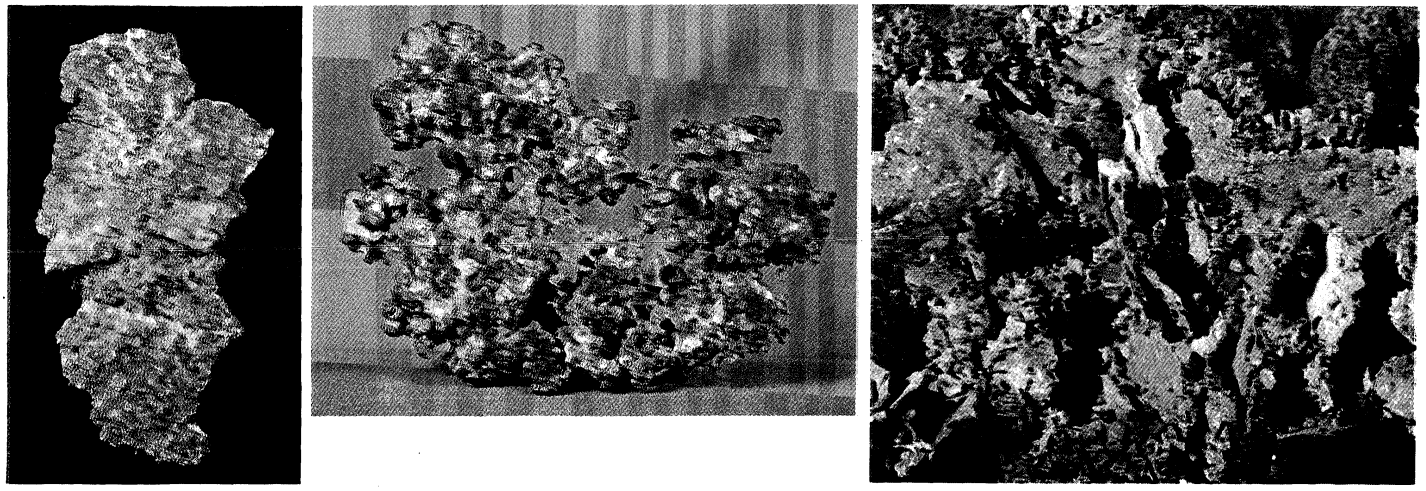


Figure 1: (Left) Dendritic silver from Ontario. (Centre) Crystalline copper from Michigan. (Right) Gold, showing leaf formation from the Mother Lode, Nevada County, Calif.
By courtesy of (left) Joseph and Helen Guetterman collection, (centre) Ted Boente collection, (right) Demuth collection; photographs, (left, centre) John H. Gerard, (right) Mary A. Root—EB Inc.

Table 1: Native Elements

metals	
Gold group	Gold, Au Electrum (Au, Ag) Silver, Ag Copper, Cu
Mercury group	Mercury, Hg Moschellandsbergite, Ag_2Hg_3 Gold-amalgam, Au_2Hg_3 Potarite, Pd_3Hg_2 Zinc, Zn
Platinum group	Platinum, Pt Palladium, Pd Platiniridium (Pt, Ir) Aurosmiridium (Ir, Os, Au) Iridosmine (Ir, Os) Siserskite (Os, Ir)
Iron group	Iron, Fe Nickel-iron (Ni, Fe)
Tin group	Tin, Sn Lead, Pb Tantalum, Ta
nonmetals	
Sulfur group	Sulfur, S
Carbon group	Diamond, C Graphite, C
Tellurium group	Selenium, Se Tellurium, Te
semimetals	
Arsenic group	Arsenic, As Antimony, Sb Bismuth, Bi

cal properties. They are alike in crystal structure: the atoms, held together by metallic bonds, are arranged in a configuration known as cubic closest packing, in which array each atom lies on a point of a face-centred cubic lattice (cubic arrangement with atoms in centres of faces) and is coordinated with 12 others. The unit of structure is a cubic unit cell containing four atoms. The three metals crystallize in the isometric system, hexoctahedral class, space group Fm3m (see MINERALS; CRYSTALLOGRAPHY). The length of the edge of the unit cell is dependent on the radius of the atom and differs for each metal. The cell dimensions and some of the physical constants of the pure elements are given in Table 2.

The common structure of these metals gives rise to similar properties. All are soft, malleable, and ductile (can be drawn into wire). This plastic deformation is the result of four possible directions of translation of the atoms in the structures. Of all known elements, these metals are the best conductors of heat and electricity in the order of silver first, followed by copper and gold.

Table 2: Properties of the Gold Group Metals
(Angstrom, Å = 10^{-8} cm)

property	gold	silver	copper
Atomic radius	1.44 Å	1.44 Å	1.28 Å
Cell dimensions	4.0786 Å	4.0862 Å	3.6150 Å
Mohs hardness	2½–3	2½–3	2½–3
Specific gravity	19.32	10.50	8.96
Melting point	1,063° C	961° C	1,083° C

The differences in properties of the metals of this group arise from individual properties of the atoms. The yellow of gold, the white of silver, and the red of copper are a manifestation of atomic properties, as are the specific gravities.

The solubility of the gold group metals in the solid state is largely dependent on atomic size. Gold and silver have essentially the same atomic radii and can substitute for one another in all proportions (to 100 percent). Most gold contains some silver and its rich yellow colour becomes progressively lighter as more silver enters the structure. When silver is present in amounts 20 percent or more, the natural alloy is called electrum. Much gold also may be present in native silver. Although traces of copper occur in gold and silver and vice versa, there is only limited substitution because the radius of the copper atom is significantly smaller. In addition to the impure nature of these metals from mutual substitution, other elements may be present in trace amounts. Among the most common are iron in gold; mercury in silver; and arsenic, iron, or bismuth in copper.

The metals of the gold group most commonly occur in nature as irregular masses, plates, and scales. All have been found, however, in beautifully crystallized arborescent groups, frequently with dendritic (branching or tree-like) shapes. Single crystals, particularly of gold and silver, are found as plates flattened parallel to octahedron faces or as wires elongated perpendicular to such a plane. Octahedral crystals of gold with equal-area faces are rare. Crystal groups of copper may show faces of the cube, dodecahedron, and octahedron.

Mercury and zinc

The metals of the mercury group are rare. Above −39° C, mercury is a silvery-white liquid with a density of 13.6; below this temperature it crystallizes in the form of rhombohedral crystals. Natural alloys with other metals include moschellandsbergite, Ag_2Hg_3; gold-amalgam, Au_2Hg_3; and potarite, Pd_3Hg_2. These minerals, extremely rare and poorly defined, are believed to be isometric in crystallization.

The properties of native zinc are those for man-made

Table 3: Properties of the Platinum Group Metals
(Angstrom, Å = 10^{-8} cm)

properties	ruthenium	rhodium	palladium	osmium	iridium	platinum
Crystal system	hex	iso	iso	hex	iso	iso
Atomic radius	1.32 Å	1.34 Å	1.37 Å	1.34 Å	1.35 Å	1.38 Å
Cell dimensions	a=2.706 Å c=4.282 Å	3.803 Å	a=3.882 Å	a=2.714 Å c=4.314 Å	a=3.831 Å	a=3.916 Å
Mohs hardness	6.5		4.8	7.0	6.5	4.3
Specific gravity	12.41	12.41	12.02	22.57	22.42	21.45
Melting point	2,250° C	1,960° C	1,552° C	3,000° C	2,443° C	1,769° C

zinc. Its structure is hexagonal closest packing, with unit cell dimensions: $a = 2.66$Å and $c = 4.94$Å. It is a bluish-white metal with hardness of 2 and specific gravity of 7.14. Although native zinc has been reported from Australia, New Zealand, and the United States, no occurrence is well authenticated.

Platinum group metals

Of the six platinum group metals, only platinum and palladium occur in sufficient purity to be considered native elements. Platinum is the only common element in the group and the other metals generally are found alloyed with it. The atoms of both platinum and palladium are arranged in cubic closest packing, and these metals have structures that are analogous to those of the gold group. The cell dimensions and physical properties of the pure platinum group elements are given in Table 3.

Chemical analyses rarely report more than 90 percent purity for native platinum. The remainder is made up of varying percentages of the other platinum group metals plus gold, copper, and nickel. The principal impurity, however, is iron, which in some analyses exceeds 10 percent and makes much native platinum magnetic. The specific gravity of pure platinum is reduced by impurities to a range of 14 to 19. The colour is silver-white to steel-gray. Platinum is malleable and ductile but less so than the metals of the gold group. Although distorted cubes and octahedrons of platinum have been reported, crystals are rare.

Natural alloys of the platinum group metals are found associated with native platinum but are very rare. Platiniridium (Ir,Pt) and aurosmiridium (Ir,Os,Au) are silver-white minerals with isometric crystallization and high hardness (6–7); their specific gravity, also high, varies with composition. Iridosmine (Ir,Os) and siserskite (Os,Ir) have hexagonal crystallization; they form a solid solution series (metal–metal solution) containing from about 77 percent Ir to 80 percent Os and sometimes ruthenium, rhodium, or platinum.

Iron and nickel-iron

Iron with 2 to 7 percent nickel, and nickel-iron with 30 to 75 percent nickel are found in terrestrial deposits and in meteorites. Ordinary iron (α iron) has a body-centred cubic structure (cubic arrangement of atoms with one atom at centre of cube). The unit cell with dimensions $a = 2.86$Å, contains 2 (Fe,Ni). Iron is malleable and has cubic cleavage (tends to cleave along planes at right angles), a Mohs hardness of 4, and a specific gravity of 7.3 to 7.87. Crystals are rare, and terrestrial material is in small to large regular masses. In meteorites known as kamacite it occurs in platy masses and in irregular intergrowths with nickel-iron.

Nickel-iron, like the metals of the gold group, has a face-centred cubic lattice. Its unit cell ($a = 3.59$Å) contains 4 (Fe,Ni). Nickel-iron's colour is silver-white to grayish-white, its hardness is 5, and its specific gravity 7.8 to 8.2; it is strongly magnetic. Terrestrial nickel-iron occurs in small grains and fine scales; in meteorites it is called taenite and it forms regular intergrowths with kamacite.

Tin, lead, and tantalum

Tin is very rare as a native metal. It is tetragonal in crystallization, and its unit cell ($a = 5.82$Å, $c = 3.18$Å) contains 4 Sn. Tin is malleable and ductile, its colour is tin-white, its hardness is 2, and its specific gravity is 7.3. It has been found in small, rounded grains associated with gold, platinum, and cassiterite in sands from New South Wales, Australia.

Lead is very rare as the native metal and authenticated occurrences are few, but large crystal groups have been found at Långban, Sweden. Native lead crystallizes in the isometric crystal system and has a face-centred cubic lattice; its crystals are octahedral with subordinate dodecahedron and cube faces. The unit cell ($a = 4.94$Å) contains 4 Pb. Native lead is relatively pure, usually containing over 99 percent Pb; small amounts of silver and antimony may be present. It is quite malleable and has a lead-gray colour, a hardness of 1½, and a specific gravity of 11.37.

Native tantalum is isometric and crystallizes in minute cubic crystals. It has a body-centred cubic lattice with the edge of the unit cell $a = 3.30$Å. Its colour is grayish-yellow, its Mohs hardness is 6 to 7, and its specific gravity is 11.2 (measured), 16.6 (calculated). Tantalum is the rarest of the native metals. It reportedly has been found in placer deposits (stream-laid gravels) in the Ural and Altai mountains, U.S.S.R., but the latter occurrence has been questioned.

Arsenic, antimony, and bismuth

Semimetals. The native semimetals (arsenic, antimony, bismuth) fall in the same family in the periodic table of the elements and thus have many similar chemical and physical properties. They crystallize in the hexagonal system, scalenohedral class, and have one direction of good cleavage; they are brittle and are poorer conductors of heat and electricity than the native metals. These properties result from a bonding mechanism intermediate to true metallic and covalent bonding (see MINERALS). Each atom is bonded to six neighbours but is more closely held to three of these than to the other three. The resulting structure has sheets of atoms that are strongly held to one adjacent sheet but weakly bonded to the other. Cleavage takes place by breaking the weak bond between sets of double sheets. Cell dimensions and physical properties for the native semimetals are given in Table 4.

Table 4: Properties of the Arsenic Group Native Elements
(Angstrom, Å = 10^{-8} cm)

property	arsenic	antimony	bismuth
Atomic radius	1.25 Å	1.45 Å	1.55 Å
Cell dimensions	a=3.76 Å c=10.55 Å	a=4.31 Å c=11.27 Å	a=4.55 Å c=11.86 Å
Mohs hardness	3.5	3–3.5	2–2.5
Specific gravity	5.7	6.7	9.7–9.8
Melting point	—	630.5° C	271.3° C

Native arsenic and antimony occur most commonly in tin-white, massive, botryoidal (globular) or lamellar (platelike) aggregates. They rarely are found in well-formed pseudocubic crystals. Native bismuth occurs usually in silver-white reticulated or arborescent shapes or in veinlets showing well-defined cleavage. These elements occur in a relatively pure state in nature, but each of them may contain traces of the other two. In addition, arsenic may contain iron, nickel, silver, and sulfur; antimony may contain iron and silver; and bismuth may contain tellurium and sulfur.

Native sulfur

Nonmetals. Sulfur has three modifications (forms), designated α, and β, and γ. Although β and γ sulfur (monoclinic forms) have been reported from a fumarole (gaseous emanation) at Vulcano, Lipari Islands, they are extremely rare in nature. Sulfur as a native element essentially occurs in the α modification. It crystallizes in the orthorhombic system, dipyramidal class, with space group Fddd (see CRYSTALLOGRAPHY). The unit cell contains 128 atoms and has the dimensions: $a = 10.44$Å, $b = 12.84$Å, $c = 24.37$Å.

Figure 2: (Left) Bismuth crystals from Utah. (Centre) Sulfur crystals (greatly enlarged) from Sicily. (Right) Iron meteorite found at Springwater, Sask. in 1931.
By courtesy of the Illinois State Museum; photographs, John H. Gerard—EB Inc.

The yellow colour of sulfur is characteristic. The fracture is conchoidal, but imperfect cleavages parallel to {001}, {110}, and {111} are sometimes observed. The Mohs hardness is 1½ to 2½, the specific gravity 2.07, and the melting point 112.8° C (235° F). Sulfur is an electrical nonconductor and an extremely poor thermal conductor.

Native sulfur commonly contains small amounts of clay, bitumen, and other mechanically held impurities. Some sulfur contains selenium, rarely more than one percent, which renders it orange-red or reddish-brown. Most commonly, native sulfur is massive or in spherical and reniform (rounded, kidney-like mass) shapes, or it occurs as incrustations. In some places, notably at Girgenti, Sicily, it is found in beautiful, symmetrically developed crystals that usually display a pyramidal habit.

Carbon Group. Native carbon has only two forms, diamond and graphite, and the only similarity between them is that both are composed of carbon atoms. No other polymorphic minerals present a greater contrast in physical properties: diamond is isometric, transparent, and the hardest of natural substances, whereas graphite is hexagonal, opaque, and one of the softest minerals. These differences arise from the manner in which the atoms are bonded together in the crystal structures.

Diamond

Diamond is the high-temperature, high-pressure form of carbon that crystallizes in the isometric system. Although the crystal class has been in doubt, most crystallographers assign it to the hexoctahedral class, space group Fd3m. The unit cell contains eight carbon atoms, the length of the cell edge is $a = 3.567$Å. In the diamond structure each carbon atom is linked equidistantly to four neighbouring carbon atoms by covalent bonds so that a tetrahedral linkage occurs throughout. This close-knit, strongly bonded structure gives rise to the high Mohs hardness as well as the high (for a nonmetallic mineral) specific gravity.

The scratch hardness of diamond is usually given as 10 on the Mohs' scale, with the next hardest mineral being corundum at 9. Actually, diamond is very much harder than corundum; its true proportional value on the Mohs' scale would be about 42. Diamond's hardness varies significantly in different directions; this makes some facets of a cut diamond easier to polish than others. Diamond has been found to be as much as 200 times more effective in abrasion than the synthetic abrasives silicon carbide (Mohs hardness 13) and boron carbide (Mohs hardness 14).

Diamond has a perfect cleavage (referred to as grain by diamond cutters) in four directions parallel to its octahedral crystal faces. This property is used by diamond cutters to split a big stone. Diamond has a specific gravity of 3.51, an adamantine to greasy lustre, and a high dispersion. Its index of refraction (change in direction of propagation of entering light) under sodium light is 2.4175. Diamonds have many colours. Although as a gem it is usually colourless to blue-white, more commonly it is yellow or brown; rarely, it is orange, pink, green, blue, red, or black.

Two types of diamond are known, termed Type I and Type II. Type II diamond shows a normal X-ray diffraction pattern, but patterns of Type I show extra spots and streaks. The photoconductivity of Type I diamond is poor, of Type II good; Type I diamonds are birefringent (show double refraction), and Type II are not. The two types also differ in their infrared and ultraviolet absorption characteristics. Diamonds intermediate between the two types have been noted, as well as two varieties of Type II. Large diamonds of gem quality have always been Type I.

Diamond characteristically is found in crystals, many well formed. They are predominantly octahedral, less commonly dodecahedral or cubic. Some crystals are flattened parallel to octahedron faces; others have a spherical form with radiated structure. Twinning on the octahedron is common with the twinned pair flattened parallel to the twin plane.

Diamonds were synthesized in 1955 by the General Electric Company, Schenectady, New York. They were made by subjecting carbonaceous material simultaneously to pressures over 1,500,000 pounds per square inch and temperatures over 5,000° F (2,800° C), thus simulating pressure-temperature conditions 240 miles deep within the earth. The diamonds made were of industrial quality and by 1960 were being marketed on a large scale at prices comparable to natural stones of the same quality. Continued experimentation produced (1967) a hexagonal polymorph with properties similar to those of normal (isometric) diamond. Almost simultaneously, small crystals of the hexagonal diamond were found in a meteorite and named lonsdaleite.

Bort is granular microcrystalline diamond, usually without cleavage and coloured gray to black by inclusions of other elements in the structure. The term also applies to poorly coloured or flawed diamond without gem value. Specimens of bort may contain as much as 20 percent mechanically held impurities; spectroscopic analyses reveal traces of barium, strontium, and chromium. Carbonado or carbon is black, massive diamond without cleavage and is less brittle than crystals.

Graphite

Graphite crystallizes in the dihexagonal-dipyramidal class of the hexagonal system; space group C6/mmc. It has a layered structure in which strongly bonded carbon atoms are arranged in six-membered rings in individual sheets. Adjacent sheets are widely spaced and are held together by the weak van der Waals bond. This arrangement results in graphite's characteristics of low hardness, low specific gravity, and good cleavage. The unit cell contains four carbon atoms and has the dimensions $a = 2.461$Å, $c = 6.708$Å.

The physical properties of graphite are distinctive: one perfect cleavage; hardness, 1 to 2 (it soils the fingers); specific gravity, 2.2; lustre, metallic; colour, black. It is often impure due to admixed clay and other minerals. Graphite crystals are rare but when found are usually six-sided tablets with triangular marking. The mineral most commonly occurs as foliated masses and as isolated scales.

The tellurium group comprises the rare native elements tellurium and selenium, which crystallize in the hexagonal

Tellurium and selenium

system, trigonal trapezohedral class, space group $P3_121$ or $P3_221$. Cell dimensions and physical properties of these native elements are given in Table 5.

Table 5: Properties of the Tellurium Group Elements
(Angstrom, Å = 10^{-8} cm)

property	tellurium	selenium
Cell dimensions	a=4.46 Å c=5.93 Å	a=4.36 Å c=4.96 Å
Mohs hardness	2–2.5	2
Specific gravity	6.1–6.3	4.8
Melting point	449.5° C	217° C

Tellurium has a metallic lustre, a tin-white colour, and a perfect cleavage parallel to $\{10\bar{1}1\}$. Selenium has a metallic lustre, gray colour, and a good cleavage parallel to $\{01\bar{1}2\}$. Both elements may contain small amounts of the other in the native form. In addition, tellurium sometimes contains admixed gold, silver, and iron.

Crystals of both are usually acicular (needlelike) elongated on the c crystal axis. Tellurium is commonly massive or granular; selenium may occur in sheets or glassy droplets. The formation of selenium as a native element has resulted chiefly from mine fires that produce artificial fumarolic conditions. Tellurium, on the other hand, is found in hydrothermal veins associated with sulfide and telluride minerals as at Kalgoorlie, Western Australia; Balia, Asia Minor; Hokkaido, Japan; and Cripple Creek, Colorado.

OCCURRENCE OF THE NATIVE ELEMENTS

One can make no generalization as to the occurrence of the native elements. They form under greatly contrasting physicochemical conditions and in all types of rocks. Even a single native element can occur in widely diverse environments.

Metals. *Gold.* Most gold is found in the earth's crust in the native state. Although a rare element, it occurs in two types of widely distributed deposits: hydrothermal veins (deposits produced by the reaction of high-temperature solutions and the invaded rocks) and placers. As a primary hydrothermal mineral, gold is found in all types of rocks of all geologic ages, but most commonly it occurs in quartz veins associated with pyrite and other sulfide minerals. Gold is recovered as a valuable by-product in some vein deposits mined essentially for other metals.

Gold placer deposits

Gold is liberated by weathering from the enclosing matrix of vein deposits and, if washed into neighbouring streams, accumulates to form placer deposits. Throughout history it has been free gold, located in the unconsolidated gravels of a placer, that has led to the discovery of major gold deposits. By tracing the gold upstream to its source, prospectors have located the vein deposits from which it came. Ancient placers, former stream gravel deposits that have been consolidated into conglomerates, are actively mined today. More than 60 percent of the world's gold is derived from such "fossil placers" in the Transvaal and Orange Free State, Republic of South Africa. The U.S.S.R. is believed to be the second-ranking country in gold production. Some gold is recovered from vein deposits in the Ural Mountains, but placers in Siberia are reported to be the major source. Elsewhere, gold is won chiefly from vein deposits. Canada ranks third in production, followed by the United States, Ghana, Australia, and the Philippines.

Silver. Unlike gold, native silver is not the principal source of the metal, but it is widely distributed in small amounts in the upper portions of ore deposits, where it has formed by oxidation of other silver minerals. Native silver has been mined in larger deposits formed from hydrothermal solutions. In the primary deposits it has three types of association: with sulfide minerals (*q.v.*), zeolites, calcite, barite, and quartz (such a deposit was worked for several hundred years at Kongsberg, Norway and produced many excellent, crystallized specimens of silver); with arsenides and sulfides of cobalt, nickel, and silver, and with native bismuth (the deposits at Freiberg and Schneeberg in Saxony, actively mined in the 16th century, and at Cobalt, Ontario, discovered early in the 20th century, were of this type); and with uraninite and cobalt-nickel minerals (such are the old deposits at Jáchymov, Bohemia, and those more recently discovered at Great Bear Lake, Northwest Territories, Canada).

Copper. Small amounts of native copper have been found in many localities in the shallow oxidized zones overlying copper sulfide deposits. Such occurrences of native copper are commercially unimportant because most of the metal is derived from natural compounds, particularly chalcopyrite, $CuFeS_2$. At a number of localities native copper is a primary mineral associated with basaltic lavas; the reaction of hydrothermal solutions with iron minerals results in the deposition of native copper rather than the normal copper sulfides. The only important deposit of this type is on the Keweenaw peninsula, Michigan; but for 70 years, commencing in 1845, this district was a leading world copper producer. Here, dipping lavas interbedded with conglomerates are exposed for 100 miles along the peninsula. The copper occurs principally as fillings in cavities in the lavas and as replacement for pebbles or filling space between them in the conglomerates.

Platinum. Native platinum, palladium, and natural alloys of platinum group metals usually occur together, in association with olivine, chromite, pyroxene, and magnetite in ultrabasic rocks (those with high percentages of iron and magnesium) (especially dunites, which are nearly pure olivine rocks). Because platinum is chemically inert and has a high specific gravity, it tends to collect in placers. It was found in Colombia in 1735 and in the Ural Mountains in 1822 in such deposits. Placers were the chief source of platinum until 1934, when large amounts were recovered from sperrylite, $PtAs_2$, a mineral associated with copper-nickel ores at Sudbury, Ontario. In 1924, platinum was discovered in the "Merensky Reef" of the Bushveld Igneous Complex, Republic of South Africa. By 1953, mining of this narrow but extensive layer of ultrabasic rock placed South Africa first in world platinum production. More recently, the U.S.S.R. has assumed first place; the recovery of platinum from placer deposits and more particularly from copper-nickel ores enabled that country to account for about 60 percent of the world production of platinum metals in 1969.

Iron and nickel-iron. Although terrestrial native iron is a great rarity, it has been found in igneous rocks (basalts), in carbonaceous sedimentary rocks, and in petrified wood. Some iron in basalts is considered to be a primary or original constituent, but most of it is believed to have formed from iron compounds by the reducing action of later incorporated carbonaceous material. Likewise, in sedimentary rocks and petrified wood, native iron has resulted from the reduction of iron compounds by organic material. Small amounts of native iron have been reported from Ireland, France, Poland, U.S.S.R., and in the United States from Cameron, Missouri, and New Brunswick, New Jersey. The most important occurrence is on Disko Island, Greenland, where it is found in a basalt in the form of small grains and as large masses, some weighing 20 tons.

Terrestrial nickel-iron is less common than native iron. It has been found sparingly as small plates and grains with gold in placer deposits of New Zealand, U.S.S.R., and Canada. The largest masses, some exceeding 100 pounds, have been found in Josephine and Jackson counties, Oregon.

Both native iron (kamacite) and nickel-iron (taenite) are found in meteorites. The iron meteorites called hexahedrites are almost completely composed of kamacite, and in those called octahedrites it is the principal constituent.

Semimetals. The native semimetals arsenic, antimony, and bismuth are found in hydrothermal veins formed under similar conditions. Consequently, all three elements may be found at the same locality with native silver and ore minerals containing arsenic and antimony. Common associated minerals are sphalerite, pyrite, galena, and quartz. In addition, native arsenic and native bismuth frequently accompany ores of nickel and cobalt. In this

association they are found in the old Saxon mining districts of Schneeberg and Annaberg, and in Bohemia at Jáchymov. Other noted localities for arsenic are Romania, Italy, Borneo, and Chile. The most productive deposits of native bismuth are in Bolivia; other important localities are Australia, where it is found in pegmatites (intrusive igneous rocks with exceptionally large crystals), and at Cobalt, Ontario, where it is associated with silver ores. Fine specimens of native antimony have come from France, Borneo, Australia, and Chile.

Nonmetals. *Native sulfur.* Native sulfur occurs at the crater rims of active or extinct volcanoes, where it forms from gases given off by fumaroles either as a sublimation product or by the incomplete oxidation of hydrogen sulfide. Less commonly, sulfur is found in veins associated with sulfide ore minerals where it is formed by their oxidation.

The largest deposits are found in sedimentary rocks associated with sulfate minerals (anhydrite and gypsum) and with limestone. In this type of deposit the sulfur is believed to have been derived from the sulfates by the action of sulfur-forming bacteria. Sulfur associated with gypsum and anhydrite near Agrigento, Sicily, has been mined since 1250, and for several hundred years this locality was the principal source of the element. Sulfur, associated with the same minerals, occurs several hundred feet below the surface in the cap rock of salt domes (*q.v.*) in Texas, Louisiana, and elsewhere. Production from such deposits by the Frasch process began in 1894. In this method, superheated water is pumped into the sulfur-bearing rock; the sulfur is melted and brought to the surface in molten form.

Over half of the 1969 world production of sulfur was produced as the native element. Although large tonnages of volcanic sulfur were mined in Chile and similar deposits were worked in Japan and Italy, over 75 percent of the native sulfur in that year was produced by the Frasch process in the United States and Mexico.

Carbon: diamond and graphite. Diamond occurs in two types of deposits: embedded in the rock in which it crystallized and in alluvium and placers as loose crystals washed from their primary source. All early diamonds were alluvial. Stream gravels in India were their only source until about 1725, when diamonds were discovered in similar deposits in Brazil and later, 1866, in the gravels of the Vaal River, South Africa.

Diamonds in South Africa

Only in diamond pipes, especially well exemplified by those discovered in 1871 in the Kimberley district, South Africa, are diamonds known to occur in the original rock matrix in which they formed. These pipes are funnels, more or less oval in section and narrowing with increasing depth. They extend to unknown depths (some have been mined to depths of more than 4,000 feet) and are evidently the result of eruptive action. The rock filling the pipes is an altered and brecciated (broken, angular fragments) basic rock known as kimberlite or blue ground. This must have been forced up from great depths, because it is entirely unrelated to the country rock surrounding the pipes. There is no evidence to indicate that the blue ground was extremely hot when forced into its present position. Therefore, it is concluded that the diamonds were formed in the depths of the earth under conditions of great heat and pressure and reached the surface by extrusion. Diamonds in alluvial deposits originally must have been formed in kimberlite, later released by erosion from their mother rock and redeposited in rivers and streams.

Over 700 kimberlite pipes, dikes, and sills have been located in South Africa; although many are barren, these are the principal source of diamonds mined in that country. The Premier Mine near Pretoria is most productive and since its discovery in 1903 has produced nearly 30 million carats (six tons) of diamonds. In Tanzania, production from a large diamond pipe discovered in 1940 makes that country an important producer. South West Africa is the major producer of gem-quality diamonds. These stones come from terrace deposits along the coast north of the Orange River. Other important African alluvial deposits are in the Democratic Republic of the Congo, Angola, Ghana, and Sierra Leone. More than half of the world's diamonds today come from placers in the Democratic Republic of the Congo, but this large production accounts for only 13 percent of the total value because about 95 percent are industrial grade. Diamonds are produced in the U.S.S.R. from both pipes and placers.

Metamorphic origin of graphite

Graphite is widely distributed, occurring most commonly in metamorphic rocks such as schist, gneiss, and marble, where it may be in disseminated flakes or layers and lenses forming an appreciable proportion of the rock. Such graphite is believed to be the metamorphic product of carbonaceous or bituminous material present in the original sedimentary rocks. In some places, either regional or contact metamorphism has converted coal beds to graphite; the graphitic coal of Rhode Island and Sonora, Mexico, is of this type. Graphite also occurs in veins as typical fracture fillings. Important economic vein deposits are found in Ceylon, Korea, Malagasy Republic, and India. As a minor constituent, graphite is found in some igneous rocks, pegmatites, and iron meteorites.

BIBLIOGRAPHY. B.S. BUTLER and W.S. BURBANK, "The Copper Deposits of Michigan," *Prof. Pap. U.S. Geol. Surv. 144* (1929), an account of the world's greatest deposit of native copper; J.L. GILLSON (ed.), *Industrial Minerals and Rocks*, 3rd ed. (1960), summary articles on the useful minerals dealing with their origin, occurrence, uses, and technology with extensive bibliographies; S.H. HAUGHTON (ed.), *The Geology of Some Ore Deposits in Southern Africa*, 2 vol. (1964), a modern review of the world's most important gold and platinum deposits; C.S. HURLBUT, *Dana's Manual of Mineralogy*, 17th ed. (1959), a textbook giving mineralogical descriptions, occurrences, and uses of native elements; and *Minerals and Man* (1968), a popular account of the discovery, exploitation, and uses of the more abundant native elements; C.A. LAMEY, *Metallic and Industrial Mineral Deposits* (1966), a summary of the world distribution of the useful minerals with extensive bibliography; B.H. MASON, *Meteorites* (1962), a discussion of native iron as found in meteorites; C. PALACHE, H. BERMAN, and C. FRONDEL, *Dana's System of Mineralogy*, 7th ed., vol. 1 (1944), mineralogical descriptions of the native elements; UNITED STATES BUREAU OF MINES, *Minerals Yearbook* (annual), a summary of both U.S. and world production of minerals of economic importance; R.C. WEAST (ed.), *Handbook of Chemistry and Physics*, 46th ed. (1965), brief summaries of the history, occurrence, and use, as well as physical constants of the elements.

(C.S.H.)

Natural Gas

Natural gas is an ideal fuel product and a raw material for the chemical industry. It usually occurs in association with petroleum because geological conditions favourable for the presence of oil generally are favourable for natural-gas occurrence as well. Although many natural gases can be used directly from the well without treatment, some must be processed to remove such undesirable constituents, as carbon dioxide, hydrogen sulfide, and other sulfur components.

Bitumen and asphalt, the oldest known petroleum products, and evidence of natural-gas leaks were first found between 6000 and 2000 BC in Iran. The use of natural gas is mentioned in China about 900 BC; around AD 900–1100 wells were drilled and gas was piped through bamboo tubes.

In Europe these achievements were unknown, and natural gas was not discovered until 1659 in England. Gas distilled from coal was discovered in 1670 and began to be used in 1790 because it was easier to transport, store, and process for use in internal-combustion engines and lighting of streets and houses. In 1821 the town of Fredonia, New York, was lit by natural gas. But natural-gas uses remained localized because there was no way to transport gas over long distances, and for a century natural gas remained on the sidelines of industrial development, which was based on coal, oil, and electricity.

Methods of pipeline transportation developed in the 1920s marked a significant stage in the use of gas. After World War II, there occurred a period of tremendous expansion that has continued until today. In 1950 the worldwide output of natural gas was 200,000,000,000 cubic metres (7,100,000,000,000 cubic feet), and the

United States was the leading producer. In 1960 world output was 470,000,000,000 cubic metres (16,590,000,000,000 cubic feet), and forecasts called for 1,000,000,000,000 in 1970 and for 1,630,000,000,000 in 1980. In 1950 gas accounted for 12 percent of the energy consumed in the world, a rate that increased to 14.6 percent in 1960 and is expected to be around 18.5 percent in 1980.

This article includes treatment of the properties, origin, and occurrence of natural gas and of some aspects of gas production. For additional details of gas technology see GASES, INDUSTRIAL AND DOMESTIC; for treatment of related resources, see PETROLEUM; COALS; and OIL SHALES.

PHYSICAL AND CHEMICAL PROPERTIES

Composition

Natural gas is usually a mixture consisting primarily of methane and ethane, both of which are gaseous under atmospheric conditions. It may also contain such hydrocarbons as propane, butane, pentane, and various liquids, and heavier hydrocarbons, all of which are in liquid form under atmospheric conditions but in gaseous form inside the reservoir; and other gases such as nitrogen, carbon dioxide, hydrogen, hydrogen sulfide, and rare gases. Because gas and water occur together in a reservoir, natural gas also contains water vapour, which is partially condensed during transmission from the reservoir to the processing plant. The composition of gas in selected fields in Algeria, the United States, and The Netherlands is given in Table 1.

Table 1: Characteristics of Selected Gas Fields

	Hassi R'Mel Algeria	Handle-Hugoton Texas	Groningen Netherlands
Depth	2,100 m 6,900 ft	820 m 2,700 ft	2,800 m 9,200 ft
Pressure	309 bars 4,482 psi	34 bars 493 psi	350 bars 5,080 psi
Reserves	2,200 × 10^9 cu m 77,600 × 10^9 cu ft	2,000 × 10^9 cu m 70,600 × 10^9 cu ft	1,600 × 10^9 cu m 56,500 × 10^9 cu ft
Composition of gas (%)			
Nitrogen	5.11	12.8	14.32
Oxygen		2	0.01
Carbon dioxide	0.19	0.5	0.87
Methane	79.6	76.2	81.30
Ethane	7.4	4	2.84
Propane	2.7	2.6	0.43
Butane	5	1.3	0.14
Pentane		0.6	0.08
Hexane			0.02

Behaviour of gas with pressure and temperature. The amounts of gas accumulated in a reservoir and the recoverable reserves, as well as outputs, are expressed in standard volumes (the volume at a temperature of 15° C [59° F] and a pressure of 750 millimetres of mercury). In order to obtain this standard volume from the actual reservoir pressure, temperature, and flow conditions, and to study reservoir behaviour, it is necessary to know the so-called equation of state of the gas (a relation between the pressure and volume and the temperature and gas amount). The simplest way of expressing this relationship is

$$PV = Z\frac{m}{M}RT,$$

in which P is absolute pressure (as read on a pressure meter, plus one atmosphere of pressure); V equals the volume occupied, at pressure P and temperature T, by a mass (m) of the gas under consideration, assuming it to have a molecular mass (M); Z equals a compressibility factor; T equals absolute temperature (°K = °C + 273, or °R = °F + 460); and R equals a universal gas constant.

This equation differs from the law governing the behaviour of perfect gas (a hypothetical gas that obeys the equation PV = (m/M) RT under all of the conditions) by the introduction of the compressibility factor Z, which depends on the composition of the gas and on its temperature and pressure. As Z diminishes, the effect of molecular interactions decreases. As Z reaches a minimum and then approaches 1, the gas tends to behave like a perfect gas. Variation of Z with pressure and temperature for ethane is shown in Figure 1. The value of Z is determined either experimentally or by calculation, using the properties of the gas under consideration.

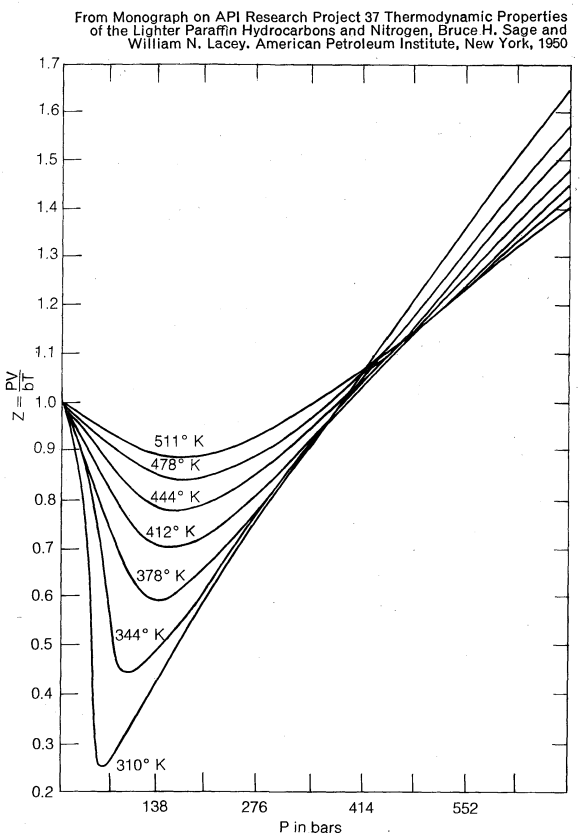


Figure 1: Compressibility factor for ethane (see text).

Condensable products

Some types of natural gas are always in a gaseous state in the reservoir and produce no liquid deposit (other than water) under atmospheric conditions at the surface; these are dry gases. Others, on the contrary, are gaseous in the reservoir and contain condensable products; these are wet gases. Such gases are ordinarily characterized by the volume or weight of the condensables contained in a given standard volume of total gas produced—*i.e.*, including both the gas and the gaseous equivalent of the condensables, expressed either in gallons per thousand cubic feet or in grams per cubic metre. The gas produced at the Hassi R'Mel gas field, Algeria, for example, contains approximately 200 grams of condensables per cubic metre (0.2 ounce per cubic foot). Some recovered condensables are of economic value: propane and butane, for example, are bottled under slight pressure, and heavier hydrocarbons are made into gasoline.

Under some pressure and temperature conditions, four paraffins—methane, ethane, propane, and butane—and carbon dioxide and hydrogen sulfide form solid hydrates in the presence of liquid water. Such hydrates may be formed during the transportation of the gas through boreholes or flow lines, whenever expansion causes cooling. Although this process sometimes results in the plugging up of pipes, there are various prevention methods.

Viscosity is an important property with respect to gas flow. The viscosity of a natural gas varies between 100 and several hundred micropoises, increases with pressure, and decreases when the temperature increases except at low pressures.

Heating power of gas. When natural gas burns completely, carbon dioxide and water are normally formed. The heating power of a gas is measured with calorimeters in which the combustion products are cooled to their initial temperature so that the measurement then indicated corresponds to the amount of heat absorbed by the cooling fluid. Because heat is given off during the condensation of water, the definition "lower heating value" can be used if all of the vapour is assumed not to condense

during cooling, and "gross heating value" if the water vapour is assumed to be condensed. The gas industry mainly uses the first definition. The values that can be found for the first paraffins, that is, methane, CH_4, and successively heavier compounds are given in Table 2.

Table 2: Heating Values of the First Paraffins

paraffins	heating power kcal/m³ (32° F 14.7 psia)	kcal/kg
Methane	9,494	13,300
Ethane	16,634	12,380
Propane	23,673	12,030
Normal butane	30,695	11,830
Isobutane	30,622	11,800
Normal pentane	37,714	11,700

The commercial gas from Hassi R'Mel, Algeria, in 1962 (83.5% methane, 7% ethane, 2.1% propane, 1.4% butane and heavier hydrocarbons, 6% carbon dioxide and nitrogen), has a heating power of approximately 9,400 kilocalories (39,330,000 joules) per cubic metre (60° F [16° C], 14.51 pounds per square inch).

ORIGIN OF NATURAL GAS

The accumulation of gas, like that of petroleum, in reservoirs is the result of successive transformation stages of the basic raw materials, followed by accumulation and trapping in reservoirs. The diversity of composition of natural gas reflects the various origins of its constituents. Most of them come from the organic matter in sedimentary rocks during long time periods. Other constituents, however, are of purely inorganic origin.

Adapted from *Revue de l'Association Française des Techniciens du Petrol* (November-December 1969)

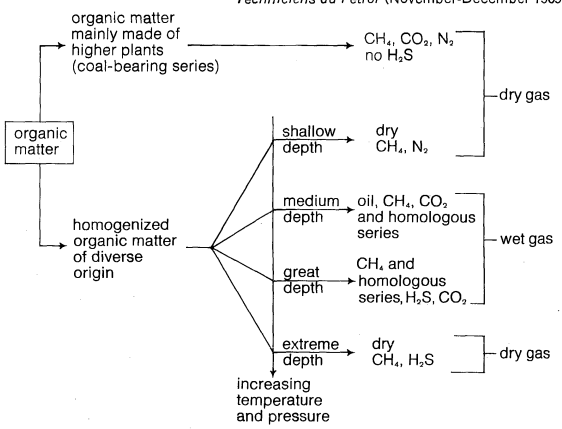


Figure 2: Derivation of natural gas from organic matter.

Organic origin. Freshly deposited sediments at the bottom of seas and lakes are the site of intensive bacterial activity capable of producing methane, carbon dioxide, nitrogen, and nitrogen oxide (NO) from the organic matter incorporated in the sediment. Although methane is formed in abundance (marsh gas, peat gas), ethane and heavier hydrocarbons are almost entirely missing. Some families of bacteria (sulfate-reducing bacteria) also form hydrogen sulfide (H_2S).

With burial of the sediment beneath succeeding deposits, bacterial activity ceases and the organic matter is transformed into kerogen, an insoluble product with a complex macromolecular structure. Through thermal degradations kerogen supplies most oil and natural gas (Figure 2). In an initial phase, at a depth of between 1,000 and 3,000 metres (3,300 and 9,800 feet), the most important occurrence is the formation of crude oil, accompanied by methane and carbon dioxide. Farther down, as the thickness and temperature of the sediment increase, the formation of light, and especially gaseous, hydrocarbons is favoured over that of oil. At very great depths (more than 5,000 metres, or 16,400 feet), molecule cracking leaves only dry methane and kerogen residue with a high carbon content.

Kerogen

This outline of how organic matter evolves covers the most general case, in which this matter has a mean composition resulting from the existence of natural associations and from being stirred by streams and currents. Nevertheless, in individual depressions located in the middle or on the edge of continents, the organic deposits may be made up mainly of lignite coming from higher plants and are then coal bearing. In this case the organic matter produces almost no oil by thermal degradation but only methane (CH_4), carbon dioxide (CO_2), nitrogen (N_2), and carbonaceous matter. The natural gas produced in The Netherlands appears to originate from such processes.

The origin of the hydrogen sulfide sometimes found at great depths raises a special problem. Apart from bacterial activity, it may be derived by a reduction of the sulfates contained in the sediments by hydrocarbons or organic matter, as well as from the action of sulfur on naphthenic hydrocarbons.

Inorganic origin. Helium and some of the argon found in natural gas are products of radioactive disintegration. Helium comes from radioisotopes of thorium and the uranium family, and argon is derived from potassium. As might be expected, the oldest gas accumulations (Paleozoic) are the richest in helium because these geological formations have been the scene of greatest radioactivity. Volcanism may act as a source of carbon dioxide and hydrogen.

Accumulation and entrapment. The sediments containing the organic matter that produces gas are clays or fine limestones—*i.e.*, rocks that are mainly compact and relatively impermeable. The gas is expelled and migrates to reservoir levels, which are porous and permeable and are made up of sands, sandstones, or even vug (or fissure) containing carbonate rocks such as limestones and dolomites. This expulsion occurs under the influence of the pressure difference that exists between porous reservoirs (in which the pressure is about the same as hydrostatic, or equal-sided, pressure) and compact source-rocks (those compressed by the weight of overlying formations).

Gas inside the reservoir circulates in the porous space, for example, between sand grains. Normally the pores, usually less than one millimetre (0.04 inch) in diameter, are filled by water, but gas, because of its very much lower specific gravity, tends to occupy the upper parts of the reservoir level, whereas water remains in the lower parts. For a gas accumulation to exist, the gas must be trapped; that is, the reservoir must be sealed off at the top —*i.e.*, covered by an impermeable stratum or cap rock (clay, salt), with the entire reservoir-cover structure being shaped in such a way as to prevent gas from leaking up to structurally higher zones, in particular to the surface.

Traps for natural gas (Figure 3) are the same as those

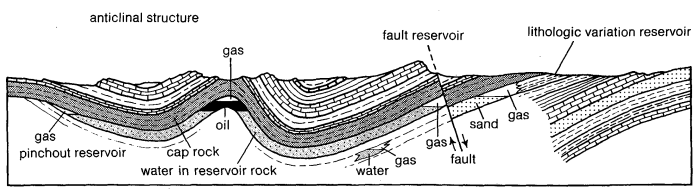


Figure 3: Principal types of natural-gas traps.

for oil accumulations and may be of several kinds. Original sedimentary deposition may form a trap by lateral lithological variation, as in a sand lens or a reef; subsequent sedimentation may form a trap by pinching or wedging out, and structural conditions after deposition may lead to folding or faulting and, hence, to the formation of anticlines and fault traps. Inside traps all of the three fluids (gas, crude oil, and water) may coexist. From the top to the bottom, they are encountered in that order, which is the same as that of their specific gravities.

OCCURRENCE AND PRODUCTION OF NATURAL GAS

World distribution. The annual production and known reserves of the world's leading countries are given in

Table 3. Among the largest accumulations are those of Urengoiskoye in Siberia (reserves of 4,000,000,000,000 cubic metres), the Texas Panhandle in the United States

Table 3: Production and Reserves of Natural Gas, 1970
(000,000 cu m)

	production*	reserves
North America		
United States	673,558	8,233,000
Canada	75,210†	1,513,000
Mexico	18,831	340,000
South America		
Venezuela	48,427	719,000
Argentina	7,665	180,000
Chile	7,628	71,000
Europe		
Soviet Union	212,943†	11,978,000
Netherlands, The	31,669	2,478,000
Romania	24,789	...
Italy	13,138	149,000
Germany, West	12,657	359,000
United Kingdom	11,099	934,000
France	10,194†	...
Africa		
Libya	20,105†	736,000
Algeria	9,628†	4,106,000
Nigeria	8,093	153,000
Asia		
Iran	30,984	3,115,000
Saudi Arabia	20,132‡	2,356,000
Kuwait	16,132‡	1,133,000
United Arab Emirates: Abu Dhabi	7,538	244,000
Iraq	5,663†	510,000
World total§	1,316,855	41,958,000

*Preliminary. †Estimate. ‡Kuwait and Saudi Arabia each include one-half of the production of the former Kuwait-Saudi Arabia Neutral Zone. §Includes production and reserves of other minor producing countries.
Source: *Minerals Yearbook 1970*, vol. I, U.S. Department of the Interior, Washington, D.C., 1972. *World Oil*, August 15, 1971.

(2,000,000,000,000 cubic metres), Slochteren–Groningen area in The Netherlands (1,600,000,000,000 cubic metres), and Hassi R'Mel in Algeria (2,200,000,000,000 cubic metres).

The fact that in all of the hydrocarbon accumulations the percentage of gas fields increases with the depth explains why gas accumulations are mostly encountered in the deeper parts of sedimentary basins. On the Gulf Coast of the United States more than 70 percent of the accumulations discovered at depths greater than 3,600 metres (11,800 feet) are gas fields. In France the extremely deep southwestern part of the Aquitaine Basin contains only gas. Hence, deeper drilling in the future is likely to mean that new discoveries will more often be of gas than of oil.

Flow of natural gas

Gas reservoirs. The laws governing gas flow in a porous medium are complex, and nonlinear terms must be introduced into the equation of movement. The total pressure drop in a circular reservoir with a radius (R), into which is drilled a well with a radius (a), with the well having a rate of flow (Qm), is written as

$$P^2_{(R)} - P^2_{(a)} = AQm + BQm^2,$$

in which $P_{(R)}$ is the mean pressure of the layer, and $P_{(a)}$ is the pressure in the well, and A and B are coefficients that depend on the characteristics of both the gas and the reservoir layer. This formula is used for well testing because it gives the relation between steady state pressure and flow. Its graphic expression is called the back pressure curve of a well. Because production from gas reservoirs began prior to theoretical research on flow, most back pressure curves are usually plotted from the following empirical relation:

$$Q_m = \alpha [P^2_{(R)} - P^2_{(a)}]^n,$$

in which the exponent (n) and coefficient (α) are empirically derived.

Although the theoretical and empirical expressions are not the same, they may be used interchangeably because the differences between them are usually less than the errors of measurement. When flow rates are quite slow, as in parts of the reservoir far from boreholes, it is possible to use simpler flow laws for liquids.

Gas wells and recovery

The outputs from gas wells are often more than 1,000,000 cubic metres (35,000,000 cubic feet) per day. The back pressure curves deduced from flow tests give the relation between flow and pressure. The aim of well tests is to evaluate the qualities of the reservoir and to determine flow with regard to the wellhead pressure for various possible types of well completion (sealing, casing, etc.). The search for the most effective completion equipment is of great importance because gas flows are sometimes enormous, and the influence of tubing diameter on pressure drops (and hence flow rates) is very great. In addition, the gas arriving at the wellhead must be under sufficient pressure to enable it to be piped off to processing installations.

The quality of the producing reservoir and of the downhole equipment is sometimes measured from the flow obtained by applying atmospheric pressure at the bottom of the well. This quality is called absolute open flow potential.

Production from gas reservoirs is facilitated by gas expansion. The recovery factor may attain 50 to 80 percent of the original gas in place, which is much higher than that for primary recovery from oil reservoirs. As a result, complex recovery methods are unnecessary. When gas pressure decreases during production to a nonprofitable level, production is stopped.

In studying the production balance of a reservoir, expansion of the gas itself is the most important factor. The high compressibility of gas practically cancels out variations in volume as the result of interstitial water expansion and reservoir contraction. Therefore, the production balance equation is relatively simple. It shows that cumulative production varies linearly with P/Z, where Z is the compressibility factor and P is pressure.

The slope of the plotted straight line gives the amount of gas initially in place. The line can be extrapolated to forecast cumulative production up to the time the wells are abandoned. Production cannot be continued until total exhaustion because of the drop in the flow through boreholes.

Water may rise into gas reservoirs associated with an aquifer (a water-bearing stratum of rock, sand, or gravel). The expression P/Z is then no longer linear, depending on cumulative production. The extent to which pressure is maintained by inflows of water is never very great because of the great compressibility of gas. Thus, expansion remains the fundamental phenomenon.

Production from condensate gas reservoirs gives rise to phenomena that are quite comparable to those observed for dry- or wet-gas reservoirs. The difference lies in the fact that condensation inside the reservoir may possibly cause the loss of a large proportion of the heavy hydrocarbons initially present in the gas. The condensate deposit is larger where pressure is lower; consequently, it is low in the most outlying parts of the field. It increases as the gas approaches the borehole where sharp pressure drops occur. The gasoline that accumulates around the borehole considerably reduces well productivity.

Depending on the region and on market conditions, condensate is often of greater value than gas itself. Therefore, when the condensate has been extracted on the surface, the dry gas is frequently reinjected into the reservoir so as to maintain pressure and push the wet gas *in situ* toward the boreholes. This process is called dry-gas recycling.

Gas injection and nuclear methods of fracturing

The oil from some reservoirs normally contains natural gas in solution. When this gas is present in large proportions, it is often used on the spot to assist oil production. It may be injected back into the reservoir to maintain pressure or improve oil recovery, or it may help increase the flow in wells by gas lift. Gas reservoirs located near an oil reservoir are sometimes made to produce for these reasons.

Some gas reservoirs located within thick and relatively impermeable reservoir formations cannot be economically exploited by standard methods. In the 1970s, plans were being drawn up to set off confined underground

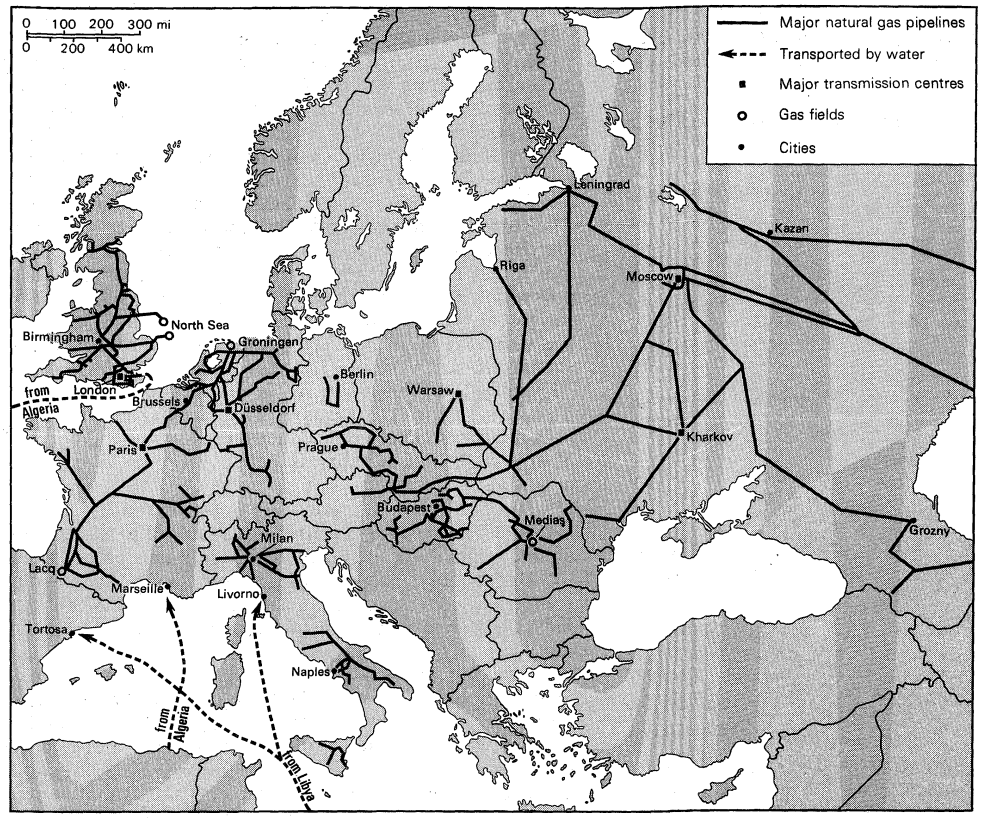


Figure 4: European gas transportation network.

nuclear explosions in the expectation that the zone fractured by the explosion and made to cave in could possibly be used to drain the reservoir. Tests were made in 1967 and in 1969, and initial results were fairly encouraging.

GAS PROCESSING

Natural gas is not directly usable when it comes from the well. Often it is extremely rich in methane and contains heavy hydrocarbons; in addition, it contains varying proportions of nitrogen and carbon dioxide, hydrogen sulfide, mercaptans, and other sulfur compounds. The aim of gas processing is to make the gas conform to standards of use and to recover the liquid or solid components, which are often of greater commercial value than the gas itself.

Treatment and purification

After its arrival at the central plant, the gas must undergo the following series of treatments: (1) expansion at processing pressure, which causes the formation of condensates; these are removed, and the gas is then dehydrated by contact with glycol, which is regenerated; (2) sweetening, which removes the hydrogen sulfide and carbon dioxide by diethanolamine treatment, and fixes them in an absorber; (3) elimination of the mercaptans by soda fixation; (4) stripping (either by cooling or by absorption in an oil), which extracts the hydrocarbons with longer chains than methane and produces a gas conforming to commercial specifications. The rich absorption oil and the purified condensate are fractionated by distillation to produce propane, butane, stabilized gasoline, and residues. These successive operations are performed continuously and require an extremely sophisticated control system. For further details, see GASES, INDUSTRIAL AND DOMESTIC.

TRANSPORTATION AND STORAGE

Transportation of gas. Natural gas is both a fuel (it accounts for more than one-third of the energy consumed in the United States) and a raw material for the chemical industry. Gas fields are often located far from major consumption centres; consequently, the gas has to be transported, although chemical industry transformation plants are frequently installed in the vicinity of production fields. Transportation of natural gas depends upon its form—in a gaseous form it is transported by pipeline under high pressure, and in a liquid form it is transported by boat.

Overland transport by pipeline

Large gas pipelines enable gas to be transported over thousands of miles. Examples are the North American pipelines, which extend from Texas and Louisiana to the northeast coast, and from the Alberta fields to the Atlantic Seaboard. Pipelines extend from Siberia to central Europe, and in the near future they will extend on to western Europe. Lengths of the world's gas pipelines are given in Table 4.

Table 4: Lengths of National Gas Main Transmission Pipelines in the World

	miles	kilometres
United States	223,700	360,000
Soviet Union	34,800	56,000
France	5,590	9,000
Netherlands, The	4,100	6,600
Italy	4,040	6,500
Germany	2,980	4,800
Great Britain	620	1,000
Algeria	500	800
Belgium	375	600

In the early 1970s transportation pressure was 70 bars (1,000 pounds per square inch) because economic studies, in keeping with technological advances, showed that transportation costs are at a minimum for pressures of this magnitude. Pipeline diameters for such long-distance transportation have tended to increase from an average of about 60 to 70 centimetres (24 to 29 inches) in 1960 to a size of 1.20 metres (about four feet) by 1970; some Soviet projects involve diameters of more than two metres (6½ feet) and pressures of about 40 bars (580 pounds per square inch).

Because of pressure losses, the pressure has to be boosted at regular intervals to keep a constant flow rate in the pipeline, and economic surveys show that a compression station should be located every 80 or 100 kilometres (50 or 60 miles) in order to obtain maximum flow under the most economically advantageous conditions. At greater

intervals, the pipeline is doubled in size if the gas flow is to be increased.

Pipelines of this sort are extremely expensive, costing approximately $90,000 per kilometre with a diameter of 70 centimetres (28 inches) and a capacity of about 1,000,000,000 cubic metres (35,000,000,000 cubic feet) per year. The cost of transportation accounts for 75 to 80 percent of the fixed charges, and the operational costs, which vary little with the output, account for the other 20 to 25 percent. Consequently, it is important for such a transportation artery to be used at peak capacity.

Despite the high pressure and circulation rates of 10 to 15 metres (33 to 49 feet) per second for gas, the energy carried by a gas pipeline is still one-third or one-half as much as that carried by an oil pipeline because one cubic metre of oil contains as much energy as 1,000 cubic metres of natural gas.

Marine transport

Petroleum prospecting has revealed the presence of large gas fields in Africa, the Middle East, Alaska, and elsewhere. Transportation from such areas is now effected by boat. The gas is liquefied to −160° C (−256° F) and transported as oil in tankers designed especially for this purpose. One cubic metre of liquid natural gas is equivalent to 600 cubic metres of gas at atmospheric pressure, with the specific gravity of the liquid being relatively low (about 0.55).

This technique has developed very quickly since the first industrial operations involving it began in 1965 from Africa to Great Britain and France. Two other lines opened up in 1970, the first from Libya to Spain and Italy, and the second from Alaska to Japan. Other large-scale projects were under way in the early 1970s from African, South American, and Indonesian fields to the United States, Europe, and Japan. There were 18 liquid natural gas, or methane, tankers in use or being built, compared with only three in 1965.

After 1965, when methane tankers with a capacity of 20,000 cubic metres began to be used, the size of these ships rose to 90,000 and even 120,000 cubic metres, an increase that enabled them to carry from 50,000,000 to 70,000,000 cubic metres (1,800,000,000 to 2,500,000,000 cubic feet) of gas per trip. The cost of a methane tanker was about $40,000,000 or $50,000,000 for the largest.

Gas storage. In the course of marine transport, necessary land storage requires double-walled tanks with special insulation. Such tanks may have unit volumes of as much as 50,000 cubic metres. The low temperatures attained require the use of special steels with a high nickel content or aluminum-base alloys. Other solutions using prestressed concrete were also being tried in the early 1970s. Because the problem of heat loss is less important at ports of arrival, where the liquid gas has to be vaporized, frozen-earth tanks, consisting of enormous holes dug in the ground and made leakproof by freezing the ground water, are used.

Storage facilities on a more extensive scale have been created by using depleted oil or gas reservoirs near consumption centres (northeastern United States, northern Italy), or by the creation of artificial gas fields in aquifer layers. The latter technique developed very rapidly and the number of storage facilities of this type in the United States increased from 125 in 1950 to more than 300 in 1968, with the storage volume being more than 130,000,000,000 cubic metres. In Europe, there are six such underground storage areas in France, five in Germany, and about ten in the Soviet Union.

BIBLIOGRAPHY. J.A. CLARK, *The Chronological History of the Petroleum and Natural Gas Industries* (1963), a detailed chronology of technical and human facts; E.L. RAWLINS and M.A. SCHELLHARDT, *Back Pressure Test Data on Natural-Gas Wells and Their Application to Production Practices* (1953), a classic report of the U.S. Bureau of Mines describing and explaining the back pressure method, with data for more than 500 gas wells analyzed; M. MUSKAT, *Physical Principles of Oil Production* (1949), a fundamental work on the basic principles of gas and petroleum; P. CHAUMET, *Écoulement monophasique des fluides dans les milieux poreux* (1968), a comprehensive book on fluid flow in porous mediums; D.L. KATZ and K.H. COATS, *Underground Storage of Fluids* (1968), a comprehensive review of various types of underground storage and technical problems; C.H. ATKINSON, D.C. WARD, and R.F. LEMON, *Gasbuggy Reservoir Evaluation 1969 Report*, No. PNE-G-54 (December 1969), an extensive report on nuclear explosion effects and production tests.

(J.D.Po./H.Bre./J.Mar./P.Mon./B.T.)

Natural Law

"Natural law," or "law of nature," is the term used by philosophers and jurists, not always precisely in the same sense, to refer in general to a system of right or justice held to be common to all mankind. As a body of generally recognized principles of right conduct, it is often contrasted with "positive law," the legal rules formally promulgated by the state and enforced by definite sanctions. Throughout the history of the concept, from Greek antiquity, there have been disagreements over the meaning of natural law and over its relation to positive law.

The Greek and Roman period. The development of the Greek city-states was accompanied by much reflection on the part of philosophers on the nature of the order that governed the relations of men. In the 5th and 4th centuries BC, a group of such thinkers known as the Sophists distinguished between *physis* ("nature") and *nomos* ("law"), the former being wise and eternal, the latter arbitrary and born of expediency. Law was thus taken to be utilitarian, designed to serve the interests of the strong, according to one variant of this view, or of the weak, according to another. This kind of relativism, though it seems now to have been an anticipation of some modern theories, was not at that date conducive to social stability. In an alternative approach, some Sophists appealed to the law of nature, which was thought to govern the evolution of natural phenomena. Man, by his faculties of reason and will, in this view has insight into goodness and badness and can develop in either direction; but he ought to devote his powers to developing according to a natural morality. Socrates, Plato, and, after them, Aristotle asserted the discoverability of immutable standards with reference to which positive law—that is, the law as actually formulated—might be evaluated.

Sophist view of natural and positive law

Neither Socrates nor Plato concerned themselves explicitly with natural law. But Socrates did initiate the search for definitions of things and qualities. And Plato held that if this search succeeded, man should know what justice, courage, and so forth really were. Since the word know implied that what is known is independent of the individual mind of the knower, he held that there must be such independent realities as justice and courage, and he concluded that such realities as justice and courage (which he called "Forms" or "Ideas") were permanent and unchanging entities, existing independently of particular things and actions. He held that these realities could not be apprehended by perception or empirical observation but must be pursued by pure reason, by the special activity that he called "philosophy," the goal of which was certain, final, and infallible knowledge. Since the "Forms" alone are real, he held that man's reason, which leads him to knowledge of them, is the highest side of his nature and most akin to the eternal and the divine. Since such knowledge is the only firm guide to action, reason must be trained and directed to achieve it; and, in the state, philosophers must be kings.

In *The Republic*, Plato asserted that insight into justice (or natural law) is confined to philosophers, an elite to whom legislation and government should thus be exclusively entrusted. In the real rather than the ideal world, however, such politics risk the dangers of dictatorial rule. In his last work, *Laws*, he thus dropped the idea of the philosophers' rule and advocated instead a mixed constitution combining monarchy and democracy.

Aristotelian concept of perfect natural law

Aristotle held that there is a natural law or justice that "everywhere possesses the same authority and is no mere matter of opinion" and that is discoverable by the use of reason. This perfect natural law, he said, embraced citizens and their life in society, though slaves could not share in it but could receive only its blessings, more or less second-hand, in the form of good treatment from their masters. Barbarians, too, were excluded. Like Plato, Aristotle supported the tendency to restrict perception of high principles and enjoyment of the good life to a few people

especially favoured by talents, birth, wealth, and education. The education that aided the understanding of justice and natural law was not a literary, technical, or intellectual education but the character training of a man brought up in a good family and a good city.

The Stoic school of philosophy, founded at the beginning of the 3rd century BC, introduced a new attitude and conceived an egalitarian law of nature. Previous thinkers had made familiar the part played by reason in molding behaviour. The Stoics in their turn asserted that reason is an attribute of all men and not of citizens alone. The state of nature, they said, is one of harmony, regulated by reason. This had been perverted by selfishness, and the task confronting mankind was to fashion its laws in such a way as to recover the antique ideal. To live according to reason was to live naturally. The laws of nature and morality were considered identical. Reason, moreover, was common to men in all countries, and natural law was therefore all-pervading.

The era of Roman ascendancy, which superseded the Greek, contributed hardly at all to natural law theory. A significant development, however, was the evolution of the *jus gentium*, or law of nations, applicable to both Romans and non-Romans (see ROMAN LAW). Although the *jus gentium*, as a gradual synthesis of practical rules and precepts, was not deduced from any general presupposition, it was applied to persons of all races, and there was a tendency in later times to identify it, to some extent, with natural law. Another significant development of the Roman period was the work of the Christian Fathers, notably St. Augustine of Hippo, who developed the idea that man had lived freely under the natural law before his fall and had since lived in bondage under sin and the positive law.

The medieval period. In Europe during the Middle Ages, the church was the main centre of learning, and it fell to the churchmen to propound a theory of law to suit the needs of the times. The classical philosophy was clearly adaptable, but to have followed the teachings of the pagan philosophers would have detracted from the authority of the church. The solution was to identify the law of nature of the ancients with the law of God, and this was the method in the 12th century of Gratian, the so-called father of the science of Canon Law, who equated the natural law with the revealed law of the Old and the New Testaments, particularly the Christian version of the Golden Rule, "So whatever you wish that men would do to you, do so to them."

Combining the law of God and the law of nature

In the 13th century St. Thomas Aquinas worked out a highly systematized interrelation between the eternal law of God, natural law, and human law. The eternal law of the divine reason, he maintained, though it is unknowable to man in its perfection because it is in God's mind, is yet known in part not only by revelation but also by the operations of human reason. The law of nature, which is "nothing else than the participation of the eternal law in the rational creature," comprises those precepts that mankind is thus able to formulate—namely, the preservation of one's own good, the fulfillment of "those inclinations that nature has taught to all animals," and the pursuit of the knowledge of God. Human law must be the particular application of the natural law. The Thomist system thus combines the philosophy of the Greeks, the law of the Romans, the teachings of the Christian Fathers, and contemporary Pragmatism with consummate power and skill. What is most striking in it, however, is its uncompromising appeal to reason, the law of God itself being declared to be "nothing else than the reason of divine wisdom."

Thomistic groups of principles

St. Thomas distinguished three groups of principles of natural law. The supreme group of principles called for love of God, love of neighbour according to the Golden Rule, and, generally, the performance of good and the avoidance of evil. The secondary principles, derived from the first, were contained in the Ten Commandments, involving obedience toward parents and authority, protection of the person and life of another and of his property, truthfulness, and faithfulness to one's pledged word. The third group of principles consisted of concrete rules of justice that were formed by applying the first two groups of principles to individual situations. St. Thomas believed that in every man there is a natural disposition to recognize the first two groups of principles. For him, this disposition could not err and was incorruptible, though the application of those principles to a given situation, made by the conscience, was far from infallible. The mistake was thought to occur always in the minor premise of a moral syllogism; *i.e.*, everyone knows that murder is wrong, but it is not always easy to say whether a particular act is murder.

Other Scholastic philosophers—for instance, John Duns Scotus, William of Ockham, and Francisco Suárez—emphasized the divine will instead of the divine reason as the source of natural law. Duns Scotus' view is illustrative. He asserted that God can will whatever does not imply a contradiction; hence he could have created another universe in which morality might differ somewhat from the morality of this universe; but once granted that he has created man in this particular universe, it would imply a contradiction in God if he ordered men to hate one another. What gives the natural law stability is the nature of God and of the creatures to whom it applies, even though there is an arbitrary element consisting of God's free choice of this world as the object of his creative activity. Phrased another way, God has two distinct powers—his absolute power and his regulatory power—and there is no necessary connection between the one and the other. Thus, for instance, the eternal law of God (involving his absolute power) may not be reflected precisely in the natural law of the Ten Commandments (involving his regulatory power); and this is evident because he has occasionally exempted certain men from observing certain commandments, and he would not do so if they were a part of eternal law. This so-called voluntarism of Duns Scotus and others influenced Catholic jurisprudence of the Counter-Reformation, but the Thomistic doctrine was later revived and reinforced to become the main philosophical ground for the papal exposition of natural rights in the 19th and 20th centuries.

The Renaissance and the Age of Reason. The intellectual atmosphere in which the Middle Ages ended was dominated by three factors: the growth of rival Christian sects, the birth of the nation-state, and the rise of science. Each of these affected the concept of natural law. The breakup of Christendom stimulated such a variety of selection, application, and interpretation of ideas and laws as almost to involve a principle of private moral judgment. There arose a demand for a moral standard independent both of ecclesiastical authority and of biblical texts. The rise of independent nation-states posed new questions regarding not only relations between the ruled and their rulers but also relations between the states, such as the question of whether they were to be governed by naked power politics. Finally, science had begun its work of subjecting all natural phenomena to free rational inquiry. It was inevitable that social and moral phenomena should be subject to the same critical review, and the result was a demand for a rational basis for natural law.

The man who first met these demands in a way that could be called "modern" was Hugo Grotius, a 17th-century Dutch jurist. Echoing the Stoics, he saw the universe as dominated by a rational law of nature. For man, the law of nature was the collection of those rules that followed inevitably from his essential nature. Natural law was therefore intelligible and immutable. Man was distinguished from the animals by his "appetite" for tranquil association with his fellows and by his power to act on general principles. If men were to live at peace with their fellows there were certain rules that must be observed: the keeping of promises, the recognition of human equality, the principles of equity and justice, of parental responsibility, and of marital fidelity. Here then was that set of moral standards, independent of church and of Scripture, that the new Europe demanded. Here were rules not tied to narrow national tradition but based on the universal nature of man, rules to which princes and governments no less than private individuals should be subject.

The law of nature had long ago been identified with the *jus gentium*. In one sense this had meant that these rules could be discovered not only by deduction from the nature of man but also inductively by noting their universal acceptance in all human societies. But, for Grotius, *jus gentium* had a deeper meaning as a rule governing the relations between peoples considered as moral agents. Some of the natural laws cited above (marital fidelity, parental responsibility) could not apply to the behaviour of sovereign states; and indeed Grotius was inclined to emphasize one single rule as relevant here—the rule that promises are to be kept. Men—and nations—as essentially social beings, desired a peaceful social order.

Natural law and the social contract

A concept that was to become important at this time (a version of which Hugo Grotius himself propounded) was the social contract. The individual, it was said, was born into a state of nature, which was happy or unhappy according to the particular version of the theory; he then, by the exercise of natural reason, formed a society by means of a contract or compact with other individuals. The Englishman Thomas Hobbes, who lived through the instability of the English Civil Wars of the mid-17th century, took the social contract theory and made it a vehicle for authoritarianism. According to him, men in the state of nature were selfish and cruel and, in order to inhibit their brutality and give themselves any kind of future, needed to come under the governance of an absolute authority. They thus entered into a social contract. The resultant sovereign, whose actions had been authorized by the subjects, became absolute, a social leviathan, whose power guaranteed stability, safety, and contentment. All that Hobbes was prepared to concede was that the sovereign (who was responsible only to God) should act according to natural law, which consisted of moral injunctions that sovereigns were expected to observe. Apart from this, the individual had no guarantees. Positive law was thus distinguished from natural law, its distinctive feature being the fact that it emanated from sovereign authority. Hobbes made the form of promulgation, rather than its content, the determining feature of law.

A milder version was forthcoming a generation later from John Locke. The reign of the Stuarts in England had revealed the danger of absolutism to the individual. Accordingly, Locke maintained that individuals had not surrendered all their rights to the sovereign but only the powers appurtenant to the preservation of law and security. Other "natural" rights relating to the integrity of the person and property were retained because they were fundamental. What Locke was seeking to do was to protect the individual from the arbitrary acts of a sovereign.

In France the continuing need to protect the individual against an oppressive monarchy found expression in the 18th century in the theories of Montesquieu and Jean-Jacques Rousseau. Montesquieu argued that natural laws predated society and were superior to those of religion and of the state. Rousseau went on to assert that early man or the savage had been virtuous in isolation and was actuated by two principles "prior to reason"—namely, self-preservation and compassion (innate repugnance against the sufferings of others), or freedom and equality. In forming a social contract, individuals did not surrender their natural rights to any single sovereign but to society as a whole, and this was their guarantee of continued freedom and equality. Society, having come into being for this purpose, was expected to confirm these rights as civil liberties. Society and law were thus dependent upon general will, on popular, as distinct from parliamentary, sovereignty. Little wonder that this theory was utilized as the philosophy of the French Revolution.

19th and 20th centuries. By the beginning of the 19th century a reaction against the natural law theories had developed for many reasons. First, and chief among them, was the critical spirit of a new scientific age that looked askance at unproved hypotheses. The social contract, in particular, was said to be palpably false, historically untrue. Second, it was asserted that deductive reasoning from a priori assumptions was incapable of yielding solutions to the problems of complex societies. Third, pure reason had been a slogan of the French Revolution with all its attendant excesses, and the reaction against the Revolution caused a reaction against using reason as the exclusive guide to social reform. Fourth, there was a growing conviction that improvement of the law would be more fruitfully approached on a practical basis, and not by appealing to reason. It was thought needful, above all, to understand existing institutions by means of a thorough analysis of them in order to bring about their improvement.

Waning of natural law theory

If there was a diminution of theorizing about natural law, such theorizing never disappeared altogether. Particularly in the 20th century, when there was a search for new values amid the instabilities of wars and mass society, there was a revival of natural law theory among some writers. This interest was stimulated by the need for guidance amid the conflicting ideologies that had sundered the world, so that the legality of regimes such as that of Nazi Germany might be opposed from the outset, before they could unleash terrible suffering. The collectivist tendencies of the age were reflected in certain new emphases. Léon Duguit, professor of law at the University of Bordeaux, preached that social solidarity and interdependence was the all-important fact of society. Human activity should be directed at ensuring the fulfillment of this social solidarity. Men owed duties to help to promote it, but no one had any right unless it were the right to do his duty within the social organism. The validity of laws was to be judged according to social solidarity as the criterion. All this was very far removed from those theories that insisted on fundamental rights.

An interesting feature of modern natural law was the endeavour to establish evaluative criteria empirically—that is, to derive them from general social experience. Critics contended that this kind of empirical search for common values—such as the United Nations' attempts to define human rights—is not the same as the explication of an immutable natural law. But proponents assert that, although their criteria had to be expressions of value judgments, they did not stem from ideologies or emotions. A dispassionate judgment founded on a study of all the available data and opposing considerations was far removed from confessions of faith or spontaneous reactions. In any event, the task that still confronts proponents of the search for common values is twofold: to evolve the values and to win for them a measure of acceptance.

BIBLIOGRAPHY. The chief ancient classics are PLATO, *Laws, Republic,* and *Phaedrus*; and ARISTOTLE, *Nicomachean Ethics, Rhetoric,* and *Topics.* The chief medieval classics are ST. AUGUSTINE, *City of God* and *De Libero Arbitrio*; ST. THOMAS AQUINAS, *Summa Theologica;* and JOHN DUNS SCOTUS, *Opus Oxoniense.* The chief modern classics are HUGO GROTIUS, *On the Law of War and Peace* (1625); THOMAS HOBBES, *The Leviathan* (1651); JOHN LOCKE, *Two Treatises on Government* (1690); MONTESQUIEU, *De l'Esprit des lois* (1748; Eng. trans., *Spirit of the Laws*); ROUSSEAU, *Du Contrat social* (1762; Eng. trans., *Social Contract*); and L. DUGUIT, *Traité de droit constitutionnel,* 2 vol. (1911; Eng. trans., *Treatise on Constitutional Law*).

(Ed.)

Nature, Philosophy of

The discipline here entitled "the philosophy of nature" consists in the investigation of substantive issues regarding the actual features of nature as a reality and is divided into two parts: the philosophy of physics and the philosophy of biology. In this discipline, the most fundamental, broad, and seminal features of natural reality as such are explored and assessments are made of their implications for man's metaphysics, or theory of reality; for his *Weltanschauung,* or "world view"; for his anthropology, or doctrine of man; and for his ethics, or theory and manner of moral action. These implications are explored on the assumption that man's understanding of the natural setting in which his life is staged strongly conditions his beliefs and attitudes in many fields.

In its German form, *Naturphilosophie,* the term is chiefly identified with Friedrich Schelling and G.W.F. Hegel, early 19th-century German Idealists who opposed it to *Logik* and to the *Phänomenologie des Geistes* ("of

the spirit or mind"). Employment of the term spread, in due time, beyond its narrower historical context in German Idealism and came to be used, particularly in Roman Catholic parlance, in the sense that it bears in this article (*e.g.*, the philosophies of physics and biology). Despite a notable decline in its usage in more recent years, the term is here employed, in the interest of the clear delineation of topics, as a complement to the philosophy of science (see SCIENCE, PHILOSOPHY OF), the discipline to which its subject matter has been allocated by recent philosophers. Thus in this work, the article on the philosophy of science is largely restricted to man's approach to nature, and thus to epistemological (theory of knowledge) and methodological issues, while that on the philosophy of nature encompasses the more substantive issues about nature as it is in itself.

I. Philosophy of physics

PHYSICS AS A FIELD OF INQUIRY

Essential features. Physics is concerned with the simplest inorganic objects and processes in nature and with the measurement and mathematical description of them. Inasmuch as the binding forces of chemistry can now, at least in principle, be reduced to the well-known laws of physics, or calculated from quantum mechanics (the theory that all energy is radiated or absorbed in small unitary packets), chemistry can henceforth be considered as a part of physics in theory if not in practice. Moreover, it has become clear, through the general theory of relativity (which formulates nature's laws as viewed from various accelerating perspectives), that there is an aspect of geometry, too, that can be regarded as a part of physics. The fact that, over a wide range of circumstances, Euclidean, or ordinary uncurved, geometry presents a good approximation to reality is considered today not as a fact stipulated by a necessity of thought, nor a derivative from such a necessity, but as a fact to be established empirically; *i.e.*, by observation. In their application, the laws of Euclidean geometry refer to those experiences that arise with measurements of length and angle and optical sightings as well as with surface and volume measurements. The possibility—already extensively elucidated in antiquity—of deriving geometrical propositions by deduction from a few axioms, assumed without proof to be correct, had given rise in earlier philosophy to the opinion that the truth of these axioms must and could be guaranteed by a kind of knowledge that is independent of experience. The recognition of such a priori knowledge, however, has been superseded by the modern development of physics. While it is granted that a pure geometry is free to posit any axioms that it pleases, a geometry purporting to describe the real world must have true axioms. Today it is considered that, if Euclidean geometry is true of the world, this truth must be established empirically; the axioms would be true because the conclusions drawn from them correspond to experience. Actually, the world appears Euclidean, however, only when this experience is limited to cases in which the distances are not too great (not much greater than 10^9 light-years) and in which gravitational fields are not too strong (as they are in the vicinity of a neutron star).

Axiomatic versus empirical geometry

The possibility of deducing all known laws or regularities as logical inferences from a few axioms, which was discovered in Euclidean geometry, became a model for the construction also of another chapter in the history of physics. The classical physics of Newton, the 17th–18th-century father of modern physics, had employed Euclidean geometry as a foundation and had portrayed the solar system as a system of mass points subject to his mechanical axioms. The laws for falling bodies framed by the 16th–17th-century Italian physicist Galileo are the simplest logical consequences of Newton's axioms, and the laws framed by Johannes Kepler, a 16th–17th-century German astronomer, which precisely describe the motions of the planets, follow from them.

In addition to the laws of mechanics there are those of the broad sphere of electromagnetic phenomena as summarized in the equations of James Clerk Maxwell, a 19th-century Scottish physicist, which describe both the electric and magnetic fields and the laws of their mutual changes, equations that may thus be considered as the axioms of electrodynamics. Because they assume the mathematical form of partial differential equations—which express the rates at which differentials (small or infinitesimal distances or quantities) in several dimensions change with respect to their neighbours—electrodynamics is a local-action theory rather than an action-at-a-distance theory as in older formulations modelled after Newton's law of gravitation. The principle of local action states that the variations of electromagnetic magnitudes at a point in space can be influenced only by the electromagnetic conditions in the immediate vicinity of this point. The finite velocity of propagation for electromagnetic disturbances, which follows from this principle, leads on the one hand to the existence of electromagnetic wave events and on the other hand to conformity with the requirements of special relativity (a theory that formulates nature's laws as viewed from the perspectives of various velocities), which demand a maximum finite velocity for signals—the velocity of light in a vacuum.

Macroscopic versus microscopic physics

The most important division of physics today is one that replaces the traditional distinctions between mechanics, acoustics, and other classical branches of physics with that between macroscopic and microscopic physics, in which the latter investigates the conformity of atoms to law and their reactions in discrete quantum jumps, whereas the former extends from the level of ordinary human experience into astronomy to a total comprehension of the universe, attained through theoretical endeavours in the field of cosmology. Because it is now possible to observe especially bright objects (quasars) that are located perhaps 10^{10} light-years from the Earth, the possibility of empirically testing cosmological models is beginning to arise. In particular, the application of non-Euclidean, or curved, geometries to the cosmos has suggested the conception of a finite, yet boundless, world space (positively curved), in which the maximum possible distance between two points would no longer be much greater than 10^{10} light-years.

Historical sketch. In the historical development of physics before the 17th century, geometry was the only field in which extensive advances were made; besides geometry, only the rudiments of statics (the laws of levers, the principle of hydrostatics of the 3rd-century BC scientist Archimedes) were clarified. After Galileo had discovered the laws of falling bodies, Kepler's laws describing the motions of the planets and Newton's reduction of them to a set of dynamical axioms established the science of classical mechanics, to which was annexed the investigation of electromagnetism. These developments culminated in the discovery of induction by Michael Faraday, an English physical scientist, the knowledge of local action by Faraday and Maxwell, and the discovery of electromagnetic waves by a German physicist, Heinrich Hertz. It was not until the 19th century that the law of the conservation of energy was first recognized as a general law of nature, through the work of two physicists, Julius von Mayer in Germany and James Joule in England, and that the concept of entropy (see below *Entropy and the direction of time*) was formulated by Rudolf Clausius, a mathematical physicist. At the beginning of the 20th century, the German physicist Max Planck introduced the so-called quantum of action, $h = 6.626 \times 10^{-27}$ erg-seconds, which, when multiplied by the vibration frequency, symbolized by the Greek letter nu, ν, demarcates a basic packet of energy. Albert Einstein then extended the quantum theory to light. The real existence of atoms was proved by him and other investigators, and the science of microphysics thus arose. The researches of Niels Bohr on the quantum-theoretical significance of atomic spectra paved the way for broader search into the fine details of quantum laws, the final comprehension of which was introduced by Werner Heisenberg in 1924 and then systematically developed by Max Born, Heisenberg, and Pascual Jordan, of Germany, and by P.A.M. Dirac, of England. Moreover, Erwin Schrödinger, an Austrian physicist, pursuing a line of thought pointed out by Einstein and Louis de Broglie, arrived at results that were

outwardly quite different from those of Heisenberg *et al.*, but were mathematically equivalent. The quantum mechanics, or wave mechanics, created by these men, which formulated quantum phenomena, were later extended to quantum electrodynamics.

General and special theories of relativity

Einstein's theory of relativity, first formulated in 1905, which was eventually extended from a special to a general formulation, brought about a revolutionary transformation in physics similar to that induced by quantum theory. The Newtonian mechanics of mass points turned out to have been merely an approximation to the more exact relativistic mechanics. The most important consequence of the special theory of relativity, the equivalence of mass (m) and energy (E),

$$E = mc^2, \qquad (1)$$

in which c is the velocity of light, was formulated by Einstein himself.

After 1916 Einstein strove to extend the theory of relativity to the so-called general theory, a formulation that includes gravitation, which was still being expressed in the form imparted to it by Newton; *i.e.*, that of a theory of action at a distance. Einstein did succeed in the case of gravitation in reducing it to a local-action theory, but, in so doing, he increased the mathematical complexity considerably, as Maxwell, too, had done when he transformed electrodynamics from a theory of action at a distance to a local-action theory.

The great importance of physics for the technology that depends upon it—which has become a leading factor in the rapidly increasing development in the conditions of human existence—is shown historically in the close connection of decisive technical developments with basic advances in physical knowledge. Einstein's equivalence of mass and energy—to cite but one example—pointed to the atomic nucleus as an energy source that could be opened up through the study of nuclear physics. Moreover, the intellectual influence proceeding from physics and affecting the development of modern thought has become especially strong through the deepened grasp of the concept of causality that has followed from quantum theory (see below *Modalities of the natural order*).

BASIC CHARACTERISTICS AND PARAMETERS OF THE NATURAL ORDER

Framework of the natural order. Earlier mathematicians and particularly Richard Dedekind, a pre-World War I number theorist, have precisely defined the concept of real numbers, which include both rational numbers,

Measurements in $t; x, y;$ and x, y, z

such as $\frac{277}{931}$, expressible as ratios of any two whole numbers (integers), and irrational numbers, such as $\sqrt{27}$, π, or e, which lie between the rationals. By reference to these numbers, the Newtonian concept of space and time, which presupposes a Euclidean geometry of space, may be made precise: the values of the time t, ordered according to the ideas of earlier and later, can be made to correspond to the single real numbers, ordered according to those of smaller and larger. Also, the points on a straight line can be brought into correspondence with the real numbers in such a manner that the location of a point P between two other points P_1 and P_2 corresponds to a number assigned to P that lies between those assigned to P_1 and P_2.

Cartesian analytic geometry. Guided by the wish to find a method that allows the systematic proof of all philosophical truths, René Descartes, often called the founder of modern philosophy, established in the 17th century the analytic geometry of Euclidean planes. In it the points of a plane can be designated by two numbers x, y, their coordinates. One chooses two orthogonal coordinate axes, $x = 0$ and $y = 0$, like those of a graph, and, with any point P, associates its two projections, one upon each coordinate axis, which define the location of P. A curve in the x–y plane is then expressed by an equation $f(x,y) = 0$, shorthand for any equation ("function") containing x's and y's. In the context of analytic geometry, every theorem of plane Euclidean geometry may be expressed by equations and thus be analytically proved.

This procedure can also be extended to three-dimensional Euclidean space by introducing three mutually perpendicular axes x,y,z. In this case, there are two different axis systems—either congruent or mirror reflections—analogous to right-handed and left-handed screws.

Metrology in modern physics. The simple space–time relationships of Newtonian physics, which was brought to completion in Maxwell's electrodynamics, have been changed in many ways by modern developments. The concept of simultaneity has been made relative by the special theory of relativity; every time measurement t is thus tied to a definite inertial system or moving frame of reference. It is accordingly appropriate to speak not primarily of points in time but of events, which are defined in each case by giving both a point in space and a point in time.

More specifically, an inertial system is a coordinate system that, relative to the fixed stars, is in uniform, straight-line motion (or at rest) with no rotation. In all inertial systems, Newton's principle of inertia, which states that all mass points not acted upon by some force persist in uniform motion with a constant velocity, is valid.

Hyperspherical space–time: x,y,z,u

Moreover, cosmological theories make it probable that space in the real astronomical universe corresponds only approximately to the relationships of Euclidean geometry and that the approximation can be improved by replacing Euclidean space with a space of constant positive curvature. Such a space can be mathematically defined as a three-dimensional hyperspherical "surface"

$$R^2 = x^2 + y^2 + z^2 + u^2 \qquad (2)$$

in a hypothetical Euclidean space of four dimensions with mutually perpendicular x,y,z, and u coordinate axes.

The assertion that the foregoing statement has no operationally comprehensible content—*i.e.*, no content provable by performable measurements—is designated conventionalism, a view that is based on a remark by a French mathematician, Henri Poincaré, who was also a philosopher of science, that a fixed non-Euclidean space can be mapped point by point on a Euclidean space so that both are suitable for the description of the astronomical reality. The range of this remark is limited, however, in that this mapping, though it can indeed carry over points into points, can in no way carry over straight lines into straight lines. Hence, many philosophers of science have held that, as long as astronomical light rays are held to be straight lines, the question of a possible curvature of space (*i.e.*, a deviation from Euclidean conditions) will by no means be solved by some arbitrary convention; that it signifies, instead, a problem to be solved empirically. If the universe in fact has a positive constant curvature, then every straight line has a length that is only finite, and its points no longer correspond, as in the Euclidean case, to the set of all real numbers.

In a very definite manner, cosmological facts have further indicated that time is by no means unlimited both forward and backward. Rather, it seems that time as such had a beginning about 10^9 or 1.2×10^9 years ago; thus, with an explosive beginning, the cosmic development began as an expansion.

The foregoing discussion has considered only the replacement of Euclidean spatial concepts by an elementary non-Euclidean geometry corresponding to a space with a constant curvature. According to Einstein, however, the fundamental idea of a still more generalized Riemannian geometry, so-called after Bernhard Riemann, a geometer and function theorist, must be brought into play in order to produce a local-action theory of gravitation.

Riemannian space: g_{kl}

Riemannian geometry is a further development of the theory of surfaces created by the 18th- and 19th-century German mathematician and astronomer Carl Friedrich Gauss, often called the founder of modern mathematics, a theory that aimed to investigate the curved surfaces of three-dimensional (Euclidean) spaces with exclusive regard to their own inner dimensions and no consideration of their being imbedded in a three-dimensional space.

Gauss thought that the points on such a surface could be

specified by reference to two arbitrary coordinates u and v defined with the help of two single-parameter families of curves, u = constant and v = constant. The square of the infinitesimal distance between two adjacent points of the surface, ds^2, is then a quadratic form of the differentials du and dv, belonging to the pair of points, namely,

$$ds^2 = g_{11}du^2 + 2g_{12}dudv + g_{22}dv^2, \qquad (3)$$

in which the coefficients g_{kl} are functions of position. One can then calculate the curvature corresponding to the location of the pair of points according to a prescription given by Gauss, a curvature that measures the deviation from Euclidean behaviour that exists at this point. The curvature is a definite function of the g_{kl} and their derivatives.

Riemann has extended Gauss's considerations to the case of a three-dimensional space that can have different curvature properties from place to place (expressed by several functions of position that are collectively called the curvature tensor); and Einstein has generalized these ideas still further, applying them to the four-dimensional space–time continuum, and has thereby attained a reduction of the Newtonian action-at-a-distance theory of gravitation to a local-action theory.

Contents of the natural order. Among the most basic constituents of the physical world are symmetries, fields, matter, and action.

Symmetries. Symmetry is one of the chief concepts of modern mathematics, which combines the different symmetries belonging to an object or a concept into groups of relevant symmetries. The a priori investigation of the totality of possible groups, defined with respect to some operation (such as multiplication), comprises a division of modern mathematics called group theory.

Three-dimensional Euclidean space displays several important symmetry properties. It is homogeneous; *i.e.*, arbitrary shifts in the origin or zero point of the coordinate system produce no change in the analytic expression of the geometrical laws. It is also isotropic; that is, rotations of the coordinate system leave all geometrical laws in effect. Further, it is symmetric with regard to mirror reflections. It is tempting to suppose that these symmetry properties of space are also valid for the physical processes that occur in space, and this is indeed true over a wide range of cases, but not in all cases (for exceptions, see below *Charge/parity/time symmetry*).

Vector fields: directional quantities. That Newtonian mechanics and Maxwellian electrodynamics display in fact all of the symmetries of Euclidean space is revealed by the fact that they can be formulated in the language of vector analysis (see ANALYSIS, VECTOR AND TENSOR). Passing over the more familiar Newtonian mechanics (see MECHANICS, CLASSICAL), a few points about Maxwell's theory may be mentioned. This theory can be made to satisfy the requirements of operational thinking by ascribing to the electric and magnetic field strengths the significance of measurable physical realities, which makes it unnecessary to interpret them as states of a mysterious, hypothetical substance or ether, for which, in any case, the special theory of relativity (with the equivalence of all inertial systems) has no place.

Field; line integral; gradient

Mathematically interpreted, a vector $\boldsymbol{a}$ represents a quantity with both magnitude and direction, which preserves its length or value and its direction when displaced. The vector field—*i.e.*, the association of a vector with every point in space (*e.g.*, electric field strength, or electric current density)—and the line integral (or summation) of a vector field V along a curve K leading from a point P to a point P' are basic concepts in vector analysis. To obtain the line integral, the curve K is divided into infinitesimal elements $\boldsymbol{ds}$, the scalar (numerical or nonvector) product of ds with the value of V at that point is taken, and the results are summed with an integration.

A small surface area envisioned with a given sense of rotation around its boundary curve can also be described by a vector. In this instance, the vector, $\boldsymbol{dF}$ is perpendicular to the surface and forms, with the sense of rotation about the boundary, a right-handed system. Its magnitude is the area of the surface. The flux of the vector field V through the surface dF is called the scalar product $V \cdot dF$.

If $\mathbf{V}$ has the property that the line integral along every closed curve K is equal to zero, then $\mathbf{V}$ is said to be irrotational. This property is equivalent to the requirement that the vector field be a so-called gradient field; *i.e.*, that there exist a scalar field quantity W with the property that the difference in the value of W at two points P and P' is equal to the line integral of the vector field $\mathbf{V}$ from P to P' (along any arbitrary curve K). If $\mathbf{V}$ is, for example, an electrostatic (charged) field, the significance of being irrotational is that one can gain no mechanical work in leading a small test charge around any closed curve; the work involved is equal to zero. For an unclosed curve K, however, the movement of the test charge yields an amount of mechanical work that is proportional to the potential difference between the endpoints of the curve. The components of the gradient of W, expressed in partial derivatives ∂, are

$$\text{grad } W = \frac{\partial W}{\partial x}, \frac{\partial W}{\partial y}, \text{ and } \frac{\partial W}{\partial z}. \qquad (4)$$

Curl; divergence; charge; and current density

If the vector field is not irrotational, there can then be constructed from it an adjunct rotational field, called curl V, by considering a small (infinitesimal) surface area dF located at a point P and forming the line integral of $\mathbf{V}$ along the boundary curve of dF. Then, when this line integral is divided by the magnitude of the surface area, the component of the curl $\mathbf{V}$ parallel to the vector dF is obtained.

On the other hand, the flux of a vector field $\mathbf{V}$ out of a closed surface can be formed by integration. If this flux is always zero (for every choice of a closed surface), $\mathbf{V}$ is called source-free. Otherwise, there is a so-called divergence of $\mathbf{V}$ at a point P, which is defined as follows: one divides the net flux of $\mathbf{V}$ out of a small surface that surrounds P by the volume enclosed by the surface. The limit of this quotient for infinitesimally small surfaces is called the divergence of $\mathbf{V}$ at P or the source field div $\mathbf{V}$.

The formulation of the basic laws of electrodynamics given by Maxwell is called the Maxwell equations. These equations contain, for example, the statement that, in a vacuum, the source field of the electric field strength is proportional to the spatial electric charge density, symbolized by the Greek letter rho, ρ, and that the magnetic field strength is source-free (divergence equal to zero). Thus, magnetic monopoles having no correlate of opposite sign do not exist. Remembering that every source-free vector field may be expressed mathematically as a rotation field (and vice versa), it is possible to derive the magnetic field strength $\boldsymbol{H}$ as a rotation field from a vector field $\boldsymbol{A}$, which is usually called the vector potential of $\boldsymbol{H}$

$$\boldsymbol{H} = c \cdot \text{curl } \boldsymbol{A}. \qquad (5)$$

The fundamental law of the conservation of charge results from Maxwell's equations in the form of the continuity equation

$$\dot{\rho} + \text{div } \boldsymbol{i} = 0, \qquad (6)$$

in which ρ is the time derivative of the charge density, and the vector field $\boldsymbol{i}$ is the electric current density.

In the case of a vacuum, the Maxwell equation that expresses Faraday's law of induction takes the form of a proportionality between the rotation field of the electric field strength and the time derivative of the magnetic field strength.

$$\dot{\boldsymbol{H}} = c \cdot \text{curl } \boldsymbol{E}. \qquad (7)$$

Matter and action. It is a significant fact that Maxwell's theory leads to a localization of energy, which in electromagnetic fields is propagated somewhat in the manner of a substance, with a density that, for the vacuum case, is

$$W = (\boldsymbol{E}^2 + \boldsymbol{H}^2)8\pi. \qquad (8)$$

There remains also the unsolved problem of clarifying the relation of gravitation to quantum theory, which is

much aggravated by the fact that gravitational energy allows of no similar localization.

In both mechanics and electrodynamics, the fundamental equations have such a form that they can be understood as the conditions for a variational or an extremal principle: that, through the fulfillment of these conditions, a certain integral receives an extreme value. This integral, which has the dimensions of action—*i.e.*, of energy times time—is one of the most fundamental quantities of nature. Although the concept of action is less obvious to man's physical intuition than that of energy, it is of even greater significance, as it appears also in connection with the quantum laws. For the basic constant of all of quantum physics, which always occurs in the laws of this domain, is likewise of this dimension: namely, Planck's quantum of action

$$h = 6.626 \times 10^{-27} \text{ erg-seconds.} \quad (9)$$

Modalities of the natural order. In a purely phenomenalist theory of matter—*i.e.*, a theory that does not go into the details of atomic physics but considers matter only in a first approximation as a spatially extended continuum—numerous material properties are ascribed to every type of matter, properties such as density, electrical conductivity, magnetizability, dielectric constant, thermal conductivity, and specific heat. To be complete, a theory must provide a means of deriving all of these material properties theoretically from the laws of atomic physics.

Causality and statistical approach

The hiatus-free causality envisioned throughout the science of physics before the rise of quantum theory cannot be separated conceptually from the far-reaching assumption that all physical processes are continuous. It had been supposed that continuous changes in antecedent causal processes would issue in continuous changes in the sequence of processes that are causally dependent upon them. Quantum physics, however, has expressly breached the old philosophical axiom that *natura non facit saltūs* ("nature does not make leaps") and has introduced a granularity not only in the matter filling space, but also in the finest processes of nature. It is therefore only logical that, with respect to causality, the quantum theory would arrive at new and modified ideas as well. Renouncing unbroken causality, it speaks only of a probability that is statistical and a predetermination for the discrete saltatory events of which physical processes consist—a view that must now, in spite of Einstein, be regarded as irrevocable.

The special theory of relativity demands that the fundamental validity of the local-action principle be acknowledged: all actions have only finite velocities of propagation, which cannot exceed the velocity of light. Thus, in relativistic cosmology it is quite possible that two partial regions of the total spatial manifold may exist between which no causal interaction can occur: causal influences could then assert themselves only inside the so-called interdependent regions in the space–time manifold. These remarks also apply to the quantum theory, in which, however, instead of a causal dependence of physical processes upon each other, there is only an induction of statistical probabilities for possible quantum transitions.

Levels of the natural order. Moving in quite different directions, the theory of relativity on the one hand and the quantum theory on the other have diverged far from the earlier basic ideas of classical physics, which were considered unalterable. There are also some physical problems, however, that can be thought through only by appealing to both the relativistic and the quantum-theoretical modifications. A so-called joint relativistic and quantum-mechanical theory suitable for such problems is quantum electrodynamics, the development of which, however, is not yet complete. Its development was greatly hindered at first by certain mathematical difficulties (so-called divergences), which it later became possible to mitigate by renormalization; *i.e.*, by a technique of correcting the calculated results. The more generally conceived quantum theory of wave fields finds a broad area of possible application in the physics of the different kinds of elementary, though short-lived, particles produced by the huge high-energy accelerators. In its final form, the theory of elementary particles should not only formulate, in general, the laws valid for all known elementary particles but should also allow a deductive derivation for all possible kinds of elementary particles—analogous to the derivations of elements in the periodic table. Heisenberg has endeavoured to set up this far-reaching problem, which has been called the world formula, for a solution. Imposing mathematical difficulties, however, have arisen in the attempt to clarify its consequences for a quantitative comparison with experience, and considerable further work may still be required.

SPECIAL PROBLEMS IN THE PHILOSOPHY OF PHYSICS

Special topological properties

Problems at the formal level. Euclidean space, in contrast to imaginable spatial structures that deviate from it, is distinguished by the simplicity of the topological properties (those preserved through rubberlike distortions) that arise from its unusually simple continuity relationships. One may ask, then, whether the empirical knowledge of modern physics gives any cause to consider deviations from the topological relations of Euclidean space. The U.S. physicist John A. Wheeler, author of a new theory of physics called geometrodynamics, has speculated about this question. In particular, he has pointed to the possibility of so-called worm holes in space, analogous to the way in which the cylindrical surface of a smooth tree trunk is changed topologically if a worm bores a hole into the trunk and emerges from it again elsewhere: the surface of the trunk has thus obtained a "handle." Similarly, one can envision certain handles being added to three-dimensional Euclidean space. Whether this hypothesis can be fruitful for the theory of elementary particles is yet to be determined. From the methodological and epistemological standpoints, it is obvious that a geometrical structure is here being assumed, the measurement of which is fundamentally hindered by the lack of rulers with calibrations smaller than the structure itself. Presumably, the practical possibility of appealing to such topological modifications of the ordinary notion of space is to be found in astrophysics rather than in elementary particle physics. Viktor A. Ambartsumian, a Soviet astrophysicist, is convinced that the processes involved in the origins of galaxies are connected with explosions in which the matter of new stellar systems arises from prestellar material; and it has been found tempting to suppose that this prestellar material exists in regions with unusual topological properties.

Lorentz transformations

The basic idea of the special theory of relativity can also be understood as a statement about the symmetry properties of the four-dimensional space–time manifold. The special principle of relativity states, in fact, that the same physical laws are valid in all of the various inertial coordinate systems—in particular the law that the velocity of light in a vacuum always has the value c. This equivalence of the space–time coordinates x,y,z,t with other coordinates x', y', z', t' that are linear, homogeneous functions of the unprimed coordinates can be expressed by the equation

$$x^2 + y^2 + z^2 - c^2t^2 = x'^2 + y'^2 + z'^2 - c^2t'^2. \quad (10)$$

In this formulation, the isotropy of space—its sameness in all directions—appears as a special case of a more comprehensive symmetry property of the space–time manifold. When $t = t'$, the special case of a purely spatial rotation of coordinates is obtained; and in the general case, in which the primed coordinates are moving with velocity u with respect to the unprimed, the famous Lorentz transformations are obtained, which, to adjust to the finiteness of c, add a factor, symbolized by the Greek letter gamma, $\gamma = \sqrt{1 - \frac{u^2}{c^2}}$, to the ordinary Galilean transformation, $x = x' + ut'$, thus yielding

$$x = (x' + ut')\gamma \quad (11)$$

(see RELATIVITY: *Relativity of space and time*). The groups of symmetries of the four-dimensional space–time manifold thus produced are called Poincaré groups.

Problems at the quantum level. Problems of particle theory, complementarity, and symmetry have arisen in studies at the quantum level.

Elementary particle theory. Whereas the atomic nuclei beyond hydrogen-1 (the proton) are compounded structures, consisting of neutrons and protons, modern physics also deals with numerous elementary particles—neutrinos; π (pi), μ (mu), and K mesons; hyperons; etc.—that are thought of as uncompounded. The elementary particles of each particular kind show no individual differences. Except for the case of light quanta, for each kind of particle a corresponding antiparticle has come to light (which, for charged particles, always carries a charge of opposite sign). For instance, to each of the two known forms of neutrino there corresponds an antineutrino with charge zero and zero rest mass; *i.e.*, with a mass that would be zero were it not moving at a velocity of the order of (in fact, at) that of light. Whether Heisenberg's world formula can provide a complete framework for all possible kinds of elementary particles is still undecided.

Every type of elementary particle has a definite value for its spin, which is either integral (*e.g.*, light quanta) or half-integral (*e.g.*, electrons, protons, neutrinos). Particles with half-integral spin obey Fermi–Dirac statistics; those with integral spin obey Bose–Einstein statistics, which differ in form as $u/(1+u)$ differs from $u/(1-u)$—u being any function. The conformity to law that underlies the Fermi–Dirac statistics for electrons was first recognized by Wolfgang Pauli, an Austrian-Swiss physicist, and formulated as the Pauli exclusion principle, which played a decisive role in settling upon the shell structure in the periodic system of the elements.

Waves and corpuscles

Heisenberg's uncertainty principle. The basic duality of waves and corpuscles is of universal significance for all kinds of elementary particles, even for composite particles in those experiments that cannot lead to a breakup of the particles into their component parts.

An electron (and analogously any other elementary particle or even, for example, an alpha particle) can appear just as well in the form of a wave as in that of a localized corpuscle. In an idealized thought experiment, one can imagine that the position of an electron can be ascertained with a gamma-ray microscope. If the electron is described in terms of wave processes in the sense of Schrödinger's wave mechanics, a very sharply concentrated wave packet appears at the stated position. In an investigation of this packet by Fourier analysis—a technique that analyzes a function into its sinusoidal components—wave components of quite different wavelengths occur; thus an electron in this condition has no definite value for its de Broglie wavelength and consequently none also for its translational momentum. Then, as stated in the so-called de Broglie relation,

$$\lambda = h/mv, \qquad (12)$$

there is for an electron moving free from impinging forces a corresponding wavelength, symbolized by the Greek letter lambda, λ, that is inversely proportional to its momentum mv. And conversely, an electron moving inertially with a definite momentum (which in the limiting case of small velocities is equal to the product of the mass and the velocity vector) has no definite position. If an electron that is moving inertially (especially an electron at rest) is constrained by the use of a gamma-ray microscope to "make up its mind," as it were, on a location, then the probability of its appearance at a point in space is the same for all locations. More precisely stated, the probability of the appearance of the electron in a definite volume is proportional to the magnitude of this volume.

In particular, it will be helpful to consider an electron moving in the x-direction and to suppose that it has a wave amplitude that depends only upon x, an electron in which the most representative wavelengths are confined to a narrow interval while the amplitudes that are discernibly different from zero are likewise confined to a certain interval Δx. If, on the other hand, Δp is the range of discernible momentum values—computed from the discernible wavelengths that represent them according to the de Broglie relation (12)—then the product of the uncertainties Δx and Δp cannot be smaller than Planck's fundamental quantum of action h. This statement comprises the famous Heisenberg uncertainty relation, which expresses the "complementarity" of position and momentum—as Niels Bohr characterized it.

If one assumes, as above, that all physically possible states of an electron can be represented by Schrödinger's wave mechanics, then the complementarity of the position coordinate x and its corresponding conjugate momentum p_x is a simple mathematical fact. When one thinks primarily of physical-measurement experiments, it should be emphasized that stringent limitations are imposed on the simultaneous measurement of the position and momentum of a particle that, according to the uncertainty principle, make it impossible to measure simultaneously both of these complementary quantities with unlimited precision. In an experiment that measures its position, the electron is forced into a sharper localization; and its particle nature comes to light. By contrast, in an experiment that measures its momentum, an interference experiment is involved; the electron must be able to display a certain wavelength, which requires an adequately extended region in space for its reacting. These two complementary and opposing demands can only be brought into harmony in the sense of a compromise; and the Heisenberg uncertainty relation formulates the best possible compromise.

Thus, it becomes at the same time clear that the state of the electron given in the wave-mechanical description before carrying out a new measurement experiment can establish only a statistical prediction for the result. The probability density for the appearance of an electron at a point in space is given by the square of the absolute value of the (complex) Schrödinger wave amplitude; for a definite result in measuring a wavelength or a momentum, the square of the absolute value of the (complex) Fourier coefficient belonging to it provides the standard. The general statistical transformation theory of quantum mechanics (as developed by Dirac and Jordan) gives a complete review of the measurable physical quantities for a microscopic mass point (or a system of such points). According to this theory, two different measurable quantities A and B can be simultaneously determined with unlimited precision only if the operators or matrices that describe A and B commute; *i.e.*, if $AB = BA$.

Charge/parity/time symmetry. A transformation in which the nucleus emits an electron and a neutrino is called beta decay, an example of nuclear radioactivity (*q.v.*). The forces that thus come to light are those of the so-called weak interactions. It has been experimentally determined that for these forces the symmetry associated with reflections in a mirror does not hold. At least in certain circumstances, however, a remnant of this symmetry continues to hold, in which the so-called CPT (for the initials in charge/parity/time) theorem applies. This theorem states that basic physical laws remain of unaltered validity when a reflection of the space coordinates as in a mirror is combined with an interchange of positive and negative charge (which is largely synonymous with the interchange of particles and antiparticles) and with a reversal of the direction of time. Whether or not this symmetry law is valid without exception is by no means fully clarified at present (see CONSERVATION LAW AND SYMMETRY.)

Weak interactions

Problems at the macrophysical level. Proceeding from the properties of atoms and molecules that are described in terms of quantum theory, a theory of macrophysical substance aggregates has been built using statistical mechanics. The theories of heat, of gases, and of solid-state aggregates (crystal lattices) have been extensively clarified. Only the liquid state still poses certain unsolved problems for the statistical theory of heat.

Mach's principle of cosmic influence. In any case, Newtonian mechanics may be derived as a macroscopic consequence of the laws of the mechanics of atoms, and its validity for the motions of astronomical bodies presents no problem. It is not so simple to prove, however, that the statements of Newtonian mechanics for rotating bodies (*i.e.*, the mechanical laws of centrifugal force and the Coriolis force) may be established from Einstein's general theory of relativity. Ernst Mach, a physicist and

philosopher of science whose train of thought has substantially fertilized the modern development of physics from the point of view of the theory of knowledge, raised objections against Newton's idea that centrifugal force is a consequence of the absolute rotation of a body; he asserted instead that the rotation of a body relative to the very distant giant mass of the universe was the true cause of centrifugal force. This idea, often referred to as Mach's principle, has been corroborated, though in a different form, within the conceptual framework of Einstein's general theory. An irrotational coordinate system —specifically, a system not rotating with respect to the fixed stars (or, better, to the spiral nebulae)—is distinguished from a rotating system by the difference in the metric field for the two cases (*i.e.*, in the properties of their respective space–times as expressed by the equation —(3) above—for the interval between two events). It is true that the local metric field (in the vicinity of the solar system) is influenced by the distant masses of the universe, but of course only in the sense of a local-action principle and therefore in no way such that the metric in the solar system is directly given as a function of these distant masses and of their motions.

The question of the precise circumstances in which Mach's principle can still be defended on the basis of Einstein's theory is somewhat complicated and thus remains obscure. In any case, it is certain that a deduction of this principle from Einstein's theory can only be given in conjunction with a complete solution of the cosmological problem; *i.e.*, of the problem of what are the overall geometric and dynamic properties of the universe considered in its totality. The remaining problems involved in justifying the application of classical Newtonian mechanics in astronomy by means of Einstein's theory contain, however, no additional fundamental difficulties.

Entropy and the direction of time. There is one more influence of cosmological relationships upon macroscopic physics, which arises in connection with thermodynamics. The existence of irreversible processes in thermodynamics indicates a distinction between the positive and negative directions in time. As Clausius recognized in the 19th century, this irreversibility reflects a quantity, first defined by him, called entropy, which measures the degree of randomness evolving from all physical processes by which their energies tend to degrade into heat. Entropy can only increase in the positive direction of time. In fact, the increase in entropy during a process is a measure of the irreversibility of that process. In contrast, it is true of the quantum theory of the atom that the positive and negative directions in time are equally justifiable (in the sense of the principle of CPT symmetry). Consequently, it is difficult to understand how statistical mechanics can make possible a thermodynamics in which the entropy grows with time.

It is true that there are fluctuating thermodynamic phenomena, even in a system in overall thermodynamic equilibrium—and here theory and experiment agree. Thus, the states that arise within any small partial volume of the system may be not only those that are thermodynamically most probable but also transitory deviations from the most probable state. The mention of these fluctuations, however, does not help to remove the above paradox.

Most physicists now hold that, until recently, this problem was treated erroneously in the usual textbook presentations. In the statistical theory of heat, entropy was regarded as proportional to the logarithm of the thermodynamic probability, and students came to regard it as a necessity of thought that nature progresses from states of lower probability to states of higher probability. In truth, however, the increase of entropy is a real physical property of the positive direction in time. Nonetheless, it was supposed that shaking a vessel containing red and white balls (or even grains of sand) in an originally ordered condition with the two colours neatly separated must result in a condition of extreme intermixing of balls. This result does not correspond, however, to some necessity of thought but to an empirical property of the real universe in which men live and experiment.

Boltzmann's theorem

This interpretation, which was held for a long time and has only quite recently been recognized as erroneous, was allegedly supported by a famous mathematical theorem of Boltzmann, which seemed to show that, in an ideal gas for which the entropy—measured by its number of particles and its total energy—was not yet at its maximum value, the entropy must increase. If collisions of gas molecules are characterized by velocity vectors that are mechanically allowable (both before and after the collision), and if these vectors must satisfy both energy and momentum conservation, then what Boltzmann actually proved is that the entropy increase follows only when a correct count of the collision rate is made, according to which every kind of collision of gas molecules has a frequency of occurrence proportional to the product of the number of collision pairs that were present and the velocities that existed before the impact.

As one can subsequently see, the positive direction in time is already marked out by the collision rate count in a manner that no longer corresponds to the CPT principle. Although it is in fact possible to reason out the continuous increase in entropy on this basis, the paradox is not overcome. The question then remains of how it is physically justifiable—*i.e.*, how it can correspond to reality—to regard this principle of collision rates as valid even though it fundamentally contradicts the CPT principle.

An answer to this paradox can now be given, thanks to the insight of modern theoreticians—among them Hermann Bondi, a mathematician and cosmologist—who have shown that the entropy principle must be understood in the sense that in the universe as a whole, one definite time direction is singled out, namely, the one for which the universe expands. The thermodynamic distinction of a positive direction in time—with increasing entropy on the macroscopic level and with collision rate counts on the microscopic level—results from an expansion of the universe. Surprisingly, the Hubble expansion of the system of all the galaxies—so named after Edwin Hubble, an extragalactic astronomer—thus displays physical effects right down to the level of everyday physics; specifically, when two bodies at different temperatures are brought into thermal contact, the temperature equalization that results is an irreversible process corresponding to an asymmetry of the positive and negative directions in time that depends upon the expansion of the universe.

Problems at the cosmological level. A mathematical discovery by Alexander Friedmann, a Russian meteorologist, has become of great significance for the mathematical derivation of cosmological models from Einstein's general theory of relativity. According to Friedmann, if the average mass density is constant throughout space, the gravitational field equations can be satisfied by a metric that embraces a three-dimensional space of constant curvature together with a time coordinate t such that the radius of curvature $R(t)$ is a definite function of time; and these cosmologies turn out differently depending upon whether the curvature of space is positive, negative, or zero. Among the models of the universe that are mathematically allowable are models in which the time coordinate may run through all values from zero to infinity, models in which the time is limited to a finite interval, and models in which it may run from minus infinity to plus infinity.

The steady-state theory

For a time, many specialists working in the field of cosmology found the so-called steady-state theory, first projected by an astronomer, Sir Fred Hoyle, especially convincing. In a modified version, this theory was adapted to the Friedmann model by Bondi. By adopting the so-called perfect cosmological principle, which holds that the broadest features of the universe are the same at all times as well as at all places, the theory then satisfied the unusually high symmetry or homogeneity requirements not only of a three-dimensional space with constant time but also of the entire space–time manifold. This high-degree homogeneity was so convincing to many authors that, in deference to it, a fundamental deviation from Einstein's field equations was tolerated: Bondi and Hoyle supposed that a small but constant creation of hydrogen

occurs in the intergalactic vacuum. This hypothesis was introduced in order to achieve, in spite of the Hubble expansion of space, a mass density that remained constant in the universe.

This theory, which in spite of its deviation from Einstein's field equations certainly advocates an allowable hypothesis worthy of consideration, no longer seems tenable, however, because of the discovery of background radiation with a present temperature of 3° Kelvin, which is interpreted as a remnant of an original "big-bang" beginning of the universe. It thus appears that it is no longer possible to uphold the steady-state theory or the perfect cosmological principle upon which it is based. Instead, one must favour either a Friedmann model, which has a beginning, from which it expands monotonically and without limit; or a Lemaître model, in which a quantity lambda, λ, called the cosmical constant, arises that is, mathematically, a constant of integration, and physically, a force of cosmic repulsion that partially neutralizes that of gravitational attraction, and which lends a curvature to space even in its empty regions. For both of these models the time coordinate increases without limit from some initial value, which would naturally be called zero. For the beginning of time, one thinks, moreover, of a singularity $R(0) = 0$ and thus of a space that at the null point of time is still a mass point. Cyclical models that alternately expand and contract in an endless sequence have also been discussed.

The empirical cosmological data, some of which, indeed, are more estimated than ascertained, seem to suggest that, in the present-day universe, the positive energy corresponding to the total rest mass of all the material existing in the universe may be exactly equal to the negative gravitational energy existing in the universe; thus, the total energy would then be equal to zero. This interesting singularity, however, needs further support. At one time, Dirac advocated the speculation that the total mass of the universe is not constant in time but is increasing—at a rate somewhat slower, however, than that in the steady-state theory. Ambartsumian's notion concerning prestellar material, which was mentioned above (see *Problems at the formal level*), could perhaps be considered support for this idea, which Dirac himself pursued only temporarily. Many further discussions have followed another conjecture by Dirac, according to which the gravitational constant G should be liable to change in the course of cosmic development. This constant would thus have to be considered a scalar field quantity, which in a Friedmann universe is approximately independent of the three space variables but dependent on the time variable. In spite of extensive theoretical deliberations on this theme, no decision has yet been reached.

The way has been opened for some fundamental conjectures on certain emerging themes by the fact that the product of the mean mass density in the universe and the gravitational constant has the same order of magnitude as the square of the reciprocal of the radius of curvature of the universe. The aforementioned relation between the mass and gravitational energy in the universe presents a different expression for this ratio. The total mass of the universe divided by the proton mass probably has approximately the order of magnitude 10^{80}, according to present cosmological notions. The order of magnitude of the radius of curvature of the universe is approximately 10^{40}, when expressed as a multiple of an elementary length of which the order of magnitude is approximately that of the nuclear radius. Whether it is justifiable to presume that there is here a functional dependence—*i.e.*, a proportionality of M to R squared—is a question for the present still undecidable. The speculative attempt of Dirac to find an answer, however, is still—at least provisionally—judged with skepticism by the majority of physicists. (P.W.J.)

II. Philosophy of biology

The sharp increase in man's understanding of biological processes that has occurred in recent years has stimulated philosophical interest in biology to an extent unprecedented since the development of evolutionary theory in the 19th century. Biologists and philosophers alike have devoted much attention to a variety of issues regarding the subject matter and the methodology of biology, resulting in a sizable output of written material, formulating philosophical questions that are still arising and framing answers to acknowledged difficulties. Most of the problems of the philosophy of biology are old questions now being investigated afresh in the light of biological advances and new standards of philosophical rigour. In this account contemporary questions will be stressed (for historical background, see BIOLOGICAL SCIENCES: *The philosophy of biology*).

THE RANGE OF TOPICS

An investigation of recent writings in biophilosophy reveals a continued preoccupation with unanswered—some say unanswerable—questions about evolutionary theory and a growing concern for a critical reappraisal of the question of whether biology is an autonomous discipline unamenable to reduction to mere physical and chemical underpinnings. Until the mid-20th century the biological sciences suffered from a lack of attention by philosophers; the principles that were generated were far less rigorously examined than were those of the physical sciences. There is now renewed hope, however, for a fresh approach to the age-old puzzles regarding life and its raison d'être. This hope rests on the recrudescence of interest in all biological matters as a direct result of an increased understanding of biological processes, of the changing quality of life, of the growing awareness of man's stewardship of the Earth, and of the exploration of space. Biology has just begun to make the sort of impact that the physical sciences have already made. It has generated a life technology with genetic engineering, organ transplants, and artificial organs. Each innovation, each technical masterstroke, each conceptual knot untied emphasizes the need for a definitive philosophy of biology, and developments toward this goal are now under way. Good biological work has been accomplished by investigators with varied philosophical outlooks ranging from Neo-Thomism to skeptical naturalism. No inevitable metaphysics evolves from the study of biology or any other natural science; nevertheless, some of the general conclusions of biology have a philosophical interest, defining the limits of reasonable belief about the nature of the living world.

Natural distinctions

Categorical discontinuities that are recognized for the purpose of scientific methodology often seem impossible to justify as "natural" distinctions. Many biologists have noted, for example, that it is easier to study life than it is to define it. Properties such as metabolism and reproduction undeniably characterize organisms and might be said to define them, yet such a definition is arbitrary to the extent that such properties are logically independent. What is true of all life forms today may not have been true of the very earliest ones and, what is more, might not be true of extra-terrestrial ones that might be encountered in the future. There is not as yet a set of nonarbitrary characteristics that mark the distinction between living and nonliving systems. Moreover, in the course of analysis, it becomes necessary to arrange all of the phenomena of nature in a more or less linear, continuous sequence of classes and then to describe events occurring in the class of more complex phenomena in terms of events in the classes of less complex phenomena (principle of hierarchical continuity). Within each class, however, there are numerous interrelations observed between events of the same order of complexity. It is thus possible to recognize a number of more or less autonomous disciplines, each permitting generalization, but ordered so that the more complex events treated by one discipline can also be analyzed in terms of less complex events treated by another discipline. It is possible, for example, to establish a body of generalizations about human society independent of the behaviour of individual persons; a number of generalizations about individual behaviour without consideration of the physiology of the sensory, conductor, and effector mechanisms involved; and a large body of generalizations about muscle or nerve physiology

Living and nonliving systems

without considering the molecular mechanisms involved. A particularly striking feature of the hierarchy is that an increase in complexity is coupled with the emergence of new characteristics. The origin and development of life from small systems that synthesized biochemicals to organisms that perform highly complicated functions suggests that the hierarchical arrangement of nature and the sciences is correlated with the temporal order of evolution. The maintenance of a steady state by metabolism, reproduction, responsiveness, modification of response by experience, tradition, and social phenomena are just some of the more dramatic examples of emergent phenomena. The emergence of new qualities as evolution proceeds might generally characterize the universe.

Moreover, photosynthesis, on the one hand, and reproduction followed by natural selection, on the other, provide a mechanism by which physically less probable systems can emerge locally from physically more probable ones. Though it frequently has been supposed that physical evolution is at an end, there is no reason to suppose that this is true of social development, for which Sir Julian Huxley, a biologist, philosopher, and educator, provided an evolutionary context. In his Romanes Lectures, published in 1943, Huxley wrote:

> It is only through social evolution that the world-stuff can now realize radically new possibilities. Mechanical interaction and natural selection still operate, but have become of secondary importance. For good or evil, the mechanism of evolution has in the main been transferred [in man] onto the social or conscious level. . . . The slow methods of variation and heredity are outstripped by the speedier processes of acquiring and transmitting experience. . . .
>
> And in so far as the mechanism of evolution ceases to be blind and automatic and becomes conscious, ethics can be injected into the evolutionary process. Before man that process was merely amoral. After his emergence onto life's stage it became possible to introduce faith, courage, love of truth, goodness—in a word moral purpose—into evolution. It became possible, but the possibility has been and is often unrealized.

It may well be that social evolution is only in its early stages. These stages, moreover, have for the most part taken place in a period during which systematic knowledge was undeveloped. A Russian mineralogist, Vladimir Vernadsky, the founder of biogeochemistry, regarded the envelope of the Earth as passing from a stage determined primarily by biological processes to one determined by conscious human effort. He called this layer of consciousness the noösphere. The concept was later extended, notably by Pierre Teilhard de Chardin, a French paleontologist, who began the building of a new philosophic bridge between biology and religion.

THE NATURE OF BIOLOGICAL SYSTEMS

Questions about the character of biological systems and of the biological world merge with those about the concepts and methods required for their understanding. Although the two cannot be wholly separated, in this account those matters most clearly related to the substantive philosophical aspects of biology will be stressed. Methodological issues will be touched on in respect to the concepts being elucidated.

Origin and definition of life. Space exploration has directly influenced the development of life-detecting devices. This technological need spurred intensive study regarding the kinds of evidence living things display reflecting their aliveness. In his *Chance and Necessity* (1972) Jacques Monod, a biologist, deals with the invariance of genetic endowment, morphological autonomy, reproductive invariance, and teleonomy (the tendency to have a purpose or project written into their molecules) as the major properties of living systems; he considers that they involve the chance and necessity that determine the course and character of the entire biological world.

Philosophic views

Philosophers have long deliberated over the definitive features of living systems. The distinction between living and nonliving, which was widely discussed at the turn of the 20th century, has lost much of its interest for current biology. A growing conviction, intuitively felt by many biologists, is that no clear line can be drawn between the living and the nonliving. The bridge between what is and is not obviously alive consists of a range of problematic agents, including viruses and genes, which appear to be living at times and nonliving at other times.

Viewpoints on the nature of life. Basically and traditionally, there are three distinct philosophical stands regarding the biological nature of life: vitalism, mechanism, and organicism.

Vitalism. Essentially, vitalism holds that there exists in all living things an intrinsic factor—elusive, inestimable, and unmeasurable—that activates life. In its classic form, as espoused by many biologists at the turn of the 20th century—in particular, by Hans Driesch, a German biologist and philosopher—it has suffered severe criticism. Ernest Nagel, a philosopher of science, rang its death knell in 1951, when he wrote in *Philosophy and Phenomenological Research* (11:327 ff.):

> Vitalism of the substantival type . . . is now a dead issue . . . less, perhaps, because of the methodological and philosophical criticism that has been leveled against the doctrine than because of the infertility of vitalism as a guide in biological research and because of the superior heuristic value of alternative approaches.

And whereas most biologists concur in renouncing this so-called naïve vitalism, some continue to espouse a so-called critical vitalism, perhaps indistinguishable from organicism (see below).

Mechanism. Simply stated, the view of the mechanists is that organisms are no different from subtle machines: the whole is the sum of its parts, which are arranged in such a way that an internal energy source can move them in accordance with a built-in program of purposeful action. In the mechanist's view, advances in molecular biology corroborate this claim and demonstrate that in principle organisms are no more than complicated physical systems. This is, in essence, the reductionist position, which states that biological principles can be reduced to physical and chemical laws. Antireductionists, of course, contend that molecular biology cannot explain all aspects of living forms.

It has often been said that, whereas biologists may think as vitalists—and hold the conviction that organisms are more than just complex machines—they perforce become practicing mechanists in the laboratory, required by the demands of scientific inquiry to view their experiments in terms of the measurable parameters of physics and chemistry. K.F. Schaffner, an American philosopher, suggested in 1967 that, even though reductionism may be correct, a better strategy may be to strive toward an independent biology.

Organicism. The basic claim of organicism is that organisms must be interpreted as functioning wholes and cannot be understood by means of physics and chemistry alone. Few scientists today call themselves organismic biologists or endorse the doctrines put forward by such organismic theorists as Ludwig von Bertalanffy and Edward Stuart Russell. Nevertheless, most antireductionists subscribe at least to part of the organismic doctrine, in particular to its wholistic claim. Russell, a foremost proponent of organicism, stated in his work *The Interpretation of Development and Heredity* (1930):

The views of E.S. Russell

> Any action of the whole organism would appear then to be susceptible of analysis to an indefinite degree—and this is in general the aim of the physiologist, to analyze, to decompose into their elementary processes the broad activities and functions of the organism. But . . . by such a procedure something is lost, for the action of the whole has a certain unifiedness and completeness which is left out of account in the process of analysis. . . . In our conception of the organism we must . . . take account of the unifiedness and wholeness of its activities . . . since . . . the activities of the organism all have reference to one or other of three great ends, and that both the past and the future enter into their determination. . . . Bio-chemistry studies essentially the *conditions* of action of cells and organisms, while organismal biology attempts to study the actual modes of action of whole organisms, regarded as conditioned by, but irreducible to, the modes of action of lower unities.

In some special sense, then, an organism is regarded as being more than a simple sum of its parts; an additional

"something" has accrued to it as a result of the unique arrangement of its components. As Morton Beckner, a philosopher of biology, asserted in an article in the *Encyclopedia of Philosophy* (5:549):

> In the history of biology it is difficult to disentangle vitalistic and organismic strands, since both schools are concerned with the same sorts of problems and speak the same sort of language. The distinction between them was drawn clearly only in the twentieth century. Organismic biology may be described as an attempt to achieve the aims of the murky organismic-vitalistic tradition, without appeal to vital entities.

Further (p. 551):

> Organismic biology is to be interpreted as a series of methodological proposals, based on certain very general features of the organism—namely, the existence in the organism of levels of organization with the biological ends of maintenance and reproduction. These features are sufficient to justify "a free, autonomous biology, with concepts and laws of its own," whether or not the higher levels are ultimately reducible to the lower ones.

Organisms as systems. The concept of an organism as a cybernetic, or automatic-control, system is currently influential in biology.

Holism and emergence. The holistic concept of an organism—*i.e.*, the theory that the determining factors in biology are its irreducible wholes—owes its success primarily to the existence of control and regulation mechanisms operating at the molecular level that determine development and behaviour. The character of such systems at all levels of analysis—molecular through total organism—is nothing other than a sophisticated kind of cybernetics. Holism and reductionism are similar in this respect. Closely allied to organicism is the old problem of emergent properties dealt with earlier: at each successive level of organization, qualities emerge that cannot be anticipated by the components and that confer an added dimension to each hierarchical level in the biological world.

General systems theory. A theoretical and methodological program called general systems theory—presented in its fullest and most persuasive form by Bertalanffy—is an extension of the tenets of organismic biology. It is an attempt to provide a common methodological approach for all of the sciences, based upon the idea that systems of any kind—physical, biological, psychological, and social—operate in accordance with the same fundamental principles. Ideally, it should be possible to deduce the principles applying to a particular sort of system from the more general ones. This approach is one still very much in need of development.

Function and purpose in biological systems. Attributions of purpose (teleology) appear frequently in biological writing. Not only do biologists say that parts of organisms have a purpose with respect to the whole, but some hold that life itself is inherently purposive. But the term purpose is both vague and ambiguous. That every biological system—from self-replicating molecules (DNA) to biotic communities—involves specific and identifiable functions is undeniable. But whether, or in what way, functional ends like the reproduction of a cell resemble human intentions or purposes is a matter of some controversy. Even if this matter were settled, a larger question would still remain, viz., whether a biological system as a whole can have a goal that is in some way similar to a human goal—*i.e.*, whether it is programmed with an utlimate purpose. Although resolution of this matter has long been and will continue to be a critical point in the philosophy of biology, much has been done in the last few years to clarify the issues involved.

PHILOSOPHY IN EVOLUTION

Teleology and determinism. Vitalists and those who subscribe to a Lamarckian view of evolution involving the inheritance of acquired characteristics claim that evolution involves a deterministic finalism, or directedness toward an end. Most evolutionary biologists—among them George Gaylord Simpson—reject that claim and hold that natural selection is the non-random element in evolution, that which gives evolution direction. Other evolutionists—among them Theodosius Dobzhansky—argue that the chance factors in mutation and selection, in addition to the unpredictability of environmental change, make it impossible to formulate deterministic laws even in experimental populations, let alone in natural populations. Similar considerations by others have led to the claim that evolutionary biology is a paradigm of an after-the-fact exploratory science and that the course of evolution can never be predicted.

The species problem. Whether biological species can be said to have a real existence in the world is a question that has received much consideration in recent years. The issue may be posed in the words of Benjamin Burma, a paleontologist, who, writing in *Evolution* (3:369), asked:

> What, then, is a species? It would seem thus far to be the whole of any one series of breeding populations. . . . [But the] definition as it stands unfortunately puts all living and fossil animals in one species, since there is a continuity of germ-plasm back from John [an individual animal] to the original primordial cell, and from it forward to every living animal (not to mention plant). Thus, if we ignore time, we end up with only one species. . . .

The temporal difficulty, however, is not the only stumbling block to the question of species reality; for, if the species is redefined as the whole of any one series of breeding populations as it exists at any one time, then there is an infinity of species, since time itself is infinitely divisible. On the basis of these and other objections, some biologists have concluded that species have only a subjective existence merely as convenient labels for arbitrary assemblages and have only a minimum of biological significance. On the other hand, there are proponents of the idea that species have an objective reality. Ernst Mayr, a U.S. evolutionist representing the latter group, has written—also in *Evolution* (3:372):

> In all multidimensional situations an inference has to be made (Simpson, 1943) on the basis of the objective species of the non-dimensional system. The subjectivity of this exmade (Simpson, 1943) on the basis of the objective species concept per se. The species of the local naturalist or of the paleontologist within a given horizon is clearly delimited against other species and can thus be considered as having objective reality.

Although the controversy is confused by semantic difficulties, one of the chief contributions of the philosophy of biology has, in fact, been to separate mere linguistic puzzles from matters of substance. Many taxonomists are guilty of ambiguity of reference; they often fail to distinguish their entities clearly, with the result that there is widespread befuddlement over just what stand is held by whom. The problems are now clearer than they have ever been, and with few exceptions biologists and philosophers tend to agree about the nature of biological species and the definition of the species category (see also SPECIES AND SPECIATION).

Evolutionary theory. Although most of the issues connected with evolution as a theory are methodological ones, two issues go beyond the limits of logic. Some philosophers have tried to demonstrate, for example, that evolutionary theory is circular and offers no real understanding of the process of evolution. Others have argued that the notions of types of organisms must be used to understand evolution and that evolutionary change takes place when a new type emerges.

The views of Simpson and Schindewolf

Two clear viewpoints regarding evolutionary theory have come to the fore since 1950. One is expressed in detail by George Gaylord Simpson, in his work *The Major Features of Evolution* (1953), and the other is put forward by a paleontologist, Otto Schindewolf, in his *Grundfragen der Paläontologie* (1950). In 1959, Marjorie Grene, a philosopher of biology, writing in the *British Journal of the Philosophy of Science* (9:11 ff.), summarized their positions as follows:

> Professor Simpson is the principal American spokesman of neo-Darwinism. . . . He sees evolution as a continuous series of minute changes in innumerable directions, in which all alterations of any significance, larger as well as smaller, quicker as well as slower, are determined by the great cooperating "pressures" of mutation, geographical isolation, and

selection, with adaptation as the universal effect, and criterion, of systematic change. The basic concept, ultimately is variation in the occurrence of genes; out of such variations all the systematic relations of living things have been gradually evolved.

Schindewolf's principles are simpler. He sees typical shapes, and sees again and again what appear to be new shapes. Therefore he assumes that living things are able to originate novel types. Mutation, he agrees, must have been the mechanism by which they originated; but the adaptive control of mutation occurs only within, not between types. The basic pattern is of change from type to type, and always, as we have seen, with the more general appearing before its specialised subdivisions.

The controversy between these two opposing viewpoints is a complex one filled with both philosophical and scientific issues. In the opinion of most biologists, Schindewolf's view is persuasive only with respect to the paleontological evidence and is not supported by the experimental study of evolution in current organisms. Most of them thus tend to accept the synthetic theory in more or less the form expressed by Simpson. It remains possible, however, that the process that Schindewolf is talking about is fundamentally different from that explored in population genetics and that typostrophic mutations are so rare, on the time scale of man, as to be beyond hope of detection in the laboratory.

EVOLUTION AS A WORLD VIEW

Very few attempts have been made in recent years to employ the concept of evolution as a scheme for viewing all knowledge and experience. Sir Julian Huxley, who remains one of the best representatives of such an effort, continues to claim that the entire universe is in a process of evolution, which, however, has different aspects, viz., physical, biological, and social. Life and nonlife alike must be understood as part of the process of cosmic evolution, and from this follows a host of metaphysical and ethical implications. The other chief representative of this viewpoint is the evolutionist priest Teilhard de Chardin, who has woven into the fabric of cosmic evolution the panoply of a Christocentric religion that sees the perfection of all things in an "Omega" point toward which evolution is moving.

Metaphysics of the more piecemeal kind—exploring the implications that biological knowledge has on beliefs and attitudes—is fostered by Simpson, a consistent antagonist of Huxley and Chardin. Simpson suggests, for instance, that knowledge of man's origins and of the process that has brought him to his current state in no way threatens belief in his own uniqueness. Man is an animal, but a very special sort of animal. Other matters of a similar kind—purpose in nature and man's evolutionary future—are considerations that constitute the implications of biology in general and of evolutionary biology in particular.

BIOLOGY AND ETHICS

The question of innate aggressiveness. One of the best known issues threatening accepted beliefs about moral responsibility is probably that raised by the proponents of the theory of innate aggression, in particular by such spokesmen as Konrad Lorenz, an Austrian student of animal behaviour, and Robert Ardrey, a U.S. writer. If there is an instinct for aggressiveness, then the notion that it is acceptable to blame individuals and society for outbreaks of violence or war loses its validity. The thrust must then be elsewhere: not in faultfinding but in shoring up against what is felt to be pedestrian and inevitable. As Ardrey puts the theory in his *African Genesis* (1961):

Views of the ethologists

> But we were born of risen apes, not fallen angels, and the apes were armed killers besides. And so what shall we wonder at? Our murders and massacres and missiles, and our irreconcilable regiments? Or our treaties whatever they may be worth; our symphonies however seldom they may be played; our peaceful acres, however frequently they may be converted into battlefields; our dreams however rarely they may be accomplished. The miracle of man is not how far he has sunk but how magnificently he has risen. We are known among the stars by our poems, not our corpses.

Raymond Dart, a South African anatomist and anthropologist, in an article entitled "The Predatory Transition from Ape to Man," published in the *International Anthropological and Linguistic Review* (1:201–208), expressed the thesis of innate depravity on which Ardrey's more popular presentation is based.

Another aspect of the innate aggression inherited from man's primate forebears is militant enthusiasm, which Lorenz described in *Das sogenannte Böse: zur Naturgeschichte der Aggression* (1963; Eng. trans., *On Aggression*, 1966):

> In reality, militant enthusiasm is a specialized form of communal aggression, clearly distinct from and yet functionally related to the more primitive forms of petty individual aggression. Every man of normally strong emotions knows, from his own experience, the subjective phenomena that go hand in hand with the response of militant enthusiasm. A shiver runs down the back and, as more exact observation shows, along the outside of both arms. One soars elated, above all the ties of everyday life, one is ready to abandon all for the call of what, in the moment of this specific emotion, seems to be a sacred duty. All obstacles in its path become unimportant; the instinctive inhibitions against hurting or killing one's fellows lose, unfortunately, much of their power. Rational considerations, criticism, and all reasonable arguments against the behavior dictated by militant enthusiasm are silenced by an amazing reversal of all values, making them appear not only untenable but base and dishonorable. Men may enjoy the feeling of absolute righteousness even while they commit atrocities. Conceptual thought and moral responsibility are at their lowest ebb. As a Ukrainian proverb says: "When the banner is unfurled, all reason is in the trumpet."

Montagu's view

Equally notable opponents of the theory of innate aggression see it much as M.F. Ashley Montagu, a British–U.S. anthropologist, does, as "original sin revisited," and deplore the tendency to neglect authoritative studies in favour of simplistic popularization. In *Man and Aggression* (1968), he writes:

> While the findings of these disciplines [anthropology and the behavioral sciences] are wholly opposed to the deeply entrenched view that man is an innately aggressive creature, most people tend to dismiss these findings out of hand or ridicule them as a rather eccentric idealistic heterodoxy, which do not deserve to become generally known. In preference to examining the scientific findings they choose to cast their lot with such "authorities" as William Golding who, in his novel *Lord of the Flies*, offers a colorful account of the allegedly innate nastiness of human nature, and Robert Ardrey who, in *African Genesis* and more recently in *The Territorial Imperative*, similarly seeks to show that man is an innately aggressive creature. . . .
>
> . . . when through the distorting glass of his prejudgments he looks at a tool it becomes not simply a scraper but a weapon, a knife becomes a dagger, and even a large canine tooth becomes "the natural dagger that is the hallmark of all hunting animals," while in "the armed hunting primate" it becomes "a redundant instrument." "With the advent of the lethal weapon natural selection turned from the armament of the jaw to the armament of the hand." But the teeth are no more an armament than is the hand, and it is entirely to beg the question to call them so. Virtually all the members of the order of primates, other than man, have large canine teeth, and these animals, with the exception of the baboons, are predominantly vegetarians, . . . that such teeth may, on occasion, serve a protective purpose is entirely secondary to their main function, which is to rip and shred the hard outer coverings of plant foods.

Further responses to Ardrey's and Lorenz' thesis are the interpretations of field studies of primate groups, such as those of George Schaller on the gorilla, of Jane van Lawick-Goodall on the chimpanzee, and of Barbara Harrison on the orangutan. These researches suggest that the majority of such groups are singularly free of belligerence. According to Montagu,

> The myth of the ferocity of "wild animals" constitutes one of Western man's supreme rationalizations, for it not only has served to "explain" to him the origins of his own aggressiveness, but also to relieve him of the responsibility for it—for since it is "innate," derived from his early apelike ancestors, he can hardly, so he rationalizes, be blamed for it! And some have gone so far as to add that nothing can be done about it, and that therefore wars and juvenile delinquents, as Mr. Ardrey among others tells us, will always be with us! From one not-so-minor error to another Mr. Ardrey sweeps on to the grand fallacy.

The matter remains moot; but there appears to be a growing consensus that, given a certain genetic constitution—and within the bounds of that endowment—whatever man is, he learns to be, especially in respect to values, morality, and customs. Baser appetitive needs, however, may have a genetic component that is greater than an environmental one.

New understanding of environmental factors and the consequences of man's actions with respect to them has made it clear that man has acquired responsibilities that he did not recognize before. It has become increasingly accepted that standards and values with respect to the environment must be established; this is perhaps the most dramatic case in which recent biological knowledge has generated a crisis of a moral kind. The classic work *Science and Survival* (1966) by a biologist, Barry Commoner, is particularly noteworthy in connecting theoretical and philosophical issues about reductionism and holism to practical matters of environmental understanding and problem solving.

The metaphysical issue of man's place in nature is now being construed as one that requires that man make value decisions, assign responsibilities, and plan for the future of his planet. Environmental problems have become intertwined with problems of social planning, racial tension, transportation and housing crises, genetic engineering, and a host of other current concerns.

Evolutionary ethics. The question of whether nature provides guides to the actions of humankind has held a fascination for many biologists. Those who call themselves evolutionary ethicists say that it does. The defenders of evolutionary ethics contend that external moral standards exist in the facts and process of evolution.

Social Darwinism

Toward the end of the 19th century, Herbert Spencer, in England, and others advanced a series of principles that came to be called Social Darwinism. It espoused such ideas as the inevitability of progress, survival of the fittest, and the struggle for existence, expressions that have become bywords even though they have since been discredited in their original sense, as applied to social phenomena. Social Darwinism as C.H. Waddington, a biologist, explains in his book *The Ethical Animal* (1960), has been superseded by

> . . . the more recent phase of evolutionary ethical thought beginning in the early 1940s, [which] comprises a number of rather different methods of approach. At one extreme we have discussions framed in terms of extremely wide scope, which treat of evolution not only in the animal world but throughout the cosmos, and attempt to relate such broad concepts to man's religious and spiritual life. The pre-eminent example of this tendency in recent years is Teilhard de Chardin, but a rather similar approach can be found in the works of several biologists, such as Conklin, Holmes, and Huxley. The opposite tendency, which of course is also found expressed to various extents in these authors, particularly in Julian Huxley, is the attempt to demonstrate, in a logically coherent argument, a real connection between evolutionary processes and man's ethical feelings.

Some biologists continue to insist, therefore, that biological facts can provide a yardstick by which to measure the morality of a given course of action. Julian Huxley, for one, has long claimed that moral principles can be found in nature and in the evolutionary process in particular:

> When we look at evolution as a whole, we find, among the many directions which it has taken, one which is characterized by introducing the evolving world-stuff to progressively higher levels of organization and so to new possibilities of being, action, and experience. This direction has culminated in the attainment of a state where the world-stuff (now moulded into human shape) finds that it experiences some of the new possibilities as having value in or for themselves; and further that among these it assigns higher and lower degrees of value, the higher values being those which are more intrinsically or more permanently satisfying, or involve a greater degree of perfection.

Huxley further asserts that, although the Golden Rule, the policy of action based on sympathy—doing as you would be done by—may be an immediate good, it ultimately leads to the suppression of those qualities most needed for survival and the continuation of a species. Rather, he argues:

> The facts of nature, as demonstrated in evolution, give us assurance that knowledge, love, beauty, selfless morality, and firm purpose are ethically good. . . . In the broadest possible terms evolutionary ethics must be based on a combination of a few main principles: that it is right to realize ever new possibilities in evolution, notably those which are valued for their own sake; that it is right both to respect human individuality and to encourage its fullest development; that it is right to construct a mechanism for further social evolution which shall satisfy these prior conditions as fully, efficiently, and as rapidly as possible.

Simpson, however, contends, in the article "Biological Sciences," in *The Great Ideas Today* (1965):

> The facts and the processes of evolution are neither ethical nor unethical. The questions of good or bad are simply irrelevant to this field, with the important reservation that evolution has produced a species, *Homo sapiens*, concerned with ethics. Denial of man's naturalistic origin and animal nature is flatly false, and any ethic based on such denial is invalid. Evolution controverts primitive creation myths, but it is consistent with higher values in the Judeo-Christian tradition and those in most now-current religions and philosophical systems. One need only think of the brotherhood of mankind—a biological fact, not only an ethical idea.
>
> Beyond such considerations as those, efforts to combine science and religion may be noble in intention but usually end up distorting or stultifying both. One of the most striking examples at present is the cult, as it may fairly be called, of Pierre Teilhard de Chardin. He preaches necessarily posthumously, for the Roman Catholic Church suppressed his views during his life—a mystical Christianity ostensibly derived from evolutionary principles. But since the mysticism is primary, the evolutionary principles are distorted and downright falsified for seeming coherence with the nonscientific, nonnaturalistic premises. In turn, the mystical views advanced as having that false basis are thereby vitiated. The result (in my opinion) has been a disservice to true religion and to true science.
>
> At the same time, no one can deny the purity of Father Teilhard's intentions or the correctness of his view that evolution and religious feeling should be considered congruent aspects of the nature of man. It is almost as irrational to deny evolution as to deny gravity. The management of life and the goals of aspiration, to be sane, must take account of all such truths of nature. They need not thereby become brutal or earthbound.

(J.R.Mn./R.C.Y.)

BIBLIOGRAPHY

Philosophy of physics: R. HARRE, "Philosophical Aspects of Cosmology," and W. DAVIDSON, "Philosophical Aspects of Cosmology," *Br. J. Phil. Sci.*, 13:104–119, 120–129 (1962); J.D. NORTH, *The Measure of the Universe* (1965); OTHMAR SPANN, *Naturphilosophie*, 2nd ed. (1963); PASCUAL JORDAN, *Albert Einstein: Sein Lebenswerk und die Zukunft der Physik* (1969) and *Atom und Weltall: Einführung in den Gedankeninhalt der modernen Physik*, 2nd ed. (1960); E.T. WHITTAKER, *From Euclid to Eddington: A Study of Conceptions of the External World* (1949); JACQUES MERLEAU-PONTY, *Cosmologie du XX^e siècle* (1965); LOUIS DE BROGLIE, *La Physique nouvelle et les quanta* (1937; Eng. trans., *The Revolution in Physics*, 1953); R.G. COLLINGWOOD, *The Idea of Nature* (1945, reprinted 1960); P.K. FEYERABEND, "Philosophie de la nature," in M.F. SCIACCA (ed.), *Les Grands Courants de la pensée mondiale contemporaine*, pt. 2, vol. 2, pp. 901–927 (1961); A.G.M. VAN MELSEN, "The Philosophy of Nature," in RAYMOND KLIBANSKY (ed.), *Contemporary Philosophy: A Survey*, vol. 2, pp. 151–161 (1968); ERROL E. HARRIS, *The Foundations of Metaphysics in Science* (1965); JAGJIT SINGH, *Great Ideas and Theories of Modern Cosmology* (1961); THURE VON UEXKUELL and ERNESTO GRASSI, *Wirklichkeit als Geheimnis und Auftrag* (1945); P.A. SCHILPP (ed.), *Albert Einstein: Philosopher–Scientist*, 2 vol. (1959); PHILIPP FRANK, *Philosophy of Science* (1957); ALFRED NORTH WHITEHEAD, *Process and Reality* (1929) and *The Concept of Nature* (1920, reprinted 1964); MARY B. HESSE, *Forces and Fields* (1961); CARL FRIEDRICH VON WEIZSAECKER, *Zum Weltbild der Physik*, 7th ed. (1958); ADOLPH GRUENBAUM, *Philosophical Problems of Space and Time* (1963); WERNER HEISENBERG, *Physics and Philosophy* (1958) and *Das Naturbild der heutigen Physik* (1955); A.S. EDDINGTON, *Fundamental Theory* (1946) and *The Nature of the Physical World* (1928, reprinted 1955); HENRI POINCARE, *La Science et l'hypothèse* (1903; Eng. trans., *Science and Hypothesis*, 1905); HENRY MARGENAU, *The Nature of Physical Reality* (1950) and *Open Vistas* (1961); JAMES CLERK MAXWELL, *A Treatise*

on Electricity and Magnetism, 3rd ed., 2 vol. (1892, reprinted 1953); ERWIN SCHROEDINGER, *Expanding Universes* (1956) and *Mind and Matter* (1958); M.K. MUNITZ, *Space, Time and Creation* (1957); ALOYS WENZL, *Die philosophischen Grenzfragen der modernen Naturwissenschaft*, 3rd ed. (1960); NICOLAI HARTMANN, *Philosophie der Natur* (1950); C.G. JUNG and W. PAULI, *Naturerklärung und Psyche* (1952); S.E. TOULMIN and JUNE GOODFIELD, *The Fabric of the Heavens* (1961), *The Architecture of Matter* (1962), and *Discovery of Time* (1965); HERMAN WEYL, *Philosophie der Mathematik und Naturwissenschaft* (1934, 3rd ed., 1966; Eng. trans., *Philosophy of Mathematics and Natural Science*, 1949, rev. ed., 1963) and *Raum, Zeit, Materie* (1919, 5th ed., 1923; Eng. trans., *Space–Time–Matter*, 1922); G.J. WHITROW, *The Structure and Evolution of the Universe*, rev. ed. (1959); ERNST MACH, *Die Mechanik in ihrer Entwickelung* (1883, 9th ed., 1933; Eng. trans., *The Science of Mechanics*, 1893, 6th ed., 1960). See also *The Monist*, vol. 47, no. 1 (1962), an entire issue on cosmology; and *Archiv für Naturphilosophie und die philosophischen Grenzgebiete der exakten Wissenschaften und Wissenschaftsgeschichte*, a journal with special interest in cosmology.

Philosophy of biology: M.O. BECKNER, "Biology," in *The Encyclopedia of Philosophy*, vol. 1, pp. 310–318 (1967), an excellent review of some of the basic philosophical problems in biology; LUDWIG VON BERTALANFFY, *General System Theory* (1968), an attempt to unify the disparate aspects of biology and formulate general principles; BARRY COMMONER, "In Defense of Biology," *Science*, 133:1745–1748 (1961), an argument for the autonomy of biology decrying the emphasis on molecular biology to the neglect of general biology; FRANCIS CRICK, *Of Molecules and Men* (1966), expresses the attitudes and positions taken by many molecular biologists in defense of reductionism; A.G.N. FLEW, *Evolutionary Ethics* (1967), a significant attempt to relate evolutionary theory to ethics; T.A. GOUDGE, *The Ascent of Life* (1961), a work devoted solely to philosophical problems of evolutionary theory; MARJORIE GRENE, *Approaches to a Philosophical Biology* (1968), a review of the European viewpoints on the philosophy of nature. (*Naturphilosophie*): CLIFFORD GROBSTEIN, "Organizational Levels and Explanation," *J. Hist. Biol.*, 2:199–206 (1969), a technical article that treats the doctrine of emergence in connection with recent biological knowledge; THOMAS H. and JULIAN HUXLEY, *Touchstone for Ethics, 1893–1943* (British title, *Evolution and Ethics, 1893–1943*; 1947, reprinted 1969), opposing views on the same theme by two famous biologists; HANS JONAS, *The Phenomenon of Life: Toward a Philosophical Biology* (1966), a collection of essays presenting an existential interpretation of biology in the tradition of *Naturphilosophie;* HUGH LEHMAN, "Are Biological Species Real?" *Philosophy Sci.*, 34:157–167 (1967), an article that reviews some of the major philosophical problems of the species question; KONRAD LORENZ, *Das sogenannte Böse* (1963; Eng. trans., *On Aggression*, 1966), a controversial book that deals with the function of aggression in the animal kingdom, with extrapolations to man; JACQUES MONOD, *Le Hasard et la nécessité: essai sur la philosophie naturelle de la biologie moderne* (1970; Eng. trans., *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*, 1971), a significant addition to the biological literature by a Nobel Prize-winning biologist; ASHLEY MONTAGU (ed.), *Man and Aggression* (1968), a collection of essays and reviews critical of the notion of innate aggressiveness as expounded by Konrad Lorenz and Robert Ardrey; RONALD MUNSON (ed.), *Man and Nature: Philosophical Issues in Biology* (1971), a collection of writings that provides an excellent introduction to current thinking in biophilosophy, with incisive commentary by the editor; ERNEST NAGEL, *The Structure of Science* (1961), portions of which are devoted to a review of various viewpoints on the philosophy of biology; JEAN PIAGET, *Biologie et connaissance* (1967; Eng. trans., *Biology and Knowledge*, 1971), concerned primarily with the evolution of mental faculties; V.R. POTTER, *Bioethics* (1971), a collection of articles dealing with the influence of biology on practical moral problems; BERNHARD RENSCH, *Biophilosophie auf erkenntnistheoretischer Grundlage* (1968; Eng. trans., *Biophilosophy*, 1971), emphasizes the rise and nature of consciousness within the framework of the evolutionary process; M.A. SIMON, *The Matter of Life* (1971), a good review of the current approaches to the philosophy of biology; GEORGE GAYLORD SIMPSON, *This View of Life: The World of an Evolutionist* (1964), a collection of essays by a renowned scholar, some of which have shaped the direction of philosophical biology today, and the especially pertinent chapter, "Man's Place in Nature," in his *Meaning of Evolution*, pp. 281–294, rev. ed. (1967); PIERRE TEILHARD DE CHARDIN, *Le Phénomène humain* (1956; Eng. trans., *The Phenomenon of Man*, 1959), a controversial attempt to meld science and religion by presenting a mystical viewpoint of cosmic evolution; C.H. WADDINGTON, *The Ethical Animal* (1960), a defense of the claim that biological facts can provide grounds for moral principles.

(P.W.J./R.C.Y.)

Nature Worship

In the history of religions and cultures, nature worship as a definite and complex system of belief or as a predominant form of religion has not been well documented. Among nonliterate or primitive peoples the concept of nature as a totality is unknown; only individual natural phenomena—*e.g.*, stars, rain, animals, plants, water, and fire—are comprehended as natural objects or forces that influence them and are thus in some way worthy of being venerated or placated. The concept of nature as an entity in itself, in contrast with man, human society and culture, or even God, is a philosophical or poetic conception that has been developed among advanced civilizations. Nature worship, therefore, is a working concept limited primarily to scholars involved in or influenced by the modern (especially Western) study of religion.

NATURE AS A SACRED TOTALITY

To students of religion, the closest example of what may be termed the worship of nature as a sacred totality, or an entity in itself, is perhaps most apparent in cultures—*e.g.*, ancient or primitive—in which there is a high god as the lord in heaven who has withdrawn from the immediate details of the governing of the world. This kind of high god—the *Deus otiosus*, or hidden god—is one who has delegated all work on earth to what are called "nature spirits" (*i.e.*, the forces or personifications of the forces of nature). High gods exist, for example, in such indigenous religions on Africa's west coast as that of the Diola of Guinea. In such religions the spiritual environment of man is functionally structured by means of personified natural powers; *i.e.*, the nature spirits.

Pantheism (a belief system in which God is equated with the forces of the universe) or deism (a belief system based on a non-intervening creator of the universe), as was advocated in the rationalistic philosophy of religion of western Europe of the 16th to 18th centuries, is not appropriate in studies of concepts of nature in preliterate cultures. Worship of nature as an omnipotent entity, in the pantheistic sense, has not as yet been documented anywhere.

The concepts of mana, orenda, wakan, and manitou

The power or force within nature that has most often been venerated, worshipped, or held in holy awe is mana. Mana, often designated as "impersonal power" or "supernatural power," is a term used by Polynesians and Melanesians that 19th-century Western anthropologists appropriated to apply to that which affected the common processes of nature. Mana was conceptually linked to North American Indian terms that conveyed the same or similar notions—*e.g.*, orenda of the Iroquois, wakan of the Dakotas, and manitou of the Algonkin. Neither the designation "impersonal power" nor "supernatural power" implies what mana really means, however, because mana usually issues from persons or is used by them, and the concept of a supernatural sphere as distinct or separate from a natural sphere is seldom recognized by preliterate peoples.

Thus, a better designation for mana is "super force" or "extraordinary efficiency." A person has mana when he is successful, fortunate, and demonstrates extraordinary skill; *e.g.*, as an artisan, warrior, or chief. Mana can also be obtained from the *atuas* (gods), providing that they themselves possess it. Derived from a root term that has aristocratic connotations, mana corresponds to Polynesian social classifications. The ariki, the nobility of Polynesia, have more power (mana), and the area that belongs to them and even the insignia associated with them have mana. Besides areas and symbolic elements associated with the ariki, many objects and animals having special relationships with chiefs, warriors, or priests have mana.

The concept of *hasina* of the Indonesian Hova (or Meri-

na) on Madagascar is very similar to mana; it demonstrates the same aristocratic root character as mana in that it indicates that the word mana is derived from the Indonesian *manang* ("to be influential, superior").

The Iroquoian term orenda, similar to mana, designates a power that is inherent in numerous objects of nature but that does not have essential personification or animistic (soul) elements. Orenda, however, is not a collective omnipotence. Powerful hunters, priests, and shamans have orenda to some degree. The *wakanda*, or wakan, of the Sioux Indians is described similarly, but as Wakan-Tanka it may refer to a collective unity of gods with great power (wakan). The manitou of the Algonkin is not merely an impersonal power, comparable to the wakan, that is inherent in all things of nature but is also the personification of numerous manitous (powers), with a Great Manitou (Kitchi-Manitou) at the head. These manitous may even be designated as protective spirits that are akin to those of other North American Indians, such as the *digi* of the Apaches, *boha* of the Shoshones, *tamanoas* of the Chinook, and *maxpe* of the Crow, as well as the *sila* of the Eskimos.

The super forces (such as Mulungu, Imana, Jok, and others in Africa) that Western scholars have noted outside of the Austronesian–American circles of peoples are often wrongly interpreted as concepts of God. Only the *barakah* (derived from the pre-Islāmic thought world of the Berbers and Arabs), the contagious superpower (or holiness) of the saints, and the power Nyama in western Sudan that works as a force within large wild animals, certain bush spirits, and physically handicapped people —appearing especially as a contagious power of revenge —may be added with a certain justification to that force of nature that is designated by mana. A striking similarity with mana may also be noted in the concepts of *heil* (good omen), *saell* (fortunate), and *hamingja* (luck) of the Germanic and Scandinavian peoples.

HEAVEN AND EARTH AS SACRED SPACES, FORCES, OR PROCESSES

Heaven and earth as personified powers of nature

Heaven and earth, as personified powers of nature and thus worthy of worship, are evidently not of equal age. Though from earliest times heaven was believed to be the residence of a high being or a prominent god, the earth as a personified entity is much rarer; it probably first occurred among archaic agrarian civilizations, as it still does in some primitive societies in which agriculture is practiced. Gods of heaven, however, are characteristic spiritual beings of early and contemporary hunter and collector cultures and are found in almost all cultures.

Primitive world views generally assume the earth to be simply given (*i.e.*, as continuously existing). Sometimes it is believed to have emerged out of chaos or a primal sea or to have come into existence by the act of a heavenly god, transformer, or demiurge (creator). Even in such world views, however, the earth usually remains without a divine owner, unless through agriculture and the cult of the dead the earth is conceived as the underworld or as the source of the renewing powers of nature. The fact that heaven is animated by rain-giving clouds (with lightning and thunder) and by a regular chorus of warming and illuminating celestial bodies (sun, moon, and stars) led to concepts of the personification of heaven from earliest times.

Heaven. Heavenly deities, as the personification of the physical aspects of the sky, appear in variations that are adapted to the types of cultures concerned. The listing offered below does not represent a unilinear development that is applicable everywhere.

The father of the family. The god of heaven is often viewed as an ever-active father of the family, often called upon but rarely the recipient of sacrifices. He is able to intervene in human and natural affairs without the aid of an intermediary (*e.g.*, priest, medicine man, or ancestors). As a numinous (spiritual) being, he is closer to man than other spiritual powers. He sends lightning and rain and rules the stars that are at most essential aspects of himself or are members of his family subject to him. He is the creator and the receiver of the dead. Modern scholars have designated such a being as the "high god," "supreme god," the "highest being" of the "original monotheism" (according to the theories of a German scholar, P.W. Schmidt), the idealized god of heaven (according to the views of an Italian historian of religion, Raffaele Pettazzoni), or the familiar father deity (according to the views of a British anthropologist, Andrew Lang). Very human, often comical, or even unethical and repulsive traits of such deities are often represented in myths that also sometimes include legends of animal or human ancestors.

Relationship of the "high god" to men

This type of deity is generally found in its most developed form among the old hunter and collector peoples of the temperate and arid areas (*e.g.*, forest Indians of North America, Indians of California and of Tierra del Fuego in South America, Australian Aborigines, and African Bushmen) and of the tropic primeval forests, where he is usually conceived as a storm and thunder being (*e.g.*, Tore of the Ituri Pygmies or Karei of the Semang of the Andaman Islands). He is also worshipped among the pastoral peoples as the "blue" or "white" sky of the wide pastures in the steppes of northeastern Africa (*e.g.*, Waka of the Galla) and of Central and North Asia (*e.g.*, Torem, Num, and Tengri of the Ugrians, Samoyeds, and Mongols).

Among such peoples, heaven is often merged with an old hunting deity, the Lord of the Animals, or it allows the latter to exist as a hypostasis by his side.

The withdrawn god. The god of heaven also may be a *Deus otiosus*, who has, after completing the creation, withdrawn into heaven. He has abandoned the government of the world to the ancestors of men or to nature spirits that are dependent on him and act as mediators between him and men. This type of the god of heaven, who is able to intervene directly only when there are widespread existential necessities or needs (*e.g.*, drought, pestilence, or war), can be found primarily where worship of the dead or worship of individual local "earth spirits"—not yet integrated into an all-inclusive earth deity—obscures everything else. This type of god occurs especially in areas of so-called primitive agriculture (*e.g.*, large parts of Africa, Melanesia, and South America).

Relationships of the "high god" to other deities

The first among equals. The god of heaven also may be the head of a pantheon of gods, the first among equals, or the absolute ruler in a hierarchy of gods. This occurs in polytheism (belief in many gods) in its purest form. The deities associated with him are often related to him by family ties (genealogies of gods). Occasionally, the heavenly phenomena are distributed among members of the clan of gods, the god of heaven himself thus becoming rather vague. The divine pair heaven–earth represents only one among many possible combinations—*e.g.*, Dyaus-pitṛi (= heaven, male) and Pṛthivī (= earth, female) in Vedic India or, with an unusual distribution of the sexes, Nut (= heaven, woman) and Geb (= earth, man) in ancient Egypt.

Occasionally, generations of gods succeed each other (*e.g.*, Greece, western Asia). In such instances, the more universal god of heaven is often replaced by the younger god of thunderstorms (*e.g.*, Zeus of the Greeks, Teshub of the Hittites, or Hadad of the Semites) or is even relegated to the background by a goddess, such as Inanna-Ishtar (the love or fertility goddess) in Babylonia or Amaterasu, the sun goddess of Japan.

In ancient China, Heaven (T'ien, or Shang Ti, the highest lord) ruled over the many more popular gods and was even closely related to the representatives of the Imperial household. Deification of the celestial emperor is a cultic practice that extends from Korea to Annam (part of Vietnam). The roots of the worship of heaven in Asia are probably the beliefs of central and northern Asiatic nomadic peoples in a solitary god of heaven. Gods of heaven, above or behind a pantheon (grouping of gods), probably originated in areas in which a theocratic stratified bureaucracy existed or in which sacral kingdoms exist or have existed—*e.g.*, in The Sudan or northeastern Africa (Akan-Baule, Dahomey, Yoruba-Benin, Jukun, Buganda, and neighbouring states), western Indo-

nesia, Polynesia and Micronesia, and in the advanced civilizations of pre-Columbian Meso-America and South America (see also SACRED KINGSHIP; POLYTHEISM).

Heaven and earth deities as partners. The god of heaven in many areas is a partner of an earth deity. In such cases, other numina (spirits) are missing or are subject to one of the two as spirits of nature or ancestors. Myths depicting the heaven–earth partnership usually describe the foundations or origins of the partnership in terms of a separation of a primeval chaos into heaven and earth or in terms of a later separation of heaven and earth that originally lay close together, and they describe the impregnation of the earth by the seed of the god (*e.g.*, *Hieros gamos*, or the sacred marriage).

This type of god of heaven (plus goddess of earth) may be found in areas of Africa that have been influenced by advanced civilizations (especially The Sudan and northeastern Africa), in eastern Indonesia, and in some areas of America under influence of advanced civilizations.

The god of heaven viewed dualistically. In several religions the god of heaven has an antagonistic evil adversary who delights in destroying completely or partially the good creative deeds of the god of heaven. This helps to explain the insecurity of existence and concepts of ethical dualism. In most such cases, the contrasts experienced in the relationship between heaven and earth deities have been re-evaluated along ethical lines by exalting the heavenly elements at the expense of the earthly ones (especially in Jewish, Christian, and Islāmic sects in Europe, west central and northern Asia, and certain areas in the northern half of Africa). The figure of an antagonistic trickster or demiurge with a somewhat ethical component may be the result of diffusion and is rather rare in primitive cultures—*e.g.*, African Bushmen, Australian Aborigines, and North American Indians (see also DUALISM, RELIGIOUS).

Exaltation of heavenly elements over earthly elements

The god of heaven viewed monotheistically. The god of heaven, viewed in his ethical aspect, is always an active, single god—*e.g.*, Christian, Jewish, and Islāmic monotheism (see also MONOTHEISM).

Earth. Although in almost all polytheistic religions earth is usually represented as a goddess and associated with the god of heaven as her spouse, only rarely is there an elaborate or intensive cult of earth worship. There are in many religions mother goddesses who have elaborate cults and who have assumed the function of fertility for land and man, but they hardly have a chthonic (earth) basis. Some mother goddesses, such as Inanna-Ishtar, instead have a heavenly, astral origin. There are, however, subordinate figures of various pantheons, such as Nerthus (Jord) in Germanic religion or Demeter and Persephone (earth mother and corn girl) in Greek religion, who have played greater roles than the world mother (Gaea). Among Indo-Europeans, western Asiatics (despite their various fertility deities), Chinese, and Japanese, the gods of heaven, sun, and thunderstorms have held a paramount interest.

"Mother earth"

When the common people have displayed intensive attention to "mother earth" (such as the practice of laying down newborn babies on the earth and many other rites), this partially reflects older cults that have remained relatively free from warrior and nation-building peoples with their emphasis on war (as in western Sudan, pre-Aryan India, and the Indian agrarian area of northern Mexico). The Andean earth-mother figure, Pacha-Mama, worshipped by the Peruvians, stands in sharp contrast to the sun religion of the Inca (the conquering lord of the Andes region). Earth deities are most actively venerated in areas in which people are closely bound to ancestors and to the cultivation of grain.

Mountains. Especially prominent mountains are favourite places for cults of high places, particularly when they are isolated as island mountains, mountains with snowcaps, or uninhabited high mountain ranges. The psychological roots of the cults of high places lie in the belief that mountains are close to the sky (as heavenly ladders), that clouds surrounding the top of mountains are givers of rain, and that mountains with volcanoes form approaches to the fiery insides of the earth.

Mountains, therefore, serve as the abodes of the gods, as the centres of the dead who live underground, as burial places for rainmakers (medicine men), and as places of oracles for soothsayers. In cosmogenic (origin of the world) myths, mountains are the first land to emerge from the primeval water. They frequently become the cosmic mountain (*i.e.*, the world conceived as a mountain) that is symbolically represented by a small hill on which a king stands at the inauguration. Pilgrimages to mountain altars or shrines are favourite practices of cults of high places.

The larger mountain ranges and canyons between volcanic mountains—especially in Eurasia from the Pyrenees to the Alps, the Carpathian Mountains, the Caucasus Mountains, the Himalayas, the mountainous areas of northern China and Japan, and the mountainous areas of North and South America (the Rocky Mountains, the Andes)—are most often centres of cults of high places. Elevations of the East African Rift Valley (Kenya, Tanzania, Uganda), volcanic islands of the Pacific Ocean (*e.g.*, Hawaii), and the mountains of the Indian Deccan have also served as centres of the cult of high places.

The high places as the abodes of the gods

In the early civilizations, the cult of high places was closely combined with that of the earth (*e.g.*, Olympus of the Greeks, the mountains of Enlil or of the Mountain Mother Cybele in western Asia, and the Meru mountain of India) were believed to bring heaven and earth into a close relationship and were often viewed as the middle pillar of the world pillars upholding the sky. Bush and wild spirits (Lord of the Animals) of the cultures of the hunters and collectors were often believed to reside in inaccessible mountainous areas (*e.g.*, Caucasus).

In addition to other mountain deities of a more recent date (*e.g.*, God of the Twelve Mountains and the One-Legged Mountain God), the mountain deity Yama-no-kami has been demonstrated to be a deity of the hunt (*i.e.*, god of the forest, Lord of the Animals) of ancient Japan. Through the worship of farmers, Yama-no-kami assumed the elements of a goddess of vegetation and agriculture. The mountain goddesses (earth mothers) of non-Aryan India still incorporate numerous features of hunt deities, and, because of indigenous influences, the Vedic (early Aryan scriptural) gods and their wives (*e.g.*, Pārvatī, Umā, and Durgā) have their abodes on mountains. The isolated mountains of East Africa, surrounded by clouds, are believed to be the dwelling places of the heaven and rain gods, and in the area of Zimbabwe (Rhodesia) there are pilgrimages to mountain sanctuaries that are viewed as the seats of the gods.

Pre-Islāmic peoples of North Africa and the extinct inhabitants of the Canary Islands (Guanches) associated mountain worship with a cult of goats (and sheep), which, when practiced in rituals, was believed to secure rain and thunderstorms in the often arid landscape. Similar cults are also found on the Balkans and in the valleys of the southern Alps.

Beliefs about the causes of earthquakes

Earthquakes. According to the beliefs of many peoples, earthquakes originate in mountains. In areas of Africa, where the concept of mana is particularly strong, many believe that the dead in the underworld are the causes of earthquakes, though in the Upper Nile, The Sudan, and East Africa, an earth deity is sometimes blamed. In some areas, a bearer who holds the world up—a concept that probably came from Arabia, Persia, India—is believed to cause an earthquake when he changes his position or when he moves his burden from one shoulder to the other. In the Arab world, on the east coast of Africa and in North Africa, an ox generally is viewed as the bearer, sometimes standing on a fish in the water. World bearers often are giants or heroes, such as Atlas, but they also may be animals: an elephant (India), a boar (Indonesia), a buffalo (Indonesia), a fish (Arabia, Georgian S.S.R., and Japan), a turtle (America), or the serpent god Ndengei (Fiji). Generators of earthquakes also may be the gods of the underworld, such as Tuil, the earthquake god of the inhabitants of Kamchatka (a peninsula in eastern Siberia), who rides on a sleigh under the earth. The earthquake is driven away by noise, loud shouting, or

poking with the pestle of a mortar. Among peoples with eschatological (last times) views, earthquakes announce the end of the world (Europe, western Asia).

Tides. The view that the tides are caused by the moon can be found over almost all the earth. This regular natural phenomenon seldom gives rise to cults, but the ebb and flow of the coastal waters has stimulated mythological concepts. Not infrequently the moon acquires the status of a water deity because of this phenomenon. The Tlingit (of northwestern America) view the moon as an old woman, the mistress of the tides. The animal hero and trickster Yetl, the raven, is successful in conquering (with the aid of the mink) the seashore from the moon at low tide, and thus an extended area is gained for nourishment with small sea animals.

CELESTIAL PHENOMENA AS OBJECTS OF WORSHIP OR VENERATION

The sun. Generally, the sun is worshipped more in colder regions and the moon in warm regions. Also, the sun is usually considered as male and the moon as female. Exceptions to these generalizations, however, are notable: the prevalent worship of the sun in hot, arid ancient Egypt and in parts of western Asia; the conception of the moon as a man (who frequently is believed to be the cause of menstruation) among primitive hunter peoples (African Bushmen, Australian Aborigines, and hunters of South America) and among certain pastoral and royal cultures of Africa (*e.g.*, the Masai and the Hottentots); and the conception of the female sun ruling northern Eurasia from the Northern Sea to Japan and parts of North America.

In many state cults of ancient civilizations, the sun plays a special role, particularly where it has replaced an old god of heaven (*e.g.*, Egypt, Ethiopia, South India, and the Andes) and especially where it is viewed as a marker of time.

Solar religions in Africa and the Orient

The sun as the centre of a state religion. In Africa, ancient Egypt was the main centre from which solar deity concepts emanated. The solar religion, promoted by the state, was concerned with the sun god Re (Atum-Re, Amon-Re, Chnum-Re), the sun falcon Horus, the scarab (Chepre), and a divine kingdom that was determined by the sun (*e.g.*, pharaoh Akhenaton's solar monotheism *c.* 1350 BC). The sun religion reached, by way of Meroe (a sun sanctuary until the 6th century AD) and the Upper Nile, as far as western Ethiopia (*e.g.*, the Hego cult in Kefa and the sun kings in Limmu) and Nigeria (*e.g.*, Jukun). In the Orient, the sun cult culminated in the religion of Mithra of Persia. Mithra was transported by Roman legionaires to western Europe and became the Unconquerable Sun of the Roman military emperors. In Japan, the Imperial deity in state Shintō is Amaterasu, the sun goddess from whom Jimmu Tennō, the first human emperor, descended. In Indonesia, where the descent of the princes from the sun also is a feature, the sun often replaces the deity of heaven as a partner of the earth. In Peru, the ruling Inca was believed to be the sun incarnate (Inti) and his wife the moon. A sun temple in Cuzco contains a representation of Inti as the oldest son of the creator god. The Natchez Indians of southeastern United States, who are culturally connected with Central America, called their king "Great Sun," and the noblemen were called "the Suns."

The sun as a subordinate deity. The sun, within a polytheistic pantheon, often is revered as a special deity who is subordinate to the highest deity, usually the god of heaven. This may be observed in the great civilizations of ancient Europe and Asia: Helios (Greece); Sol (Rome); Mithra (Persia); Sūrya, Savitṛ, and Mitra (India); Utu (Sumer); and Shamash (Babylonian and other Semitic areas).

The sun not infrequently is considered female (Shams of some Arabs, Shaph of ancient Ugarit in Palestine, Sun of Arinna of the Hittites, as well as the female Sun of the Germans). Siberian people such as the Taimyr Samoyed (whose women pray in spring to the sun goddess in order to receive fertility or a rich calving of the reindeer) or the Tungus worship sun goddesses. They sacrifice to the sun goddess, and her symbols are embroidered on women's clothes.

The sun and moon as a divine pair. A sun god is often related to a moon goddess as one member of a divine pair (in the place of heaven and earth as "world parents"). A sun-moon god exists among the Munda-Dravida in India (Singbonga); a sun-moon (earth) pair, partially seen as bisexual, exists in eastern Indonesia; and Nyambe (the sun) among the Rotse in Zambia is represented as united with the moon goddess as the ruling pair.

The sun as an attribute of the highest being. The sun sometimes is viewed as a coordinated or subordinate attribute, or hypostasis, of the highest being. This may possibly occur because of a partially weakened influence of a stronger solarism in areas of older primitive peoples, such as in The Sudan, Upper Volta, Nigeria, northern East Africa, and Australia.

The sun as a mythical being. The sun, in some religions, is conceived as a purely mythical being, who is cultically recognized in sun dances (*e.g.*, prairie Indians) and in celebrations of the solstice, with jumps over fires, sports festivals, and other events. These rites may be either survivals of an earlier local cult of a sun deity or influences of such a cult.

The moon. The moon is often personified in different ways and worshipped with ritual customs; nevertheless, in contrast to the sun, the moon is less frequently viewed as a powerful deity. It appears to be of great importance as the basis of a lunar calendar but not in the higher agrarian civilizations. The moon, infrequently associated with the highest god, is usually placed below heaven and the sun. When the moon with the sun together (instead of "heaven and earth") constitute an important pair of gods (world parents), it frequently assumes the features of an earth deity. In tropical South America, the sun and moon are usually purely mythical figures.

The moon as a female or a male deity

Between the Tropic of Capricorn and the Tropic of Cancer, the moon is predominantly female. Only some remainders of ancient peoples (hunter peoples) view the moon as a male being (*e.g.*, African Bushmen, Australian Aborigines, Congo Pygmies, Semang, Andamans, Chaco Indians, Ona, and some Brazilian tribes). In the few significant moon gods of the Oriental civilizations (Chonsu in Egypt, Sin-Nanna in Babylonia, Candra in India)—in contrast with the female Selene and Luna in the Greek–Roman culture, a more ancient substratum may possibly be present. Where the moon is considered as male, he often determines the sexual life of the woman, especially among the Aborigines of Australia.

The phenomenon of the moon that attracts all people is the sequence of its phases. The waxing and waning of the moon crescent is often interpreted as gaining or losing weight (eating, dieting). Thus, the Taulipang in Guayaná believe that the moon is first nourished well and then inadequately by his two wives, Venus and Jupiter. Where the moon is viewed as female, the phases represent pregnancy and delivery. Elsewhere, people see childhood, maturity, and dying as the phases of the moon: the first crescent is thus the rebirth or the replacement of the old by a new moon.

The appearance of the crescent or the full moon is sometimes celebrated by a rest from work, and some attempt to participate in the waxing and waning of the moon by analogous magical rites. Girls with small breasts stand in the full moonlight (in the Salzburg, Austria, area); persons who desire a tumor to decrease point to the waning moon; and newborn children often are exposed to the waning moonlight, or they and everything else that is desired to be healthy and permanent are dyed white; *i.e.*, they are made "moonlike." Nearly everywhere, connections between the moon phases and the rhythm of nature (the tides) and humans (menstruation) are recognized.

Significance of the phases and signs of the moon

The three dark days of the "death" of the moon are believed by many to be dangerous. During this period the moon is believed to be defeated in a battle with monsters who eat and later regurgitate the moon; or the moon is viewed as having been killed by other heavenly beings

and later revived. The period is a time in which people, if possible, do not engage in a new enterprise.

The halo of the moon is also viewed as a bad omen among many peoples. Moon spots are regarded as testimonies of a battle with heavenly opponents. In addition to the popular Man in the Moon, there are also other figures represented: "the woman with the basket on her back," "the spinning woman," or "the weaving woman" (in Polynesia the woman who pounds tapa). The most popular animal figure recognized in the features of the moon, the rabbit (from Europe to America), presumably earned this role because of its fertility.

Beliefs and practices centring on eclipses

Eclipses of the sun and moon. Eclipses of the sun or the moon—usually interpreted as a battle, as the dying, or the devouring of one of the two heavenly bodies—in many religions are met with anxiety, shouting, drum beating, shooting, and other noises. Many North American Indian tribes, Hottentots in Africa, Ainu in Japan, and the Minangkabau in Sumatra interpret the eclipses as fainting, sickness, or the death of the darkened heavenly body. In Arctic America, Eskimos, Aleuts, and Tlingits believe that the sun and moon have moved from their places in order to see that things are going right on earth. The explanation that heavenly monsters and beasts pursue the stars and attempt to injure and to kill them, however, is a view found over a larger area. Noise and shooting are believed to deter the monsters from their pursuit or to force them to return the celestial bodies if they have already been captured. In China and Thailand the monster is the heavenly dragon; in China, among the Germanic tribes, and among northern American Indians, dogs and wolves (coyotes) are the culprits; in Africa and Indonesia, they are snakes; in South America, the beast is the jaguar; and in India they are the star monsters Rāhu and Ketu. The belief in the darkening of one star by the other in a battle—*e.g.*, between the sun god Lisa and the moon goddess Gleti in Dahomey—is about as widespread. An eclipse may also be interpreted (as in Tahiti) as the lovemaking of sun and moon, who thus beget the stars and obscure each other in the process.

Stars and constellations. Worship of the stars and constellations in the modern world survives only in a very corrupt or hidden manner. True star worship existed only among some ancient civilizations associated with Mesopotamia, where star worship was practiced. Mesopotamia, where both astronomy and astrology reached a high degree of refinement—especially after a Hellenizing renaissance of astronomy—was the origin of astral religions and myths that affected religions all over the world. Though the view is controversial, Mesopotamian astral worship and influence may have reached as far as Central and Andean America (by way of China or Polynesia). Sumerian, Elamite, and Hurrian contemplation of the stars influenced not only Mesopotamia, Asia Minor, Egypt, Iran, and India but also other areas. Knowledge of the zodiac, the planets, and observation of precession extended from the West to south Asia—*e.g.*, the Pythagoreans and Orphics (mystical philosophers) in the Mediterranean area and astrological mystical thinkers in India, Indonesia, China, and Polynesia. West Sudan, for example, was deeply influenced by the spirit of ancient Mediterranean and Oriental knowledge of the stars.

The influence of Mesopotamian astral worship

Apart from areas in The Sudan, northeast Africa, and Rhodesia (Monomotapa), not much of Africa has had any considerable knowledge of the stars. That knowledge of the stars is relatively limited among forest peoples, unless old hunting cultures survived, is explained by an Ekoi tribesman in southeast Nigeria, as follows: "Ekoi people do not trouble themselves about the stars, because the trees always hide them." Hunting pygmies likewise have never achieved any significant knowledge of the stars, which the Bushmen on the steppe have.

Knowledge of the stars in the areas of the true primitive peoples rarely leads to a worship of the stars. True star gods are rare, for example, in large parts of Africa. In Polynesia, where significant knowledge of the stars by the seafaring people and fishermen was learned in regular schools of astronomy, there seldom occurred what can be called true religious worship of the stars. Knowledge of the stars, however, is still relatively significant among the hunting peoples in the Southern Hemisphere, especially among the African Bushmen and Australian Aborigines, who were formerly untouched by the high civilizations. Economic considerations connected with the rising and setting of the stars, however, surpasses their mythological significance by far. The stars are usually considered to be living beings, particularly animals that have been transferred to the sky. They evidently are taken seriously primarily because they indicate by their rising and setting the appearance of game to be hunted or fruits to be collected.

Widespread beliefs in the constellation Orion and the planet Venus

The widespread African interpretation of the constellation Orion as a hunter, as game, or as a dog (from East Africa to the lower Congo and in the area of the Niger) is most likely a vestige from an earlier hunting period that has survived in agricultural civilizations. In a different form, the constellation Orion is still known in Europe as a hunter, in north Asia as a hunter of reindeer and elk, and in North America as a hunter of bears. In South America—outside the Andean empires—a whole series of astral beliefs of the ancient hunting culture has been preserved: the concepts of stars and constellations as Lords of the Animals, as helpers of the hunter, or as animals themselves.

Venus, the best known planet, has probably experienced its most significant personification in the figure of the Mesopotamian goddess Inanna-Ishtar. She was viewed as a being, sometimes female and at other times bisexual. Through her identification with the Greek Aphrodite and the Roman Venus, Inanna-Ishtar, the queen of heaven, still survives in Roman Catholic iconography—*e.g.*, as the Virgin Mary standing on the moon. African cultures also have been significantly impressed by this planet, not only in the rare figure of a Zulu heavenly goddess who determines the agricultural work of the women but even more as the evening and morning star, who are the wives of the moon. In the royal culture of Monomotapa (Rhodesia) and its influences in Buganda and southern Congo, the king is related to the moon, and his wedding with the Venus women is a type of *hieros gamos* ("sacred wedding"). In large areas of Africa the concept of "Venus wives of the moon" is preserved, although the moon is usually considered as the wife (or sister) of the sun. This concept was most likely prevalent at a time when the moon-king ideology was widespread in the eastern half of Africa from the Nile to South Africa, perhaps indicating south Arabian influences.

The Pleiades, Polaris, and the Milky Way

The Pleiades, a group in the constellation of Taurus (six to seven adjacent stars), are viewed in many parts of the world as maidens who are pursued by men (*e.g.*, hunters). The Pleiades are also interpreted as a mother hen with her chickens, especially in Eurasia, where the star Aldebaran, which is located close to the Pleiades, is often included as a part of the constellation. In Africa the Pleiades designate the beginning of the agricultural year. Therefore, in many Bantu languages the verb *kulima* ("to hoe") furnishes the basis for their designation *kilimia*, the Pleiades. In addition to East and South Africa there is still a smaller area in the western Sudan that retains this belief.

Polaris (the north star) enjoys a central significance among the Finno-Ugric and Turkish Tatars as "nail of the world" or "pillar of heaven." Among Altai Tatars Polaris is viewed as the negotiator of the god of heaven Ülgan; in Japan, Polaris is a god of heaven above the ninth layer of clouds.

The Milky Way, depending on a group's economy and life style, is often simply named after hunting or domestic animals: way of the tapir, the donkey, or the camel. It also is called the seam of the heavenly tent or a water stream. As the footsteps of God or the way of God, as the way of the dead, or as a deserted way of the gods, the Milky Way reveals older mythical conceptions, among which is that of the world (cosmic) tree.

Aurora borealis, the northern lights of the polar regions, is frequently interpreted by Arctic and sub-Arctic peoples (*e.g.*, Eskimo, Athapascan, Tlingit) as the reflection of the dance fire of the ghosts or of the peoples further in

the north, as the "cooking of meat," or the ball game of these peoples. Northern Germanic tribes saw in it the splendor of the shields of Valkyrie (warrior women).

FORCES OF NATURE

The natural forces of fire and water, which evidently exclude each other, are brought together in a unity of opposites in the worldviews of early archaic civilizations. Both forces are purifying as well as protective, and are viewed by many as being connected with the cosmic powers of the sun and moon. Where they are truly combined, often genetically, fire (as the sun) is usually male, and water (as the moon), female. Where the fire is included more into the chthonic (earthly) sphere it may also receive a feminine character (*e.g.*, fire in the earth, preserved in the womb); where rain is viewed as the semen of heaven, which is usually personified as male, it takes on a male character.

The "living" qualities of water

Water. Many of the qualities of water make it appear to be animated; on this basis it is psychologically understandable that water (*e.g.*, rain, sea, lakes, and rivers) might become a natural phenomenon worthy of worship. Water is always in motion, changes color in the light of the stars, reflects the world and man, "speaks" with murmuring and roaring, brings new life to dried out vegetation, refreshes men and animals, the tired and the ill, and heals. Because it dissolves dirt it is also most suitable for purifying the soul (*e.g.*, after the violation or the commission of a sin of any kind). Under certain circumstances, even pictures (icons) of gods have to be washed. Water also demonstrates destructive forces (seaquakes, floods, and storms). The most important mythical-religious facts symbolized by water are the following: the primal matter; the instrument of the purification and expiation; a vivifying force, a fructifying force; and a revealing and judging instrument.

Water as primal matter. The conception of a primal body of water from which everything is derived is especially prevalent among peoples living close to coasts or in river areas—*e.g.*, the Egyptian Nu (the primordial ocean) and the Mesopotamian Apsu (the primeval watery abyss) and Tiamat (the primeval chaos dragon). The earth may be fished out or emerges from the primeval water; heavenly beings (*e.g.*, Ataentsik, ancestress of the North American Iroquois) appear on the emerged earth; and birds lay an egg that is later divided into two halves (heaven and earth) on the chaotic sea. Thus, water is viewed as the foundation of all things. A survival of the original primeval sea, in such myths, is the water that flows around the earth's disk (*e.g.*, Oceanus).

Water as an instrument for purification and expiation. Water is viewed as an instrument for purification and expiation, especially in arid areas. Cultic acts, in such areas, generally take place only after lustrations, sprinkling, or immersion in water. The same view holds true for entry into new communities or into life (*e.g.*, baptism). Water lustration is especially necessary after touching the dead, and as a purificatory washing for priests and kings. Pictures of the gods also have had water poured over them.

Myths of a great flood (the Deluge) are widespread over Eurasia and America. This flood, which destroys with a few exceptions a disobedient original population, is an expiation by the water, after which a new type of world is created.

Water equated with life forces in the body

Water as a vivifying force. Water is viewed as vivifying, like the heavenly rainwater that moistens the earth. Water also is equated with the flowing life forces of the body (*e.g.*, blood, sweat, and semen). In order to replace the lost liquids, water was added to the mummified dead in Egypt. The African Ashanti designate their patrilinear groups as *ntoro*, which means water, river, and semen, and the Wogeo of Papua call their patrilinear clans *dan*; *i.e.*, both water and semen. The myth of the Kasuar ancestress of the But of Papua related how Kasuar's blood became sea (and salt).

Water as fructifying. Wherever early archaic culture spread the myth of the world parents heaven and earth, there also was a belief that heaven fructifies the earth with heaven's seed. The springs, pools, and rivers on the earth, therefore, may bring not only healing and expiation, but also fertility. The river Skamandros in ancient Greece evidently was so personified; according to Aeschines, a 4th century BC Greek orator, girls bathed in it before marrying and said: "Skamandros, accept my virginity." Magical rites in which water serves as a substitute for semen or the fertility of men are numerous.

At the corn festival Nsiä of the Bamessing (in Cameroon), which is celebrated in the dry season, the festival begins with the mourning of the dead vegetation. Reminiscent of the Egyptian Osiris and the Mesopotamian Tammuz festivals, the Nsiä festival emphasizes that the god who gave the nourishment has died and is being mourned like a chieftain. The chief, dying symbolically with the god, has to be strengthened with a miraculous "chieftain water," which has to be fetched by virgins of the chieftain's clan. For two weeks the chieftain drinks from the gourds of all the maidens after the women of the tribe have drunk from the holy water place.

Battles of gods and heroes with mythical beings, beasts, and monsters that hold back the fructifying water are widespread in mythology. The liberation of water during the mythical battle is equivalent to the end of the dry season or a drought, to the reviving of vegetation. In Indian mythology Indra slays Urtra; in Syrian and Palestinian mythology Baal battles with Leviathan; and in Huron (North American Indian) mythology Joskeha, the spring hero, kills the frog that attempted to restrain the water from flowing freely.

Water as a revealing or judging instrument. Water also serves as an instrument that reveals and judges. Reflections in the water led to a whole series of oracles originating from an alleged prophetic or divinatory power of water. A visionary look into the water surface was believed to reveal the future as well as past misdeeds. This ancient custom may have been preserved in the use of crystal balls by modern fortune tellers. The custom of water divination is found in ancient Europe, North Africa, the Near East (*e.g.*, Babylonian fortune telling by means of cups), eastern and northern Asia (where the use of metal mirrors by the shamans often replaces the water as a divining means), and in Southeast Asia and Polynesia. Where such means of divination were severely repressed, such as in sub-Saharan Africa, these methods of mirror and water gazing were changed into manipulated water ordeals.

Water is also used as a judging element: in ordeals believed to demonstrate the judgment of the gods, water ordeals (*e.g.*, immersion in water), as well as the more frequent fire ordeals, appear. Here also the purifying character of the water plays a role.

Fire. Worship of fire is widespread, especially in areas where the earthly fire is believed to be the image of the heavenly fire.

Fire as a living force or power

Because of various psychological reasons, fire is considered to be a personified animated or living power: it moves vehemently, devours, and becomes hungrier; it spreads fast into a giant blaze and is red like human blood and warm like the human body. It makes the plants that it has devoured suitable for fertilizing the earth; it shines brightly in the night and, by transference, may have "eternal life" or by constant rekindling can be made into a "perpetual fire." In cremation it separates the body from the soul; it drives away predatory animals and insects that cause pestilence.

Its chief functions are similar to those of its main adversary, water: to purify and to ward off evil, especially from home and hearth. Fire magically drives away rain, but with its smoke it also attracts rain clouds during a period of drought. Fire is believed to have both heavenly and earthly origins: it is brought by lightning, and it lives in the volcano of the underworld.

Stories are told of ancestors, heroes, or animals of primeval times who purloined the fire from the higher numina (spiritual powers). Bringers of civilization, similar to the Greek god Prometheus, fetch it—often together with fruits of the field, iron, or musical instruments—from heaven. Like Prometheus, Nommo, the primal being among the Dogon in Mali, brings fire and the first fruits

of the field down to the earth. Prometheus steals the fire from the blacksmith Hephaestus, but Nommo himself is the first blacksmith. In both areas this cultural achievement is celebrated with annual torchlight parades (in Greece, called Promethea festivals). Elsewhere, birds, or animals, such as the dog (especially in Africa), who is closely allied to the hearth fire of man, are the bringers of fire. Animals often fetch the fire from the Lord of the Animals in the bush.

Where geysers and volcanos indicate that the oldest fire is beneath the surface of the earth, fire is brought forth by animals and heroes. The Maori hero Maui seizes it from his ancestress Mahuike in the depth of the earth and puts it into a tree. Since that time it has been possible to get fire from the wood of the trees (*e.g.*, the fire borer). In areas practicing a definite ancestor worship, hunters obtained the fire from the subterranean world of the dead (as in East Africa). Previous to the Iron Age (15th century–2nd century BC), the generating of fire with the aid of fire borers, or fire saws, was viewed as a sexual act (male and female fire wood), especially in East and South Africa, India, Indonesia, and Mexico. In the creation myths of the Dayak of Borneo, fire is produced by rubbing a liana (male) on a tree (female) and is interpreted as coitus. The Tlingit (of northwest America) tell a story of the magical conception of a girl by the sawdust of the fire borer. The boring for the new state fire in the Loango empire (West Africa) coincides with the public coitus of a young couple.

The generating of fire as a sexual act

This conceptual framework seems to be a late consequence of earlier ideas of fire in the body of humans, especially of women, as a centre of sexual life. Such views are probably most pronounced among the Aborigines of Papua and Australia. The Marind in New Guinea, who, in their myth of the origin of fire view it as being derived from the sexual act, undertake the new boring of fire in connection with a cultic act in which the raping of a girl is the central rite. Elsewhere in New Guinea, there is a concept that fire lies in the genitals of women, especially of the first woman.

When iron-smelting techniques by means of fire became common among Neolithic (New Stone Age) peoples of similar mentality, as in Indonesia and Africa, the making of iron in shaft furnaces (considered as female) and bellows (male) has been interpreted as coitus with a subsequent birth (especially among the Bantu).

In archaic civilizations with sacral kings, the sacred perpetual fire (*i.e.*, the state fire) of the residences and temples of the royal ancestors was believed to have a phallic element. It was cared for by virgins, who were viewed as wives of the fire. Vestal virgins of this kind are documented in ancient Rome, Monomotapa (Africa), and pre-Columbian America. Among the Maya of Central America, an order of fire caretakers was founded by a deified "virgin of the fire." Extinguishing and rekindling of fire at the inauguration of a prince points to the idea of a spirit of the princes in the state fire and also to the cyclical renewal of the state in the purifying fire, which signifies the beginning of a new era.

Fire gods are found less frequently in primitive cultures than in archaic civilizations. That fire gods are not yet known in some ancient Oriental areas is probably because of the advanced development of a sun deity.

Iranian fire worship was derived from the cult of the god Ātar, but it was made a central act in Zoroastrianism (a religion founded by the 6th century BC Iranian prophet Zoroaster). Fire worship continues to be practiced among the Parsiis (modern Zoroastrians) of India: in temples the sacred fire is maintained by a priest using sandalwood, while his mouth is bound with a purifying shawl; fire in new temples is kindled from the fire of the old temples; household fires are not permitted to go out and are greeted in the morning by the members of the household and offered sandalwood; and Aryan cremation practices are renounced because the fire would be contaminated, and the dead are thus deposited in the "temples of silence."

Weather. The worship of atmospheric powers can only with difficulty be separated from the worship of heaven. In most cases the high god in heaven is also the god of thunderstorms and rain. Specific gods of wind and storm are found especially in countries with tornadoes and hurricanes (*e.g.*, the Maya deity Huracan). People (*e.g.*, the Tuareg and Arabs) in arid countries and deserts, dried out by the wind, speak of sand funnel spirits or of a desert god, such as the "boneless Kon" of the Peruvians.

The worship of atmospheric powers

From northern Europe to the tropical forests, thunderstorm deities rule heaven and earth. The most famous group of these numina (spiritual beings) are the Indo-European thunder gods (Thor-Donar of the Germans, Taranis of the Celts, Perkunis of the Slavs, Indra of the Indians, Zeus-Jupiter of the Greeks and Romans), who throw their thunderbolts or bundles of lightning. The Finnish god Ukko and the Basque god Orko probably stem from the same root; these gods still continue in the popular beliefs of Eastern Europe or Latin America today, such as St. Elijah or Santiago. These deities are related to the west Asiatic gods Teshub and Hadad (associated with the steer and with lightning) but also to the thunder god Shango of the Nigerian Yoruba, who is accompanied by a ram (as Thor uses a he-goat for pulling his wagon). Shango, as Yakuta, throws thunderbolts (*i.e.*, stone axes) to the earth, as does the Chac Mol of the Maya.

The goat, the ram, or horses appear as companions of weather gods or as animals that pull the thundering sky vehicle. In other cultures thunderbirds are the companions of the thunder gods or are the lightning itself. The lightning bird Zu, or Imdugud, occurs in ancient Mesopotamia, and the Garuḍa (with Wadjra) in Vedic India. Thunderbirds are represented (sometimes with arrows or spears in their bill or fangs) on archaeological artifacts of the Bronze Age in Dodona, Minussinsk in Siberia, Dong Son in Vietnam, and on pots in north Peru; they are described in myths of the Pueblo and prairie Indians and among East and South Africans.

Where prayers or sacrifices to gods and ancestors in the religious cult are not effective in producing rain, rain magic, which is practiced universally in similar rites, is often able to accomplish it. Trained magicians usually perform such rites, but ancestral priests or "persons holding power" also may do so. In rain magic, sprinkling, spitting, or immersion of people or things is often used to call down heavenly moisture. Smoke clouds to attract the rain accomplish the same purpose. There also must be suitable vestments (fresh greens, skins or pelts of water animals), body painting (representing clouds), or adornment with bird down. The color black in the clothing or on a killed or exposed animal is believed to be especially effective. Animals held responsible for holding the rain or water back (frogs, snakes, or mythological dragons) must be challenged. The sound of rain or thunder is produced with "bull-roarers," whistling, noise pots, rattles, and chains. If excessive rain is to be stopped, the injunction to perform or refrain from certain acts (*e.g.*, the prohibition of washing, boiling water, burning objects, making noise, and whistling) must be observed.

Rain magic

The rainbow often is considered a being, generally in the form of an animal, who swallows and holds back rain or water. The rainbow serpent (as a double bow also conceived as bisexual) is a figure that is found especially in the tropics of Africa, south Asia, north Australia ("ungud" snake), and Brazil. Elsewhere, the rainbow is viewed as a heavenly bridge that connects the worlds of gods and men: the Bifröst bridge in the Edda (a Norse saga); the bridge of the soul boats in Indonesia or of the creator god in Africa; and the path of the Greek goddess Iris. In Christian iconography, the rainbow is the throne of Christ; among Arabs and some Bantu of Central Africa it is the bow of god, and among the Nandi, Masai, and the Californian Yuki it is the robe of god.

WORSHIP OF ANIMALS, ANIMALISM, AND TOTEMISM

Animalism. Among the numerous animals that are prominent in religion and magic, the wild animals of the forests, the sea, and the air that are most important for the hunter are the most significant. Hunters and collec-

tors, rooted in the earliest cultures of man, believed that they not only had to kill animals—which were important for their economy as nourishment and raw materials—but also that they had to avoid their revenge. The feeling of a close connection between men and animals that has been lost to the highly civilized people (broadly speaking) led to an anthropomorphizing of animals. They were not only considered as living beings but were humanized to such an extent that the borderline between man and animal became virtually nonexistent and animals were held responsible for crises.

The religious magical attitude by which primitive peoples confront animals may be called animalism, regardless of whether the animal is thought to have life (animatism) or to have a soul (animism).

Totemism. The best known form of what may be called animal worship is totemism. An anthropologist, Sir James Frazer, defined totemism as "an intimate relation which is supposed to exist between a group of kindred people on one side and a species of natural or artificial objects on the other side, which objects are called the totems of the human group" (*Totemism and Exogamy*, 1910, IV, 3ff.). Frazer added that the mentioned "species" generally is more "natural" than "artificial." In this "intimate relation" the animal has the prerogative and in true totemism this relationship represents a more or less reverent attitude of man toward his animal partner. In totemism the relationship with the animal has not only merged into the realm of religion but has at the same time become a phenomenon of social life.

Man's relationship to animals

The Algonkin Indian word totem actually refers to one's own group, or family, thus indicating that totemism is a group phenomenon. It has also become customary among scholars, however, to include here "individual totemism" (that is, the personal permanent intimate relationship of an individual with an animal) and also "sex totemism," found mainly in Australia, where the women and men of a tribe separately are associated with an animal as totem. In Southeast Australia the totem is most commonly a species of small birds or male or female lizards.

True totemism, however, is always "group totemism," and that only when the totemic group is a so-called "unilinear association," such as a clan, which may be either matrilinear or patrilinear. Neither individual nor sex totemism belong to group totemism. In tribal totemism, one animal is significant for the whole tribe, thus indicating a kind of worship of a tribal animal. The Kwena tribe of the Tswana (South Africa) has the crocodile as such a tribal animal; they call it their "father." Sex and tribal totemism would better be associated with the general worship of animals; *i.e.*, the animal cults within the framework of "animalism." The same is true of the individual "totems," which are better designated as guardian animals.

The main features of group totemism are: unilinear groups of a tribe (clan, phratries) that are named after an animal (more rarely plant or sky phenomena) and avoidance of the totem, by not killing or eating it. The totem is inherited in the paternal or maternal lines. The group also believes, through its clan legend, in a descent from an ancestor who took up the association with this animal.

In "cult totemism," which shows a stronger religio-magical relationship to the animal but is rarer than the more profane form, the following features may be added: sacrifices and nourishment for the animal (a clan on Samoa gives to the eel as the totem the first fruits of the taro), and the idea of descent may be expanded to a belief in an animal ancestor. Some clans (particularly in Australia) specialize in the fact that they care magically for the proliferation of their totem animals. (For additional information see also ANIMISM; TOTEMISM; PRIMITIVE RELIGION.)

BIBLIOGRAPHY. Apart from some older works about particular forms of nature worship, the most important discussions of this form of religion and the related nature mythology appear in journals. JAMES G. FRAZER, *The Worship of Nature* (1926), is a classic work mainly concerned with the worship of the heavens, the sun, and the earth; and *The Golden Bough*, 3rd ed. rev., 12 vol. (1907–15; abridged ed., 1922 and 1959), contains an important discussion of nature gods, especially vegetation gods, and their cults. Important references to the worship of nature may be found under various headings in *The Mythology of All Races*, ed. by LOUIS H. GRAY, 13 vol. (1916–32); and *The Encyclopaedia of Religion and Ethics*, ed. by JAMES HASTINGS, 13 vol. (1908–26, reprinted 1955).

The heavens and earth: HERMANN BAUMANN, *Schöpfung und Urzeit des Menschen im Mythus der afrikanischen Völker*, 2nd ed. (1964), and *Das Doppelte Geschlecht* (1955); MIRCEA ELIADE, "La Terre-Mère et les hiérogamies cosmiques," *Eranos-Jahrbuch*, 22:57–95 (1954); JURGEN ZWERNEMANN, *Die Erde in Vorstellungswelt und Kultpraktiken der sudanischen Völker* (1968).

Sun and moon: The solar mythology that originated with the development of Indo-European linguistic studies places the sun at the centre of religion and myth: see G. ELLIOT SMITH, *The Migrations of Early Culture* (1929); and WILLIAM J. PERRY, *The Children of the Sun* (1923). (*General material concerning sun-worship*): FRANZ J. BOLL, *Die Sonne im Glauben und in der Weltanschauung der alten Völker* (1922). (*The sun in the cults and myths of African peoples*): HERMANN BAUMANN, "Die Sonne in Mythus und Religion der Afrikaner," in *Afrikanistische Studien* (1955). (*Moon worship*): ERNEST SIECKE, *Liebesgeschichte des Himmels* (1892). (*Other celestial bodies*): HUGO WINCKLER, *Himmels- und Weltenbild der Babylonier*, 2nd ed. (1903); ALFRED JEREMIAS, *Im Kampfe um Babel und Bibel* (1903); RICHARD H. ALLEN, *Star Names* (1963); HELMUT WERNER, "Klassische Sternbilder am Himmel der Tschuktschen," *Z. Ethnol.*, 77:139–141 (1952).

Fire and water: JAMES G. FRAZER, *Myths of the Origin of Fire* (1930), the most comprehensive work about fire myths; WINIFRED S. BLACKMAN, "The Magical and Ceremonial Uses of Fire," *Folk-lore*, 27:352–377 (1916); BEN ZION GOLDBERG, *The Sacred Fire: The Story of Sex in Religion* (1930; reprinted as *Sex in Religion*, 1970), a popular collection; ARCHDALE KING, *Holy Water* (1926); ELIZABETH D. VAN BUREN, *The Flowing Vase and the God with Streams* (1933), concerned with the ancient Orient; BEATRIX HEINTZE, "Der südrhodesische dziva-Komplex," *Anthropos*, 62:337–368 (1967), covers aspects of water cults in southeast Africa; LEOPOLD WALK, "Das Flutgeschwisterpaar . . . ," *Mitt. anthrop. Ges. Wien*, 78–79:60–115 (1949).

Earthquakes: RICHARD LASCH, "Die Ursache und Bedeutung der Erdbeben im Volksglauben und Volksbrauch I," *Arch. Religionswissenschaft*, 5:236–257 (1902); BERNHARD STRUCK, "Zur Kenntnis der afrikanischer Erdbebensvorstellungen," *Globus*, 95:85–90 (1909).

Mountains: NELLY NAUMANN, "Yama no Kami–Die japanische Berggottheit," *Asian Folklore Studies*, 22:133–366 (1963) and 23:48–199 (1964).

Weather: CHRISTIAN S. BLINKENBERG, *The Thunderweapon in Religion and Folklore* (1911); JOHN LOEWENSTEIN, "Rainbow and Serpent," *Anthropos*, 56:31–40 (1961).

Animals: JAMES G. FRAZER, *Totemism* (1887) and *Totemism and Exogamy*, 4 vol. (1910, reprinted 1968); HERMANN BAUMANN," Individueller und kollektiver Totemismus," in the *Proceedings* of the 14th International Congress of Sociology, vol. 4 (1950); and "Afrikanische Wild- und Buschgeister," *Z. Ethnol.*, 70:209–239 (1938); CLAUDE LEVI-STRAUSS, *Le Totémisme aujourd'hui* (1962; Eng. trans., *Totemism*, 1963); J.F. MCLENNAN, "The Worship of Animals and Plants," *Fortnightly Review*, 11:407–427, 562–582 (1869) and 13:194–216 (1870); OTTO ZERRIES, *Wild- und Buschgeister in Südamerika* (1954); LEOPOLD SCHMIDT, "Der 'Herr der Tiere' in einigen Sagenlandschaften Europas und Eurasiens," *Anthropos*, 47:509–538 (1952); IVAR PAULSON, *Schutzgeister und Gottheiten des Wildes (der Jagdtiere und Fische) in Nordeurasien* (1961).

(Ed.)

Naval Ships and Craft

The first craft fitted purposely to make war were undoubtedly conversions of the dugouts, inflatable bladders, papyrus rafts, or hide boats used in everyday transport. It is probable that the conversion at first consisted simply of a concentration of weapons in the hands of a raiding party. In time conversions added offensive and defensive powers to the craft itself.

As vessels became more seaworthy and more numerous, warships designed as such developed both as marauders and as defenses against marauders. In succeeding millennia the warship went through many changes, some with revolutionary impact upon civilization. Though many ships are designed for a warlike purpose, the practice of converting peaceful ships to wartime uses has also survived. Any major war demands far more fighting ships than a nation can economically maintain during peace, and, even in World Wars I and II, merchant ships as large as passenger liners were converted into cruisers. The first craft designed and built especially for combat may have sailed in the fleets of Crete and Egypt 5,000 years ago. The word craft, it may be noted, has come increasingly to be applied to smaller vessels, with the term ship reserved for larger ones.

Naval vessels to World War I

THE ANCIENT WORLD

First warships

Early period. *Egypt.* The first recorded appearance of warships is on the Nile, where Egypt's history has centered since antiquity. Built of bundles of reeds lashed together to form a narrow, sharp-ended hull and coated with pitch, these were hardly suited for tempestuous seas; by 3000 BC larger wooden seagoing versions of the reed craft sailed for distant cruising, trade, and conquest.

Hieroglyphics recount the dispatch of 40 armed ships to Byblos on the Phoenician coast about 2900 BC to buy cedarwood (from the famous cedars of Lebanon) to build ships. These vessels had both oars and sails, being fitted with a bipod (inverted V) mast and a single, large, square sail. The whole mast could be lowered when under oars. Large Egyptian ships had over 20 oars to a side, with two or more steering oars. The war galley was built to the same pattern but was of stouter construction; modifications that could be easily incorporated in a merchant hull under construction included elevated decks fore and aft for archers and spearmen; in later Egyptian warships, a wider sail that could be furled; in the 12th century BC, planks fitted to the gunwales to protect the rowers, and a small fighting top high on the mast to accommodate several archers; and on some galleys a projecting ram, well above the waterline, which may have been designed to crash through the gunwale of a foe, ride up on deck, and swamp or capsize him.

Crete. By about 2000 BC Crete had evolved into a naval power exercising effective control of the sea in the eastern Mediterranean. Little record exists of Minoan seapower, yet these maritime people are known to have had an organized navy that cooperated with Egypt in suppressing piracy. They may also have been the first to build a warship designed as such from the keel up, rather than as a modification of a merchant ship. Sometime before their sudden and mysterious overthrow by the Mycenaeans from the Greek mainland in the 15th century BC, the Minoans began to differentiate between war craft and merchantmen and between the rowing galley and sailing vessel.

Sometime in this 2nd millennium BC, the commodious merchantman evolved as a beamy "round ship" powered by sails and emphasizing cargo capacity at the expense of speed. Fast fighting craft developed both as predator on and protector of maritime trade and coastal cities. This "long ship" was narrower, faster, and more agile than the tubby cargo ship. She hoisted her sails for cruising but depended on oars when in action.

The Cretan warship had a single mast and a single bank of oars. The sharply pointed or "beaked" bows suggest an emphasis on the tactical use of the ram. Relatively little else is known about the fighting craft, but the lack of defensive walls around Cretan cities may testify to the navy's effectiveness.

Phoenicia. Beginning about 1100 BC, the Phoenicians dominated the eastern Mediterranean for about three centuries and sailed beyond the Pillars of Hercules into the Atlantic. Information about Phoenician ships is fragmentary and somewhat inconclusive, though they appear to have been built primarily for trading, with a capacity to fight effectively if necessary. What little is known about their vessels comes from highly stereotyped representations.

These indicate that the Phoenician trading ships were galleys, mounting a single pole mast with a square sail and with steering oars to port and starboard. The war galleys show a Cretan influence: low in the bow, high in the stern, and with a heavy pointed ram at or below the waterline. Both these and the trading galleys are depicted as two-banked. The rowers of the upper bank may have sat on thwarts mounted on sponsons—*i.e.*, long platforms projecting outward above the waterline. This would have meant a narrow underwater hull with a high length to beam ratio, resulting in a slender and graceful vessel. Oars could be carried in a staggered, two-bank arrangement, allowing more oars to be mounted in a ship of a given length and increasing power and momentum.

Because the ram was the principal weapon, both the finer lines and greater rowing power were important in providing more speed for the decisive shock of battle. Coins of Sidon and Assyrian sculptures show the staggered-bank oar arrangement that the Phoenicians probably introduced.

Greece. Unlike the Egyptians, for whom wood was scarce and costly, the Aegean peoples had an abundance of timber for shipbuilding. Their earliest warships may have been large dugouts, perhaps 65 feet (20 metres) long, fitted with outriggers (timber frames built out from either side of the hull) for stability.

Tactically, they and their successors for centuries were used more to carry attack personnel than as fighting vessels. No mention is made in the *Iliad*, for instance, of sea warfare. Even the pirates of the time were sea raiders seeking their booty ashore, like the later Vikings, rather than in sea actions. The so-called long penteconter, mentioned by Herodotus, was employed in exploring, raiding, and communicating with outlying colonies. Light and fast, with 25 oars to a side, it even more than the "round-built" merchant ship, played an important role in the early spread of Grecian influence throughout the Mediterranean. As the Greek maritime city-states sped the growth of commerce and thus the need for protection at sea, there evolved a galley built primarily for fighting. The first galleys, called uniremes (Latin *remus*, "oar"), mounted their oars in a single bank and were undecked or only partially decked. They were fast and graceful with high, curving stem and stern. In Homeric times some carried an *embolon*—a beak or ram—that became standard in succeeding centuries.

Biremes and triremes. The bireme, probably adopted from the Phoenicians, followed and became the leading warship of the 8th century BC. Greek biremes were probably about 80 feet (24 metres) long with a maximum beam around ten feet (three metres). Within two or three generations the first triremes (ships with three banks of oars) appeared. This type gradually took over as the primary warship, particularly after the great sea victory of Salamis (480 BC).

The first triremes

Like its predecessors, the trireme mounted a single mast with a broad, rectangular sail that could be furled. The mast was lowered and stowed when rowing into the wind or in battle. Built on an entirely different system from the Egyptians—with keel, frames, and planking—these were truly seagoing warships.

The arrangement of oars has fed a scholarly controversy for generations. Two principal theories survive. One holds that there were, in fact, banks of benches for oarsmen—two for the bireme, three for the trireme—and that oarsmen of upper rows sat on thwarts slightly forward of the corresponding rower below. The other school argues that the thwarts ran at an angle obliquely from the galley's centre line. For a trireme, two oarsmen sat alongside each other on an upper thwart extending outward into a sponson with their oars projecting through ports in the sponson itself. The third oarsman in the group sat on a lower thwart with his oar projecting through a port in the hull.

After Salamis, the trireme continued as the backbone of the Greek fleet, with the ram continuing as the primary weapon. Its keel, like those of its predecessors, formed

the principal-strength member, running the length of the ship and curving upward at each end. The ram, usually shod in bronze, formed a forward prolongation that gained effectiveness from the heavy keel back of it. Additional longitudinal strength came from a storming bridge —*i.e.*, a gangway along the centre line from bow to stern along which the crew raced to board when a foe was rammed. Gradually, with ships becoming steadily heavier, boarding assumed greater importance and the ram lost some of its importance. Many war galleys also had full spar decks instead of the centre-line gangway.

A trireme of the 5th century BC may have had a length of 125 feet, a beam of 20 feet, and draft of three feet (about 38 × 6 × 1 metres). Her great length required extra longitudinal strength, which was provided by wales (belts) around the outside of the hull. Manned by about 200 officers, seamen, and oarsmen (perhaps 85 on a side), with a small band of heavily armed *epibatai* (marines), she could reach seven knots under oars and was a formidable foe.

Rowers in the ancient world, into early Roman times, were free crewmen rather than slaves. The captain of a Greek trireme was usually a wealthy man, a political appointee, who could afford to fit out and operate the warship for a year. The navigator was an experienced seaman.

With only scant room for provisions, such warships could not remain long at sea. A voyage usually consisted of short hops from island to island or headland to headland. Even the largest triremes put into shore and beached, stern first, for the night, resuming the passage in the morning, weather permitting. Light construction and little endurance made short distances between bases essential and frequent refits imperative.

Early Greek naval battle tactics were rudimentary. The basic tactical disposition was simply two opposing battle lines of galleys drawn up on line abreast, each trying to overwhelm the other by ramming and boarding. By the 5th century BC, however, skilled fleet commanders dared to essay the manoeuvre of "breaking through the line," or turning the foe's flank. Breaking through the enemy's battle line consisted of a sudden thrust at the centre of the opposing fleet. Piercing it, the attackers would then wheel to right and left behind the gap to take the rest of the line in the rear. For success, the manoeuvre required speed, surprise, and great coordination that only much training could achieve.

Evolution of naval tactics

Later developments. Athens employed quadriremes (four-bank rating) by the middle of the 4th century BC, with quinqueremes appearing soon thereafter. In the late 4th and early 3rd century BC an arms race developed in the eastern Mediterranean, producing even larger multibanked ships. Macedonia's rulers built 18-banked craft requiring crews of 1,800 men. Ptolemaic Egypt capped them with 20's and 30's. Ptolemy III even laid down a 40 (tesseraconter) with a design length of over 400 feet and calling for a crew of 4,000 rowers. The vessel was never actually used. The multiplicity of "banks," once a puzzle to historians, signifies the number of rowers on each oar rather than an almost unimaginable piling up of banks. While representations of biremes and triremes with staggered oars exist, nothing portrays any further superimposition of oars. The ratings were of "man-power" propulsion banks, similar to modern horsepower ratings.

This same arms race brought other changes of significance. Until the late 4th century BC, manoeuvre, marines, and the ram constituted a warship's offensive strength. Archers provided close-in fire, though perhaps not to the same extent as in the early Egyptian navy, in which the bow was favoured. Demetrius I Poliorcetes of Macedonia is credited with introducing heavy missile weapons on ships, at the end of the century, starting a trend that still has not ended.

Demetrius' ships mounted crossbow-like catapults for hurling heavy darts, and stone-throwing machines of the type the Romans later called ballistae. From this time on large warships carried these weapons enabling them to engage a foe at "standoff" ranges, though ramming and boarding also continued. Temporary wooden turrets—forecastles and sterncastles—were similarly fitted to provide elevated platforms for archers and slingers.

Following the fragmentation of the brief empire of Alexander the Great, sea power developed elsewhere. The city-state of Rhodes built a small but competent fleet to protect its vital shipping. Rhodian warships were reported to have made use, in at least one action, of firepots suspended on spars projecting from their bows, anticipating Greek fire by several centuries.

Meanwhile to the west, Carthage, a state with ancient maritime origins, rose to prominence on the north coast of Africa, and by about 300 BC had become the foremost Mediterranean naval power. Carthage's navy consisted probably of the same ram-galley types developed by its ancestral Phoenicians and by the Greeks.

Rome. Coincidentally, across the sea to the north the city-state of Rome expanded to include most of the southern Italian peninsula with its extensive seacoast and maritime heritage. Rome's growth southward collided with Carthage's ambitions in Sicily, leading to the First Punic War, which began in 264 BC. Unlike their seafaring opponents, the Romans were not a naval power. When in the fourth year of the war Carthage sent a fleet against Sicily, Rome realized her fatal disadvantage and moved to remedy it. The Greek city-states she had conquered had long seagoing experience. Employing their shipbuilders and learning also from the foe, Rome built a fleet of triremes and quinqueremes, the latter patterned after a Carthaginian warship that ran aground in Roman territory early in the war. While galleys were being built, oarsmen trained on dummy rowing benches ashore (an early ancestor of the training devices of modern navies). The first Roman battle fleet with effective crews soon put to sea.

Not content with copying the enemy's tactics, the Romans took land warfare to sea and forced the Carthaginians to fight on Roman terms. Each Roman galley had fitted in the bow a hinged gangplank with a grappling spike or hook (the *corvus*) in the forward end, thus providing a boarding ramp. They added to the crews many more marines than warships usually carried.

Roman techniques

The Phoenicians and Greeks had emphasized ramming, with boarding as a secondary tactic. A Roman captain rammed and then dropped the gangplank. Ram and *corvus* locked the galleys together, and the Roman marines boarded, overwhelming the opponent. The Roman fleet had extraordinary success in the great naval battle of Mylae off northeast Sicily destroying or capturing 44 ships and 10,000 men.

After other victories, and some defeats, by the end of the First Punic War, 241 BC, Rome had become the leading sea-power. A generation later, in 218 BC, Rome again engaged in war with Carthage. The genius of Hannibal almost overwhelmed Rome and probably would have had Carthage put forth major effort at sea and supported him in Italy. When the war ended in 201 BC, Carthage surrendered its navy and was never again to become a maritime power.

At the end of the Second Punic War Rome began a long effort to subdue the pirates of the eastern Mediterranean. After several generations of intermittent operations, in 67 BC Pompey was given a three-year commission to finish the task. The Mediterranean was divided into "naval districts"; regular patrols were established; the sea was systematically cleared and held. This operation helped convince Roman rulers of the need for a standing fleet.

As the Roman navy evolved, so did its warships. Though pictorial evidence is ambiguous, it seems clear that the gangplank and *corvus* disappeared as the Romans gained experience in sea warfare. Later Roman warships appear to have been conventional fully decked ram-galleys mounting one or two wooden turrets (probably dismountable) for archers. To the single mast with rectangular sail was added a bowsprit carrying a small sail, the artemon. Falces, or long spars tipped with blades, were used by Julius Caesar's fleet against the sailing vessels of the Veneti of northwest Gaul to cut their rigging and immobilize them. Catapults and ballistae served as "artillery," and it was under their fire that Caesar's legions landed in England.

The light liburnian

Early Roman warships were all large; to escort merchantmen and combat pirates Rome found need for a lighter type, the liburnian. Probably developed by the pirates themselves, this was originally a light, fast unireme; the Romans added a second bank of oars. In the Battle of Actium, 31 BC, Octavian's (later Augustus') skilled fleet commander, Agrippa, used his liburnians to good effect.

The Byzantine empire. Although polyremes continued to be built after Actium, the liburnian became the predominant Roman man-of-war. With the break-up of the western Roman Empire, naval organization and activity in the west decayed. In the eastern Roman Empire, however, the need for sea power was well appreciated. During the 11 centuries that the Roman Empire centred on Constantinople, the Byzantine rulers maintained a highly organized fleet. Their original type of warship was the liburnian, called in Greek the *dromōn*, or runner, a word that like liburnian, eventually came to denote any warship; it was built in several different sizes; the heavier designed to bear the weight of battle; the lighter single-bank dromons serving as cruisers and scouts.

Throughout the eastern Roman Empire's existence warships changed little except in rig and armament. An average large dromon measured up to 150 feet (45 metres) in length, with 100 oars and one or two fighting towers for marines. Other than some widening of the stern and its superstructure to provide more accommodation aft, the hull differed little from the Roman type. At some point early in the Christian Era, the lateen sail, three-cornered and suspended from an angled yard, probably adopted from the Arabs, came into general use. Eastern warships had two or three masts. In a departure from classical customs, these were left in place in battle. Contemporary pictures show rams above the waterline. Oars may have projected directly from the ship's side rather than from an overhanging sponson.

Missile-launching weapons grew in size, some hurling projectiles as large as 1,000 pounds (450 kilograms) up to 750 yards (700 metres). Greek fire, a combustible material for setting fire to enemy ships, was invented in the 7th century or earlier. The various compounds passing under the name used a blend of some of the following: pitch, oil, charcoal, sulfur, phosphorus, and salt. The pitch was inflammable resin, the oil, usually crude petroleum or naphtha, the salt probably crude nitrate of potash. As its composition improved, tubes shaped into the mouths of savage monsters were placed in the bows of war galleys and the flaming terror, which water merely spread, was hurled on the enemy. At close quarters marines also used a type of hand projector. In later tubes the charge was placed in a brass cup resembling a beer stein and inserted in the breech end of the tube. Significantly, the first guns (except handguns) followed the same practice.

Greek fire was an important factor in terrifying and repelling the Muslim fleet in sieges of Constantinople from the early 8th century on.

EARLY WAR VESSELS OF NORTHERN EUROPE

The earliest northern warships of which record exists were the sailing ships used against Julius Caesar by the Veneti of Gaul. During the Roman period, ships in European waters were of the Roman pattern—war galleys (mounting mast and sail but basically rowing vessels) and merchant ships. Like their Greek predecessors, the latter were slow, beamy sailing ships designed for cargo capacity rather than speed and manoeuvre. Like warships, the merchantmen mounted one mast with a single rectangular sail. They also hoisted two triangular topsails spread between the yard and masthead, as well as the artemon. Large Roman merchantmen were over 175 feet (50 metres) long, with a beam of perhaps 50 feet (15 metres).

Viking vessels. In addition to dugouts, early Scandinavian craft are believed to have included craft built of hides stretched over wood frames. These proved suitably strong and seaworthy for coastal operations. By the beginning of the Viking period, about AD 800, the early craft had evolved into the well-known Viking ship, a sturdy, double-ended, clinker-built (*i.e.*, with overlapping planks) galley put together with iron nails and caulked with tarred rope. It had a mast and square sail, which was lowered in battle; high bow and stern, with removable dragon heads; and a single side rudder on the starboard (steer-board) quarter. This method of construction had existed for about five centuries, though a keel providing strength for distant open-water operations and a rudder actually attached to the hull first appear in a recovered boat dating around AD 600; the mast and sail seem to have come into use later.

Viking vessels were essentially large open boats. Like the Homeric Greeks, the Vikings at first made no distinction between war and cargo ships, the same vessel serving either purpose as the occasion demanded. Later, however, they built larger ships specifically designed for war. By AD 1000 they sailed three categories of these: those with less than 20 thwarts (40 rowers); those with up to 30; and the "great ships" with more than 30, which might be considered the battleship of the time. Expensive and unhandy, though formidable in battle, the great ship was never numerous. The middle group, manoeuvrable and fast, proved most valuable.

These long ships played an important role in exploration (reaching Greenland and America before Columbus), in the consolidation of kingdoms in Scandinavia, and in far ranging raids and conquests. It was in them that the Norsemen early invaded the British Isles and established themselves in Normandy, whence their descendants under William I, the Conqueror crossed the Channel to win England in 1066.

English warships. English war vessels reflect developments that were going on elsewhere in western Europe. Early English warships resembled their Viking counterparts. Rather than being built specifically for fighting, most were called into national service only in time of emergency, though King Alfred built ships "full twice as long" as previous British craft, swifter and higher in the water, to battle the Danes at sea.

To about the end of the 13th century, the typical ship in northern waters remained a clinker-built, single-masted, square-rigged descendant of the long ship. In this century, and even more in the 14th, changes began that would bring an end to the long dominance of the oar in battle. Around AD 1200 came one of the great steps in the history of sail: the introduction, probably in The Netherlands, of the stern rudder. This rudder, along with the deep-draft hull, the bowsprit and, in time, additional masts, transformed the long ship into the true sailing ship that could beat into the wind as well as sail with it.

Thus the beamier round ship also became the warship. Fleets of this period, and indeed through the 16th century, were owned for the most part by feudal lords, merchants, and port cities, though some were royal ships. In time of peace the king's, as well as private, vessels were used as merchantmen and manned by civilian crews. When a need arose, the king mustered his own ships, called on vassals for theirs, and hired whatever others he needed. Temporary wooden castles added at bow and stern provided bulwarked platforms for archers and slingers. A complement of men-at-arms embarked, in addition to the ship's seamen. Tactics were usually simple and straightforward, opposing fleets closing and attempting to beat down each other's archers before grappling and boarding. At war's end, off came the castles, and the ships went back to trading.

Warships of feudal lords

THE AGE OF GUN AND SAIL

The trading vessel that could be promptly adapted to war did not, however, fulfill the need of a nation for a navy. The coming of gunpowder and the period of world exploration brought changes that were to cause the sailing man-of-war to become increasingly distinct from the merchantman.

Until the 15th century, northern ships probably continued to have single masts, though in the Mediterranean a two-mast rig carrying lateen (fore-and-aft) sails had existed for some time. Then change came rapidly in the

north, spurred on by Henry V's construction of large and strongly built warships for his cross-Channel French campaigns. The remains of one of these, "Grâce Dieu," reflected the clinker-built construction of the Viking long ship, but had a keel to beam ratio of about 2½:1 and now carried a second mast. The removable castles had become permanent parts of the ship, mounting many small guns.

Some historians believe that "Grâce Dieu" carried a third mast. At any rate, in a few decades ships had three and, by the end of the century, large vessels mounted four masts carrying eight or more sails. A three-master carried a large sail on each mast and in addition a main topsail and the spritsail under the bowsprit—the rig, in fact, of "Santa Maria" in 1492. Ships, no longer dependent on fair winds, could and did range the world.

Gun-armed warships. Employment of guns afloat, bringing a slow but progressive revolution in warship construction and naval tactics, had its first small beginnings by the 14th century. One of the compositions of Greek fire, given by Marcus Graecus in AD 846, was one part live sulfur, two of charcoal of willow, and six of saltpetre (potassium nitrate) ground into fine powder in a mortar and mixed. This recipe is a forerunner of black powder, the basic gun propellant until the late 1800s.

Gradually, saltpetre's properties became more widely known. In the 13th century, Mediterranean Muslims made hand grenades and fired small balls from hand tubes or lance heads. The first gun used at sea, undoubtedly a hand weapon, was probably in a Mediterranean galley in the 13th or early 14th century. It played a minor role. In fact, in the numerous sea battles of the Greeks, Genoese, Moors, Turks, and Venetians during this period there is no mention of guns. But by the middle of the 14th century, the English, French, Spanish, and other navies mounted guns. Most were relatively small "man-killers" located in the castles fore and aft. Later, heavier guns were added. The Mediterranean galleys of Turkey and Spain at first simply shackled a heavy gun on the forecastle to fire forward, supplementing the ram.

First guns used at sea

European guns were originally built up of wrought-iron bars welded together to form a tube, then banded with a thick iron hoop. Initially, they were breechloaders with an open trough at the rear of the barrel through which the ball was loaded and a cylindrical chamber filled with powder inserted and wedged tight. Inherently weak and low-powered, they were replaced in the British navy after 1500 by brass muzzle-loaders, cast in one piece with bores drilled out. Some of these attained great size for their day; by the mid-16th century even some 60-pounders (firing 60-pound solid shot) were mounted in the largest ships. In this century also, increasing knowledge of iron metallurgy led to the production of large cast-iron cannon that slowly replaced the bigger brass guns in ships, though brass remained predominant for the lighter calibre well into the 19th century.

Henry VII of England created the first true oceangoing battle fleet. The "king's ships" carried many guns but most of the impressively large numbers were small breechloaders. Following him, Henry VIII initiated gunports in English warships, a development that was to have a far-reaching effect on man-of-war design and, indeed, on world history. Neither stability nor structural strength favoured heavy guns in the high castles. Henry's introduction of gunports (perhaps following an earlier use across the Channel), at first low in the waist of the ship and afterward along the full broadside, made possible the true heavy-gun warship. The cutting of gunports in the hull must also have been a factor in causing the northern nations to shift from clinker-built ships to caravels with flush-fitting planks, a change that took place in the early 1500s.

The armament of an English man-of-war of the early 16th century consisted of four or five short-barrelled cannon, or curtals, a similar number of demicannon, and culverins. The average cannon, a short-range gun, hurled an iron ball of about 50 pounds, and the demicannon one of 32 pounds. The culverin, a longer and stronger gun, fired a smaller shot over a longer range and was likely to be more accurate at other than point-blank range. Supplementing these standoff "ship killers" were the smaller demiculverin, saker (quarter culverin), falcon (half saker), falconet, robinet, in descending size of ball fired, down to several ounces.

A great warship of the 16th century mounted a total of large and small pieces approximating the numbers mounted in battleships of World War II. Henry VIII's best known warship, "Henry Grâce à Dieu," had 186 guns for her original complement in 1514, mostly small, but a number of iron "great guns."

As the 17th century advanced, guns and powder improved and classification became standardized. The long gun (culverin type) ran 18 to 25 calibres (bore diameters) in length; the short cannon, 15 calibres. Gun carriages had heavy wooden sides called brackets, with sockets for the gun trunnions, joined by similar flat timbers called transoms, supported on wooden trucks and hauled out after recoil by heavy tackle.

Galleasses, galleons, and great gun battles. The coming of mighty men-of-war, like "Henry Grâce à Dieu," did not mean the immediate end of oared warships. In fact, some types of galleys and oared gunboats continued to serve well into the 19th century. Venice had galleys when conquered by Napoleon. Throughout the 18th century Russia and Sweden employed galleys and, later, oared gunboats. Indeed, Russian-oared gunboats were built in 1854 and engaged in action that year. The United States employed "row galleys" or oar-driven gunboats in the War of 1812 and the Mexican War (1846–48).

Survival of oared warships

Galleasses at Lepanto. The climactic action of the age of oar was Lepanto, fought off Greece in 1571, in which a combined European fleet defeated the Turkish fleet in an action that differed little from traditional galley warfare with two exceptions. First, galleys now carried guns, perhaps five heavy and medium guns forward to supplement the shock effect of the ram and more small breechloaders complementing bows, lances, and swords.

Second, the European line of battle included six Venetian galleasses, a compromise type developed in the transition from oar to sail. These huge vessels that depended on sail as well as oar, bristled with guns, including heavy ones in broadside. Although cumbersome to manoeuvre, their concentrated fire contributed importantly to victory.

Galleasses outside the Mediterranean differed somewhat from Venice's in that they were basically full-rigged sailing ships carrying broadsides of heavy guns and a bank of auxiliary oars for mobility. The hybrid existed only in small numbers and soon passed out of fashion to the north.

The galleon and the Spanish Armada. The "great ships" of Henry VII and Henry VIII were carracks: starting basically with the lines of beamy, seaworthy merchant ships, designers had added stronger timbers, masts, sailpower, broadside guns, and forecastles and aftercastles, with the forecastle triangular and overhanging. In the galleon, the successor to the carrack, the general principles of design of sailing men-of-war were established and ruled, without fundamental change, for three centuries. The galleon retained certain of the characteristics of the galley, such as its slender shape; it had a greater length to beam ratio than the carrack, standardizing in England for a time at about three to one but later increasing to considerably greater ratios. The triangular, overhanging forecastle of the carrack gave way to the long raised beak of the galley projecting beyond the stem and with a square-ended forecastle set back under the foremast. Like carracks, the larger galleons might carry a single mizzenmast or two relatively small masts, the second being called the bonaventure.

In the longer, leaner galleon, the number of heavy guns was increased until they ran the full length of the broadside in one or two tiers (and later three). This advance in firepower was incorporated in "Henry Grâce à Dieu" during its rebuilding in 1536, though it remained a carrack.

The galleon came into favour in the north during the following decades, leading to the first ships of the line. The far-ranging experience of mariners and improved construction techniques led to great fighting ships that

were both lower in the water and more seaworthy than their predecessors. They also carried comparatively large batteries of heavy guns, affecting the design of ships. The sides now sloped inward from the lowest gun deck up to the weather deck. This tumble home helped concentrate the weight of the big broadside guns toward the ship's centre line, improving stability. Sir Francis Drake's flagship, "Revenge," for example, displaced only 500 tons, yet she mounted a broadside of 34 guns ranging from 32-pounders to nine-pounders, plus smaller guns.

Heavy gun batteries

By that time it had become normal for northern warships to mount powerful broadsides of 28 or more ship-smashing guns, a much heavier armament, in proportion to their size, than their predecessors. For their handy, manoeuvrable ships, the British had also concentrated on the long-range-culverin type and by the time of the Armada most of their principal ships carried these in broadsides. Thus they were designed for off-fighting, permitting the English fleet to get the most out of its superior manoeuvring qualities. These early ships of the line began the true reign of the big gun at sea, maintained until World War II when the carrier and submarine became more formidable. Today the rocket missile (ballistic and guided) has taken the gun's leading role, except in amphibious warfare.

When the Spanish Armada arrived in 1588, the British sought and fought a sea battle with ship-killing guns, rather than the conventional fleet engagement of the past that concentrated on killing men, ramming, and boarding. With superior ability and long-range culverins, they punished the invading fleet outside the effective range of the heavy but shorter range cannon the Spanish favoured. This historic running battle of July 1588 closed one era and opened a greater one of big-gun sailing navies.

Ship of the line. The late Elizabethan galleon that began the true fighting ship of this type reached its culmination in England's "Prince Royal" of 1610 and the larger "Sovereign of the Seas," along with similar great ships in other European navies. These two English ships mounted broadside guns on three decks, "Sovereign of the Seas," most formidable ship afloat of its time, carrying 100 guns. This was the standard minimum armament for a first-rater long into the future.

In this mobile fortress of approximately 1,500 tons there was some reduction of height and the bonaventure mizzen disappeared, leaving the standard three masts that capital ships thereafter carried.

Soon ships began to be standardized into different categories. James I organized his ships into four ranks, and, by the mid-17th century, six rates existed as a general concept, though not yet a system. The number of guns carried determined rate, with a first-rater mounting 100 and a sixth-rater 18. An important improvement came in the standardization of batteries in the higher rates so that guns on the same deck were of the same weight and calibre rather than mixed, as originally in "Sovereign of the Seas." Near the end of the century, guns began to be described by their weight and calibre, with the 32-pounder long gun favoured as the standard lower deck weapon for British warships, though three-deckers might carry 42-pounders in the lowest tier.

Rates were periodically revised. By the middle of the 18th century, while the rate concept continued for some official purposes, ships were generally known by the number of guns they were rated to carry—*e.g.*, "Victory," 108. Similar developments took place in other navies. The French seemed to emphasize scientific shipbuilding more than the British, hence class for class their warships were often more efficient. The Dutch and Spanish also built excellent warships, the former favouring shallower draft vessels.

Tactics and guns. The frequent hard-fought sea battles of the 17th century between fleets closely matched in size and skill, particularly in the Anglo-Dutch wars, led to the column formation of heavy warships called line ahead. In 1653 the British Admiralty prescribed the line ahead as the principal combat formation, with the intention of enabling the commander to control the fleet and bring the entire force into action at the same time.

Anglo-Dutch formations

Two schools of British Navy thought existed as to the tactics of engaging the enemy. The formalists urged aligning the full line of battle from van to rear parallel with the foe before bearing down in a concerted manoeuvre that would engage all of the ships at the same time. The second school believed in prompt attack, closing the enemy quickly after a formal approach had brought the fleets near enough together, whether or not in exact alignment. They held that speed of onslaught more than compensated for lack of a completely concerted attack.

Both schools believed in fighting at close quarters where the fire of even the smallest guns could be effective. This shift from the tactics against the Spanish Armada did not do away with the long gun, but in time led to the introduction of larger bore, shorter range guns like the cannon of the early period. The change in viewpoint may have been strengthened in the world wars of the 18th century by the tendency of the French and Spanish to concentrate on damaging upperworks and rigging. They fired on the uproll, while the British, concentrating on the hull, fired on the downroll to smash the ship.

A large ball has greater crushing effect, but factors like those of recoil, weight, and metallurgy considerations required a short-bore gun. Late in the 18th century, the Scottish firm of Carron Company Ironworks brought out "The Smasher" of eight-inch (20-centimetre) bore, firing a close fitting 68-pound (31-kilogram) ball with only 5½ pounds (2½ kilograms) of powder. Carronades, as they came to be called, had a short bore. They gave ships massive close-range firepower without the weight and recoil problems of similar long guns. They could not replace the long gun and were valuable only because navies, in a sense, continued the tactics of the Romans—smashing with heavy artillery and boarding.

Ship design changes. Warships gradually improved in design through the 17th and 18th centuries. New types of sails, providing more canvas and more versatile combinations for varying weather conditions, such as staysails and the jib sail, came into use in the 17th century. Soon thereafter the steering wheel replaced the old whip staff, or tiller.

The adoption of line-ahead tactics made it necessary to standardize the battle line, which had consisted of ships of widely varying strength. Now only the more powerful warships were considered suitable "to lie in the line of battle." Hence the origin by the 1700s of the term line-of-battle ship, or the ship of the line, and, in the second half of the 19th century, the derived term battleship—ships that could hit the hardest and endure the most punishment.

Some first-raters were built to carry as many as 136 guns, but the biggest ships were often cumbersome, hence relatively few were built. The handier 74-gun third-rater proved particularly successful, combining sufficient hitting power with better speed and manoeuvrability. Most of the ships of the line of the late 18th and early 19th centuries were 74s. One of these might be approximately 175 feet (50 metres) long with two full gun decks, the lower mounting the heaviest guns, by the Napoleonic Wars usually 32-pounders. The upper gun deck customarily carried 24-pounders, while the forecastle and quarterdeck mounted lighter guns or carronades. The bigger ships were similar but had three covered gun decks instead of two. Viscount Nelson's "Victory," built in 1765 and preserved in dry dock as she was at Trafalgar in 1805, is a classic example of this powerful type.

The 74-gun third-rater

Frigates and other types. Ships of the line, first to fourth rates, had strong fast frigates as consorts. This ancestor of the modern cruiser evolved during the mid-18th century for scouting, patrol, and escort, as well as for attacking enemy merchantmen. The frigate carried its main battery on a single gun deck with other guns on forecastle and quarterdeck. Like the ship of the line, they varied in size and armament, ranging from about 24 guns in early small frigates to as many as 56 in some of the last. Two classic examples, still preserved, are the United States Navy's "Constitution," of 44 guns, and "Constellation," of 38.

Smaller vessels aided frigates in their blockade, escort, commerce raiding, and other duties. The single-masted cutter served as scout and coastal patrol craft. Brig and schooner-rigged types, generally called sloops of war, by the time of the American Revolution grew into the three-masted, square-rigged "ship sloop." Called a corvette on the Continent, the fast ship sloop complemented frigates on the fringes of the fleet. Smaller sloops, schooners, brigs, and luggers were widely used for special service. Fleets also needed ordnance and supply ships and other auxiliaries; these were usually merchantmen taken into service in war emergency. Converted merchantmen, such as John Paul Jones's "Bonhomme Richard," often played combat roles. Fleets also had various special types, such as fire ships and bomb ketches. The latter, with two large mortars hurling bombs of about 200 pounds, were developed by France in the late 1600s, and were used with devastating effect against Barbary pirate ports.

THE AGE OF STEAM AND IRON

As the Industrial Revolution unfolded in the 19th century, the age of wood and sail gave way to that of steam and iron. Phenomenal changes took place in nearly every aspect of warship design, operation, and tactics. These changes ended the reign of the majestic ship of the line by the mid-1800s, but another half century elapsed before it was clear what form her replacement as the backbone of fleets would take. The changes may be summarized under two headings: propulsion and ship construction and armament.

Propulsion. Steam for propulsion of vessels was tried with varying success in several countries during the late 18th century. Engines and supporting machinery were at first not adequate for this fundamental advance in ship capability, but useful steam craft appeared in the early 1800s, suitable for operation on inland and coastal waterways. The earliest steam warship was "Demologos" of the U.S. Navy (renamed "Fulton" after her designer). Built in the War of 1812, this well-gunned, double-hulled, low-powered ship cruised briefly in the New York Harbor area before the war ended and later was destroyed by an accidental fire.

First steam warship

The earliest steam warships in action were small paddle wheelers used by British and American navies against weak foes. USS "Sea Gull," a New York Harbor steam tug, operated against pirates in the West Indies during the 1820s.

As engines gradually improved, navies experimented with them in standard warships, first as auxiliaries to sail, which was then essential for endurance. The paddle-wheels were particularly vulnerable to enemy fire. In 1843, through the drive of Capt. Robert Field Stockton of the U.S. Navy and the inventive skill of John Ericsson, a Swede whom Stockton brought to America, the United States launched the world's first screw-driven steam man-of-war, USS "Princeton," a large 10-gun sloop.

The screw propeller was an old idea going back to Archimedes, but, with Stockton's assistance, Ericsson had made it effective for large warships, as Sir Francis Pettit Smith was doing at about the same time in England for large merchantmen. By the mid-1840s, boilers, engines, and machinery had improved to the point that thereafter practically all of the new warships had steam propulsion. The French Navy's "Napoléon" was a particularly successful application of steam to a great ship of the line. The U.S. Navy's experience in the Mexican War of the 1840s and the allied experience in the Crimean War of the 1850s demonstrated the outstanding advantages of steam warships for restricted-waters operations, as well as in adverse winds and tides.

Among the advances of this period were two other milestones. In 1834 Samuel Hall of England patented a type of steam condenser that made it possible to use fresh instead of corrosive salt water for boilers. In 1824 James Peter Allaire of the United States invented the compound-expansion steam engine, in which the steam was used in a second cylinder at a lower pressure after it had done its work in the first. Eventually it was made practical by progress in metallurgy and engineering; in 1854 John Elder, shipbuilder on the Clyde, installed a successful two-stage engine in the merchant steamer "Brandon." The higher efficiency was of great importance for ocean-keeping navies.

Ship construction and armament to the American Civil War. The change from wood to iron and armour that followed the introduction of steam propulsion came slowly, in considerable part because the new material required new techniques and experience in shipbuilding. By mid-century a handful of iron ships had appeared in various navies; the general use of iron for warships, however, awaited realization of the full value of the shell gun, the resulting development of armour, the employment of armoured batteries in the Crimean War, and the recognition of the significance of the battle between "Monitor" and "Virginia," better known as "Merrimack."

Slow change from wood to iron

Guns. "Princeton" not only set the course for propeller warships but exemplified the rapid development of armament that with steam and other factors brought navies increasing capabilities against each other and in operations against land defenses. She mounted 12-inch shell guns, which still had many unsolved manufacturing problems; in a gala demonstration for Pres. John Tyler one of the guns blew up, killing the Secretary of the Navy and several others.

Basic changes in armament that would advance rapidly throughout the 19th and 20th centuries had begun in the 18th century. In the British Navy steps to make possible heavier long-range guns began with the introduction of strong springs to take up the first shock of recoil, aided by inclined-plane wedges behind the trucks. Flintlocks pulled by a lanyard, instead of a match, fired the guns. Sights also improved. In the early 1800s navies began to employ mercury fulminate in caps to initiate firing. Efficient percussion locks came into use within a few years.

Smooth-bore guns were still inaccurate, and successful efforts were made to bring back the rifling, as well as breech-loading, of early guns, thus increasing speed and accuracy of fire. In the 1840s, Italian and Austrian inventors brought out sliding-wedge breechblocks. Later the French developed an interrupted screw system, originally an American invention. A British firm produced a rifled breech-loading gun that the British Navy used until 1864, when a number of accidents brought a temporary reversion to muzzle-loaders. But defects were eventually remedied, and breech-loading brought phenomenal increases in naval rates of fire.

French 6.5-inch (16.5-centimetre) cast-iron rifles in the Crimean War demonstrated superiority in range, destructive power, and accuracy. Along with the advent of armour, they helped impress all of the navies with the need for rifling. Slower burning powder was badly needed. Black powder was gradually improved during 600 years of use in firearms, but it still retained its primary defect, too rapid burning. This required keeping the charge down (and therefore the range) to prevent the bursting of even the best guns. Just before the American Civil War the U.S. Army developed large, perforated, dense grains of black powder that were a start toward the controlled burning ultimately achieved with smokeless powder.

Shell guns in warships' main batteries were preceded by bombs fired from mortars, small-shell guns, and solid hot shot heated to cherry red. A principal architect in bringing big-shell guns to sea was Henri-Joseph Paixhans, a general of French artillery. The first large-shell guns from Paixhans' design, chambered howitzers firing a 62½-pound shell (thicker walled than bombs to penetrate before exploding) was tested in 1824 against a moored frigate with remarkable accuracy and incendiary effect.

Big-shell guns at sea

The new guns began to come into use afloat in the 1830s, a French squadron firing them in bombardment of Vera Cruz, Mexico. The U.S. Navy began installation, including 16 eight-inch-shell guns in the three-decker "Pennsylvania," along with 104 32-pounder solid-shot guns. The British made similar installations. There was good reason for navies to proceed cautiously. Powder charges had to be limited, so shell guns had shorter range, and shell was

less accurate than shot. Improvements in metallurgy, gun construction, and fire control, along with the manoeuvrability of steam warships, at last were beginning to lead to an important extension of range that the big gun had promised from the beginning.

In 1853 the dramatic destruction of a weaker Turkish squadron by a Russian fleet in the harbour of Sinope of Turkey's Black Sea coast attracted world attention and increased interest in shell guns. England, the United States, and others built big-steam frigates (as they were misleadingly called) with large-shell guns. Their great striking power and manoeuvrability under steam made them the capital ships of the day, superseding the ship of the line for a brief time before the ironclads took over.

Larger guns, increased powder charges, and greater tube pressures were made possible by the replacement of cast iron by built-up wrought-iron guns (later, cast steel and, eventually, forged steel). Hoops were shrunk on over the powder chamber and breech end of the tube to give the strength required for the greater pressures.

Armour. The use of larger guns with more penetrating power and explosive shells made armour imperative. Among early experiments were floating armoured batteries built for the Crimean War. These flat-bottomed vessels could carry large-shell guns close inshore. Heavy wrought-iron plates over a thick wooden backing gave them outstanding protection. In October 1855 in a hard-fought duel, French floating batteries silenced a Russian fort at Kinburn, at the mouth of the Dnepr (Dnieper) River and forced its surrender, suffering little damage themselves from Russian shellfire and solid shot.

Other developments followed swiftly. The British soon built the first iron-hull floating batteries. The French followed with "Gloire," the first seagoing armoured warship, protected throughout her entire length by a wrought-iron belt of 4.3 to 4.7 inch armour backed by 26 inches (66 centimetres) of wood. Displacing 5,617 tons, she mounted 36 large-shell guns and could steam at 13.5 knots (one knot equals 1.15 miles per hour or 1.85 kilometres per hour); a three-masted sailing rig supplemented the engines. "Gloire" was the first of a series of ironclads laid down by Napoleon III; 13 similar ships soon followed, then two-decker armoured rams.

Great Britian countered with "Warrior," the first iron-hull, seagoing, armoured man-of-war. Much larger than "Gloire," she displaced 9,210 tons, mounted 28 seven-inch-shell guns, had slightly lighter armour, carried sails, and was one knot faster. These first ironclads were commissioned on the eve of the American Civil War in which ironclads were destined to take a decisive part. The war itself produced several spectacular developments, including pioneer submarines (see SUBMARINE), the first aircraft carriers (to handle balloons for observation), and the torpedo boat, one of several means the Confederates explored in trying to break the blockade. These little craft had weak steam engines and mounted a torpedo lashed to a spar projecting from the bow. Called Davids, they frightened the blockaders and might have sunk a wooden warship; but the one successful attack struck the armoured "New Ironsides." The explosion caused considerable damage, requiring shipyard repairs, but did not lift the blockade. The David was a weak but definite forerunner of the torpedo boat and then the versatile destroyer.

The ironclads

Ironclad warships were crucial, perhaps decisive in the North's victory. Partial ironclads appeared early on the western rivers, and, under Flag Officer Foote, spearheaded Ulysses Grant's victories in 1862. River and coastal ironclads (ultimately, mostly monitors) dominated the war against the South in attacks from the sea and in decisive support of land operations from the Mississippi system to the Chesapeake Bay and James River. Most memorable of the combats was the duel between "Monitor" and "Virginia" ("Merrimack"). When the Federal forces lost Norfolk Naval Shipyard, Portsmouth, Virginia, April 1861, they burned several warships, including the heavy steam frigate "Merrimack." The Confederates raised "Merrimack," installed a ram and slanting casemates made from railroad track over thick wooden backing, as had been done in "Gloire," and renamed her "Virginia." Mounting ten guns, including four rifles, "Virginia," with yard workmen still on board finishing up, sailed on March 8, 1862, for her trial run. Defying concentrated fire of ship and shore batteries, she sank two of the wooden blockading fleet before retiring with the ebbing tide. In this dramatic moment Ericsson's "Monitor" arrived from New York during the middle of the night. Displacing fewer than 1,000 tons, less than a third of "Virginia," "Monitor" had a boxlike iron hull supporting an iron-plated wooden raft on which revolved the turret. The 172-foot-long vessel had little freeboard except for the thickly armoured rotating turret that mounted two 11-inch smoothbores. (Although new on warships, the revolving turret was an old idea; a turreted warship was built in England at about the same time.)

"Monitor" had many deficiencies. Not really a seagoing warship, she had nearly sunk on her voyage down from New York and did sink on her next sea voyage. Yet she proved the equal of her rival in their duel on March 9. The battle ended in a draw with neither ship seriously injured, but repercussions swept the world.

On April 4, scarcely more time than required for a ship to cross the Atlantic, Great Britain ordered the 131-gun ship of the line, "Royal Sovereign," to be cut down, armoured, and fitted with turrets. Only three and a half weeks later she laid down "Prince Albert," the Royal Navy's first iron-hull turret ship, mounting four turrets.

Coastal and river monitors

The Union Navy ordered 66 coastal and river monitors, low freeboard ships not suitable for high-seas action and most not for long voyages. Many were not completed in time for war service. The North also laid down four double-turreted Kalamazoo-class monitors to carry 15-inch guns that would have been seagoing battleships; these, however, were not completed by the war's end and were scrapped.

Besides "Virginia," the Confederates began a number of other ironclads from Richmond south on the east coast and on the inland rivers. Several of these rendered valuable service and probably lengthened the war. But most had to be destroyed before completion; David Glasgow Farragut's decisive fleet dash into New Orleans, Louisiana, forestalled completion of two powerful ironclads that might have stopped him. Out of a combination of characteristics of the "Monitor" and "Virginia" types evolved the battleship that was next to rule the sea.

Underwater. The South developed innovative subsurface weapons, including mines (called torpedoes) that were detonated by contact or electrically, and a submarine armed with a torpedo attached to a spar. (For a detailed description, see SUBMARINE.)

Warship development to the end of the century. The later 19th century continued to be a time of great flux in warship design. European nations tried numerous arrangements of guns and armour, such as centre-line turrets, a central armoured citadel with large guns on turntables at each corner, lightly armoured big guns topsides in barbettes (open-top breastworks), torpedoes in even the largest vessels, and substitution of high speed for armour.

For a time even the ancient ram was revived. When the Austrians won the Battle of Lissa from the Italians in 1866 by ramming, its value for the future seemed confirmed. Hence for years most large ships carried rams, which proved to be more dangerous to friend than foe when ships were sunk in peacetime collisions.

Ships. The trend toward the centre-line-turret, all-big-gun battleship finally became clear. In it were combined the seagoing hull, armour, and habitability of "Virginia," "Gloire," and "Warrior" with the turret and massive big guns of "Monitor."

HMS "Monarch," 8,300 tons, mounting four 12-inch guns in two turrets, commissioned in 1869, was perhaps the first true seagoing turret warship. HMS "Devastation," 9,330 tons, four 12-inch guns in two turrets, and massively armoured, completed four years later without sail, was a next step toward the ultimate 20th-century battleship, a ship with an armoured citadel around the propulsion plant, powder magazines, and handling rooms. Rising out

of it, protecting big guns and crews, were barbettes and turrets. The main battery shrank to a few powerful guns, but these took the place of many in broadside because of their great size and ability to fire through a wide arc of bearings. The change was vividly illustrated by the "new navy" the United States began building in the 1880s, consisting not of improved monitors but of powerful seagoing capital ships with mixed-calibre main batteries. In 1891 the United States laid down "Indiana," "Massachusetts," and "Oregon." Displacing 11,700 tons, these sister vessels had 18-inch belt armour, a speed of 15 knots, and mounted four 13-inch guns in two turrets. They also mounted eight eight-inch guns in four turrets, smaller guns for defense against torpedo boats, and six torpedo tubes. The plan was, as in other navies, to employ the heavy guns against enemy machinery and magazines while the faster firing eight-inch guns attacked his superstructure.

Armoured cruisers

The armoured cruiser was developed in this period as a large, fast vessel armed with intermediate-calibre guns and protected by armoured deck and medium-weight belt armour. Designed for commerce protection and raiding, as well as to cooperate with the battle line in fleet action, it was considered powerful enough and sufficiently protected to fight any ship capable of catching it and able to outrun battleships. Some even held it should become the principal warship.

Less heavily armoured was the protected cruiser, the engines and magazines of which were shielded by an armoured deck, but which lacked an armour belt. Unprotected cruisers had little or no armour, carried fairly light guns, and were designed primarily for scouting, patrolling, and raiding.

In this period there were fundamental advances in underwater weapons. An Austrian naval captain and a Scottish engineer produced an underwater projectile driven by compressed air to strike the unprotected hull below the waterline. At first slow, inaccurate, and of short range, it was progressively improved. Advances included stabilizing devices, more powerful warheads (initially of dynamite), and, very importantly, alcohol-burning steam-turbine power instead of compressed air.

Carrying this new self-propelled weapon, the torpedo boat had great potential, particularly under conditions of low visibility. Small, unseaworthy, and useful only in restricted waters with the then short range, slow torpedoes, the new boats did not immediately live up to expectations; nevertheless, as craft and torpedo improved, they were soon regarded as a major menace. Torpedo-defense batteries of rapid-firing small guns joined the main batteries of larger warships. Heavy ships added searchlights and torpedo nets. The threat also brought a new type of warship. The torpedo-boat destroyer first appeared in Great Britain in 1893. It was essentially a larger torpedo boat, mounting quick-firing guns, as well as torpedoes, of high speed, and capable of operating on the open sea. It was intended to accompany a battle line at sea, screen it from torpedo boats, and attack the enemy line with torpedoes. HMS "Havock," of 240 tons, was the first of the new type.

General improvements in technology. Engines for all of the types of warships steadily improved. Stronger metals made possible higher steam pressures and weight reduction. With higher pressures the two-stage compound engine added a third and even a fourth stage; then as the century ended, the steam turbine appeared to provide a marked increase in the speed of ships.

Development of turbines

Early armour had been of wrought iron backed by wood. To increase resistance against ever more powerful rifled guns, compound armour of steel backed with iron was devised to combine steel's surface hardness with iron's resiliency. The firm Schneider & Cie in France invented an oil-tempering process to produce a homogeneous steel plate that had good resiliency and greater resistance than compound armour. The later addition of nickel further improved its resistance.

Steel-armour-piercing shells came into use in the late 1880s, again threatening the armoured ship. An American engineer, Hayward Augustus Harvey, perfected a face-hardening process, applying carbon to the face of the steel plate at very high temperatures for an extended period and tempering. Harvey nickel-steel armour superseded earlier types. Then, in 1894, the Krupp firm of Germany devised hot-gas tempering, based on Harvey's process, that in turn became standard with world navies. Later, the addition of chromium to nickel steel was found to be a further improvement.

The impact of developments in guns and powder exceeded even that of warship design in their effect upon navies. In the two decades after the American Civil War the main difficulties with breech mechanisms were resolved. Better guns, along with breech-loading, made possible both longer ranges and higher rates of fire.

New powders were equally important. About 1880 brown or cocoa powder appeared, employing incompletely charred wood. It burned slower than black powder, hence with a given maximum pressure the sustained pressure was much greater. To take advantage of this for longer-range firing, large gun calibres jumped to 30-35 times bore diameter.

Several nations began to achieve success with smokeless powder of nitrated cellulose and usually some nitroglycerin. Calibres of guns again advanced. With greater striking power available, armour-piercing projectiles became more formidable. These were originally solid shot designed simply to punch through armour plate. In the 1890s, better steel and fuses made it possible to add an explosive charge. The resulting semi-armour-piercing shells became highly destructive, and in time all of the armour-piercing projectiles carried explosive charges.

In 1881 the British Admiralty advertised for an anti-torpedo-boat gun to fire cased ammunition at a rate of 12 shots per minute. Benjamin Berkeley Hotchkiss, an American ordnance engineer with a factory in Paris, produced a series of one-, three-, and six-pounder rapid-fire guns that eventually went into nearly all of the navies, though he got his first order from the United States Navy. The rate of fire vastly increased for small guns. For larger ones, loading times also improved; in 1884 the average time to load a 12-inch gun was about four minutes. By World War I it had shrunk to about 20 seconds.

Increases in rate of fire

Other technical developments in these decades that had great impact on naval craft were the appearance of the internal-combustion engine and the electric motor, which at last made possible an effective submarine. Experiments in wireless telegraphy foreshadowed control of the operation of ships out of sight of the fleet commander.

Naval vessels of the World Wars

WORLD WAR I

The dreadnought battleship. By the 1890s the capital ship had evolved into what is known historically as pre-Dreadnought, characterized by its mixed main battery. Studies reinforced by battle experience in the Spanish-American (1898) and Russo-Japanese (1904–05) wars indicated that fire from large guns at longer ranges was more effective than mixed-battery fire closer in. Only big-gun shells had sufficient impact to do serious damage to well-armoured ships. Improved gunsights, new spotting techniques, and developing range finders made long ranges practicable.

Both the United States and Great Britain designed all-big-gun, centre-line-turret battleships, but the British "Dreadnought," completed in 1906, was the first to be commissioned. Equipped with ten 12-inch guns, turbine drive, 21-knot speed, she gave her name as a generic term to a new type. Other navies promptly sought to match and excel her.

Battle cruisers, essentially a further development of the armoured cruiser but with battleship guns, were designed to run down and destroy hostile cruisers and to protect friendly shipping. The battle cruiser was several knots faster than contemporary battleships, speed being achieved by reducing armour. Germany's battle cruisers had slightly less speed but better protection than their British equivalents, both in armour and compartmentation. German crews were intensively trained in damage control.

As the naval race speeded up, progressively more powerful ships appeared. All of the navies quickly adopted turbine propulsion. The United States went to oil fuel even for its capital ships. With oil supply more of a problem, European nations nevertheless followed suit because of oil's superior efficiency. Newer battleships that could hit more and more accurately at ever increasing ranges rapidly made older ships obsolete. Guns increased in size and striking power with intensive research going into propellants, shell design, and detonating explosives. Improvements in range-finding and fire-control equipment rapidly extended the effective battle range. Director firing (control of several or all of a battery from a single station) made for better coordination of fire, as well as more precise spotting and salvo shooting. World War I speeded up these developments, especially in range keepers and other fire-control equipment that made possible rapid and accurate calculation of target course and speed. The complex elements of the fire-control problem were so well-handled that early hitting became possible, sometimes on the first salvo—no mean feat when the target is many miles away racing through the sea.

This extension in range was a factor in the adoption of larger guns. Because accuracy, range, and the destructive power of guns all increase with calibre, given the same initial velocity and accuracy of fire control, it is evident that as soon as range-finding and spotting can determine the fall of shot beyond the effective range of existing guns, a larger calibre is demanded. "New York" and "Texas" of the United States Navy, commissioned in early 1914, mounted 14-inch guns. Great Britain's Queen Elizabeth class, joining the fleet in 1915–16, mounted 15-inch guns and sped through the water at 25 knots, virtually outmoding the battle cruiser. The United States' Colorado class mounted 16-inch batteries.

Steady increase in gun calibre

The battleship saw little combat in World War I as compared to the heavy service of its ship-of-the line predecessor and the role it was destined to play in World War II. Yet World War I was in many ways the zenith of the battleship's influence. Despite submarines, aircraft, and destroyers, the outcome of the war still hinged upon control of the sea by the battleship. Had superiority in battleships passed to Germany, Great Britain would have been lost, and the Allies would have lost the war. The one moment when this might have happened was the one large-scale clash of battleships, the Battle of Jutland. Fought in May 1916 in mist, fog, and darkness, the combat revealed the strengths and weaknesses of the battleships and battle cruisers; superb seamanship on both sides demonstrated that miles-long battle lines lost in the mist could be effectively controlled. Three British battle cruisers were lost. Several German battleships, thanks to watertight subdivision and efficient damage-control systems, survived despite more hits. But the British advantage in numbers of battleships was decisive. Germany turned to the submarine as a means of countering the Allied blockade.

Cruisers and destroyers. On the eve of World War I, Britain had 145 cruisers in commission or building, almost as many as the other major naval powers combined; Germany had 57, the United States 32. Cruisers of the first part of the 20th century did not show the phenomenal changes in size and power of the battleship. They benefited from the same advances in propulsion, gunnery, and materials that made all of the warships more effective. Cruisers had only a modest speed advantage over battleships until the Omaha class of the United States Navy, designed late in World War I and built afterward. Small tube boilers and superheated steam gave them a speed of 35 knots, about a 50 percent increase over other United States cruisers.

Destroyers experienced radical development. HMS "Havock" that started the type in 1893 soon had an outstanding successor. HMS "Viper," powered by improved turbines, on trials in 1899, achieved the then unprecedented speed of 34 knots. The destroyer grew in size, power, habitability, and functions into World War I, when the submarine gave it enormous importance. In August 1914 Britain had nearly 300 destroyers, Germany about half as many, the United States about 50. When Germany adopted unrestricted submarine warfare in February 1917, Great Britain felt an acute shortage of destroyers. Shipping losses soon forced diversion of destroyers from fleet duty to convoy protection and antisubmarine warfare.

When the United States entered the conflict it brought 67 destroyers in commission, with a few on the ways, and began a massive naval shipbuilding effort primarily in destroyers and smaller antisubmarine vessels and the host of merchant ships needed to sustain the vast overseas logistic flow into France and Britain. Most of the 247 United States destroyers laid down were completed after the Armistice; but these 35-knot ships of 1,200 tons, or more, a generation later played an important role in World War II. The U.S. and other Allied destroyers of 1917–18, equipped with hydrophones and depth charges, as well as guns and torpedoes, overcame the submarine threat and had a large share in the safe convoy of 2,000,000 American troops to Europe without loss of a single soldier.

Submarines. The last half of the 19th century brought the many technical advances that made possible undersea craft. Paradoxically, most of the development was carried out in France, Great Britain, and the U.S., though the craft were to be used primarily by these nations' adversary, Germany, in World War I.

Germany started building undersea craft later than other countries. In 1914 it had 38 in commission when war broke out—fewer than Britain, France, Russia, or the United States, counting those nearing completion. Despite the number in existence, since the submarine was unproven in war, its true potential was not generally appreciated.

The World War I submarine

The submarine of World War I (and indeed of World War II) was still essentially a surface ship that could dive for brief periods. It could maintain moderate submerged speed for only a few hours until its batteries became exhausted, forcing it to surface to recharge with the engines. At night, and at other times whenever possible, the submarine cruised on the surface, rigged for quick diving, depending on its low profile to escape detection before it sighted the foe—an advantage that the airplane negated. Submarines would often make the first part of approach on a target surfaced at high speed to get in position to attack. Thus systematic ship and aircraft patrol severely restricted the U-boat's capabilities in the patrolled area. Of the 372 submarines the Germans had or commissioned during the war, over half were lost (see SUBMARINE).

Mines. Like most modern naval weapons, mines have long ancestry, including British and United States efforts in the 17th and 18th centuries. In the 19th century, the Confederates employed mines (called torpedoes) with considerable success in the American Civil War. Their versions included contact mines and percussion or electrically detonated mines fired from shore when the ship was in the right position. This controlled type sank USS "Cairo" on the Yazoo River in 1862, the first destruction of a warship by a mine in combat.

When Robert Whitehead's self-propelled underwater device appeared, it took the name torpedo whereas the term mine reverted to the underwater weapon that lies and waits (or drifts with the current) for its target.

During the latter part of the 19th century the Herzhorn, an improved method of firing contact mines that began to make this one of the most effective weapons in war, appeared. The horn, or tube, protruding from the mine contains a glass tube of acid. When struck by a ship, the glass breaks and releases the acid to form the electrolyte of a small battery that generates enough of an electrical charge to fire a detonator, which in turn sets off the mine.

Early moored mines were laid with a preselected length of mooring cable to suspend them at proper depth below the surface. Later mines of this type were fitted with automatic devices that paid out the cable from a reel in the anchor (or sinker) to position the mine.

Mines were used in the Spanish-American War without result, but in the Russo-Japanese War both sides em-

ployed them effectively, destroying three battleships, five cruisers, and eight other warships. During World War I the belligerents laid nearly 250,000 mines. Submarines joined surface ships as minelayers. During the last year of the war, mines began to be employed in large numbers against submarines. Allied minefields sank 44 U-boats, more than were sunk by any other means.

Naval aviation. The airplane had just begun to go to sea on the eve of World War I. In Chesapeake Bay in November 1910, the scout cruiser USS "Birmingham" launched the first airplane ever to take off from a ship. Two months later in San Francisco Bay, January 18, 1911, a plane was landed on an improvised flight deck built over a turret of the armoured cruiser USS "Pennsylvania." The next year the United States Navy bought its first plane, as did the British. Within a few years, larger warships were able to launch planes underway—thus navies were acquiring far-seeing "eyes of the fleet" greatly surpassing the scout cruiser and its frigate ancestor.

The first carriers

In 1914 the British Admiralty converted Channel steamers to carry seaplanes and converted a tanker hull for use as a seaplane carrier for ten planes. "Ark Royal" served in the Dardanelles campaign and in the North Sea in 1918, flying her planes against Zeppelin bases.

The British modified a number of other ships to take planes, including fighters with wheel landing gear that could be launched but could not land on board. A passenger liner was completed as HMS "Argus," the world's first true flush-deck carrier that could launch and land planes on a flight deck extending from bow to stern. But although scores of aircraft went to sea in warships during the war, most naval planes of this period, both sea and land types, operated from shore bases, principally against submarines. Nevertheless, navies put a substantial effort into pioneering air warfare. The United States Navy developed a flying bomb, a pilotless, remote-controlled airplane, an ancestor of the guided missile, and giant seaplanes for distant oyerwater cruising. In 1919 one of them, NC-4, completed the world's first crossing of the Atlantic by plane, arriving in Lisbon via the Azores on May 27.

Naval planes employed bombs against ships and shore bases. The Royal Navy also added torpedoes starting early in the war, and, by 1915, had seaplanes specially built for torpedo attack. In August of that year, operating from a seaplane tender, they sank two Turkish auxiliaries in the Dardanelles.

To counter the airplane, significant changes were made in warship armament. Early in the century, navies had standardized on two gun sizes for battleships—the centreline-turret main battery providing concentrated massive fire on most bearings and a secondary broadside battery of smaller calibre guns with the typical naval characteristics of high initial velocity, low trajectory, and short time of flight to combat high-speed targets—*i.e.*, destroyers and torpedo boats. The appearance of aircraft, smaller and even faster, demanded rapid-fire, high-angle guns and new fire control. Capital ships again used many types of guns on board with a third battery in U.S. battleships of 3″/50 calibre anti-aircraft guns. These and 0.30-calibre machine guns were a beginning of intensive developments in air defense.

WORLD WAR II

Naval limitations treaties. The end of World War I found Great Britain, France, and Japan building or planning large, new capital ships and the United States continued a major construction program that would have given it the world's leading navy. To end what was turning into an arms race among former allies, the Washington naval treaty of 1922, and later London naval treaty of 1930, prescribed limits for navies of signatory powers (5:5:3 for the United States, Great Britain, and Japan). Battleships could not be armed with guns larger than 16 inches (Britain had designed the battle cruiser "Furious" with 18-inch guns during the war, and a race for new super guns was feared), could not exceed 35,000 tons standard displacement, and could only be replaced when 20 years old. Any man-of-war over 10,000 tons was designated a capital ship, thus fixing this as the top limit for cruisers.

Arms limitations

Change in warships continued to accelerate, affecting every field of naval warfare, including aircraft and missiles, communications, the revolutionary introduction of radar, anti-aircraft gunnery, propulsion plants and other elements of surface ships, submarines, torpedoes, and mines.

Battleships. Battleship construction, halted for several years, began again in the late 1930s as the disarmament treaties expired. All of the major navies laid down battleships of superior size and quality to the earlier generation. The largest were two Japanese vessels, "Musashi" and "Yamato." Displacing 70,000 tons, mounting 18.1-inch guns, with armour capable of withstanding projectiles of equal calibre, these two giants were never surpassed in battleship construction. Four turbine engines gave each a speed of 27 knots, with fuel capacity for a cruising radius of 7,200 miles (11,600 kilometres) at 16 knots. Improvement in engineering plants, including higher steam pressures, added markedly to the speed and radius of the new capital ships; the United States Iowa class had a speed of 33 knots. All of the new battleships bristled with anti-aircraft guns.

In World War II, the new U.S. battleships served primarily with fast, carrier, task forces, protecting carriers from surface attack and providing anti-aircraft protection. They were also frequently used for bombardment of shore targets and occasionally fought surface action in traditional manner, as at Guadalcanal in 1942. Late in the war, United States battleships bombarded the coast of Japan. The newest ships survived hits by torpedoes, gunfire, and Japanese suicide planes (kamikaze) and continued to operate until they could be sent to port for repair.

Great Britain, Germany, and Japan lost new battleships to air attack, in which torpedoes from carrier-based aircraft were the primary weapon, or to a combination of air and submarine or, in the case of "Bismarck," to surface-ship gunfire and torpedoes. Had any of these battleships been given reasonable air protection, it would have survived. Even old United States battleships went through campaigns from Tarawa to North Africa, Normandy, and Okinawa without loss, though a number were hit.

Importance of air protection

At Pearl Harbor (December 7, 1941) several old battleships suffered fearfully in the peacetime surprise attack. This was inevitable for any ship not in a condition of maximum gun and watertight-door-closure readiness, employed at times of possible enemy contact and relaxed at other times for the sake of rest and habitability. Yet all but two were repaired—some quickly—and served valuably in amphibious assaults in both oceans.

Naval aviation. After World War I several nations built aircraft carriers, a new type of capital ship that Great Britain had pioneered during the war by converting warships and merchantmen. HMS "Hermes," laid down January 1918, was the first ship begun as an aircraft carrier. After its launch in 1919 the Japanese laid down the second carrier, "Hosho," and pushed it to completion ahead of "Hermes." The fundamental techniques of takeoff and landing were worked out on the "Hosho" and on British and United States ships in the early 1920s. A characteristic feature, the "island," or bridge for navigation and command, was added, though the Japanese continued to operate a few flush-deck carriers for several years. Following the limitations set by the Washington naval treaty of 1922, Japan, the United States, Great Britain, and France all converted capital ships then on the stocks into large carriers, most of which saw service in World War II.

Seaplane tenders were employed for patrol-plane operations. Battleships and cruisers were outfitted with aircraft for scouting and spotting. Techniques for these functions as well as torpedo-attack, dive-bombing, and fighter operations were refined into doctrine.

Among carrier planes in various navies, the United States SBD was probably the best dive bomber flying at the time of Pearl Harbor. The United States Navy's F4F was a rugged and dependable fighter. The lighter, faster Japanese "Zero" could outmanoeuvre it; but, in combat,

American pilots had a much higher rate of success, in considerable part because of armour, self-sealing fuel tanks, and heavier gunfire. Improved types of all of the planes were developed during the war.

Dominant role of attack aircraft carriers

Carriers played a dominant role in every aspect of the war at sea. The Pacific conflict began with the Japanese carrier strike against Pearl Harbor and ended with American and British carriers operating with impunity against the Japanese homeland. In between, the Battle of the Coral Sea, May 1942, was the first in history in which opposing fleets fought without ever coming in sight of each other. A month later off Midway atoll, carriers again played the decisive role.

The Battle of Midway reinforced a conviction already clear, especially from British operations in the Mediterranean with and without air support, that control of the sea now also meant control of the air over the sea. For the rest of the war this axiom ruled with increasing force. In the autumn of 1942 the Solomon Islands campaign underlined the importance of both aircraft and submarines in fleet operations, emphasized that modern sea power is a trident of air, surface, and undersea forces.

The powerful United States task forces of the war's later stages included three to four carriers, six to seven battleships and cruisers, and 13 to 14 destroyers. Improved fighters, guided by the carriers' combat-information centres, radar plots, and fighter direction, provided the basic air defense. Supported by aircraft, destroyers gave antisubmarine protection and set up an outer screen of anti-aircraft fire. Battleships and cruisers ensured protection from surface foes and massive anti-aircraft support closer in to the carriers. The introduction of radar fire control and the proximity fuze, which exploded the projectile in the vicinity of the target, for the 5-inch anti-aircraft batteries, made it almost impossible for enemy planes to penetrate the defense.

Use of radar

The phenomenon of radar had been observed in the 1920s. By the late 1930s, prototype sets with huge antennas were operating. A plane or ship as well as land reflected the transmitted beams, bouncing an echo back to the antenna on the ship's mast and thence to the receiving-set screen. Reaching ranges up to 200 miles, radar became invaluable for observation and control of a carrier's own planes and for detection and tracking of hostile planes. It also became a principal factor in antisubmarine warfare (especially for aircraft), for surface-ship formation-keeping and manoeuvring, and for navigation. Radar added markedly to the effectiveness of gunfire, especially anti-aircraft because at any range it could give distance measurements accurate to within a few yards.

Navies needed carriers for duties that did not require the high speed and complex fighting capability of the first-line, capital-ship attack carrier. A simpler type sufficed for the demands of convoy escort, antisubmarine duties, and amphibious assault. Conversions from merchant-ship hulls initially provided the ships that carried out these duties, thus continuing the time-honoured practice of using merchantmen as warships. Later these escort carriers, as they came to be called, were laid down as such. Most ranged around 8,000 tons, though some displaced over 11,000 tons, similar to the light attack carriers. They had a speed of 18 or 19 knots, carried 28–30 aircraft, and mounted one or two 5-inch guns, plus numerous smaller anti-aircraft batteries.

Seaplane tenders also served with the fleets. The largest could transport seaplanes. Any cove or inlet not controlled by the enemy could serve as an advanced seaplane base immediately upon the arrival of the tender. Seaplanes carried out valuable search, reconnaissance, patrol, and surveillance duties. Naval land-based patrol planes served similar functions; as their ranges increased they began to supplant the slower seaplane but did not do so completely until long after World War II.

Submarine warfare. Most major nations built submarines between the wars. The Soviet Union commissioned a very large fleet. Germany began the war with only 57, of which 49 were available for service, whereas the Soviet Union had 235. Germany was also very slow in building up production, losing more submarines than it commissioned in the first 15 months of the war—thus giving Britain time to develop defenses. The new submarines were larger, faster, could dive deeper, and had longer endurance than their World War I predecessors, and carried more powerful torpedoes. The fall of the Low Countries and France in May 1940, providing long, open coastline for submarine and air bases close to British shipping routes, enabled the handful of U-boats to become very effective, but although Germany belatedly put tremendous effort into building submarines, massive anti-submarine measures defeated them.

Improved anti-submarine measures

In addition to large numbers of escort vessels, there were improved weapons. Sonar gear became more efficient. New explosives increased the destructive effect of depth charges. The British produced the hedgehog, a multiple, mortar-like device, mounted forward, that threw a pattern of 24 contact depth charges ahead of the ship while sonar operators could still detect the target on their equipment. This eliminated the blank period unavoidable in a depth-charge attack astern. Patrol aircraft and radar also contributed; the range of patrol planes steadily expanded to permit covering more and more of the convoy ocean routes where submarines might surface to recharge batteries. The extension of range together with the installation of radar on ships and planes made possible night detection and often surprise of the surfaced submarine.

The escort carriers provided air search and attack close to the convoy, forcing U-boats to remain submerged and making it harder for them to get into position for attack. Ultimately, when patrol planes gained endurance to accompany a convoy over much of its route, the escort carriers operated independently in areas of known submarine concentration, making all of the parts of the ocean near shipping routes unsafe for submarines.

Late in the war, when surfacing even at night to recharge batteries had become very hazardous, the Germans adopted the snorkel, a Dutch invention. This air-intake-and-exhaust funnel permitted running the engines while submerged, to recharge batteries without surfacing, and presented a very small radar target.

Both the United States and Japanese navies made extensive use of submarines in attack roles in the Pacific, and, as the war progressed, United States submarines played a key role in the air-naval blockade of the Japanese home islands. They sank many warships and were responsible for 55 percent of Japan's critical merchant ship losses. There were few important innovations; the most dramatic, perhaps, being the Japanese employment of small, two-man submarines, launched from a mother ship, designed to penetrate enemy harbours. These achieved little success, however.

Mines. Both Great Britain and Germany devised magnetic mines in World War I. Germany continued development, and early in the Battle of Britain small minecraft, submarines, and airplanes laid magnetic bottom mines, detonated by changes in the magnetic field caused by the passage of the iron hull of a ship. The mines caused serious loss and dislocation until one was recovered and suitable minesweeping gear devised. A supplementary countermeasure was degaussing, passing an electrical current through a cable strung horizontally around the ship, thus setting up a magnetic field to counter the mine or the magnetic torpedo-firing mechanism.

In the following years, navies brought out a wide variety of firing mechanisms for mines that included a combination of various influence types, such as the acoustic and, near the end, the almost unsweepable pressure type, actuated by a change in pressure caused by a ship's passage.

Surface vessels, aircraft, and submarines on both sides laid mines in every theatre. British air-laid mines caused many losses and severely hampered Axis operations from the Baltic Sea to the Mediterranean. The Allies lost numerous warships and 1.4 million tons of merchant shipping to mines.

In the Pacific, United States carrier planes, submarines, and surface vessels laid mines, but the largest effort was by the United States Air Force, laying naval mines. In the spring and summer of 1945, B-29s, operating from the

Mariana Islands, laid over 12,000 influence-type mines in Japan's home waters, sinking about half-million tons of Japan's merchant marine.

Cruisers and destroyers. Powerful new warships of these types appeared in all of the major navies in the 1930s. Two cruiser classifications came into common use: "heavy," armed with 8-inch guns, and "light," with 6-inch guns. Both mounted 5-inch double purpose batteries, and in the war added 40 millimetre and 20 millimetre. The United States also built anti-aircraft cruisers with 5-inch guns as the main battery. Several other nations also built anti-aircraft cruisers. Cased ammunition and automatic loading systems increased the rate of fire of all of the naval anti-aircraft. Radar, developed for gun directors, greatly improved accuracy, making possible automatic tracking of targets and precise firing in fog and darkness. Another anti-aircraft improvement was the influence or proximity fuze, a tiny electronic device that detonated the projectile at a predetermined distance from the target to produce maximum damage.

New destroyer types

Because the disarmament treaties did not limit them, most nations built a series of new destroyers of increasing capabilities. In 1939 the British Navy had over 100 of modern construction. These and older destroyers satisfied only a beginning of Great Britain's needs, despite the small number of German submarines. Beginning in 1941 the British produced convoy-escort ships called sloops and corvettes, smaller, slower, and less complex than destroyers. The United States commissioned over 400 fast destroyers between 1939 and 1945, and, starting in 1942, turned out a stream of nearly twice that many smaller and slower destroyer escorts that were effective both for convoy escort and in hunter–killer task groups.

The Japanese also improved their destroyers, notably in the areas of night optics and torpedoes. They inflicted severe losses on the United States and Allies in cruisers and to the United States in destroyers. The large United States building program soon changed the balance in destroyers as in other ships.

Service force and logistics. Logistic ships, indispensable for World War II fleet operations, included tenders with spare parts and repair shops to act as mother ships for destroyers, submarines, tankers, supply ships, ammunition ships, netlayer-tenders, and various other types. Minesweepers were indispensable, both for attack and defense.

In operational handling of service and logistics ships, the United States Navy pioneered two developments widely copied since: first, the establishment of floating advanced bases, capable of taking care of the whole fleet's needs and including floating dry docks, floating cranes, repair barges and ships, barges for degaussing (demagnetizing), and storage and issue of ammunition and supplies; and second, underway replenishment, which made it possible for task forces to remain at sea for months at a time without returning to port. Fuel for ships and planes far outweighs in tonnage any other logistic requirement of warships, followed, in combat operations, by munitions, food, and other supplies.

Amphibians. Although the failure of the Dardanelles campaign in World War I had somewhat discredited amphibious operations, both the United States and Japanese navies practiced such operations between the two wars. The Japanese opened their offensive in 1941–42 with successful amphibious assaults against Wake Island, the Philippines, Indonesia, and several other points. After Midway, the United States launched its first amphibious assault, against Guadalcanal, in August 1942, an operation that involved several new types of landing craft as well as all classes of ship, from battleship to converted merchantman. A long series of amphibious assaults followed in both Pacific and European waters, with increasing refinements of landing and support craft. The final amphibious operation of the war, against Okinawa in 1945, involved about 1,500 vessels. An advance force of battleships, destroyers, minesweepers, landing craft, and other auxiliaries entered the area a week before the main assault, sweeping mines, clearing obstructions, and establishing a repair–logistic anchorage in a nearby island group, simultaneously carrying out an intensive bombardment of the landing area. In the assault, large merchantmen types, converted or built for assault, participated. They carried numbers of landing craft with bow ramps, some for troops, some for vehicles and troops. Some landing craft were armed with guns, mortars, or rockets. Amphibious tractors and trucks also waded ashore. The highly complex operation, with its great variety of ships and craft effectively coordinated, significantly illustrated the growing ability of sea-based power to outweigh that ashore.

Amphibious assaults

Warships in the nuclear age and the future

The generation since World War II has seen the continuation of swift change in warships that has characterized the past century and a half. Among developments that may be termed epochal are nuclear propulsion, ballistic missiles with atomic warheads, guided missiles, and space satellites for naval operations. Many other advances have taken place; carrier jet planes fly at two to three times the speed of sound, and submarines extend their operations to thousands of feet below the surface.

SUBMARINES

Nuclear power in submarines. The first government support of atomic energy in the United States was the assignment of small research funds by the navy in 1939. Interested because of the promise for propulsion, the navy thereafter made important progress toward solving the problem of controlled nuclear fission until later when all of the effort was concentrated toward producing the atomic bomb.

After the war, the Navy was able to get funds for research on nuclear propulsion. Capt. Hyman George Rickover was assigned to head the project. In less than a decade a revolution, perhaps even greater than that from sail to steam, was launched with the USS "Nautilus" (January 21, 1954).

Steam had freed ships from the vagaries of wind and tide but at the same time had imposed the limitations of large fuel consumption at high speed and the requirement for frequent refueling. Overnight, atomic energy freed ships from both limitations. Nuclear power plants can drive ships at maximum speed virtually continuously.

Advantages of nuclear power

Another advantage of nuclear power has special meaning to submarine operations. Until its advent, submarines were essentially surface ships that could submerge for relatively short periods. Atomic energy, which does not require oxygen from the air for combustion, at last made possible the true submarine. Hence submarines could be designed with hulls shaped for high underwater speed, rather than on the surface-ship lines that had hitherto prevailed. Naval constructors developed and tested a streamlined hull that provided maximum speed, efficiency, and manoeuvrability submerged. The resulting experimental "Albacore" marked a notable step forward in submarine design. Highly manoeuvrable in any direction, almost like an airplane, and swift under water, even with battery propulsion, her performance gave promise of the submarine of the future that would result from joining this hull design to atomic power.

The nuclear power plant in "Nautilus" surpassed the hopes of the most optimistic. In her first two years, this revolutionary submarine steamed more than 62,000 miles, setting new speed and submerged-endurance records. Her reactor cores were replaced early, after two years. Cores in the early 1970s had a life of a decade or more and drove the boat submerged at speeds considerably greater than the fastest submarine surface speed in World War II. In 1958, "Nautilus" voyaged over the top of the world, under the polar ice cap. In 1960, one of her successors cruised submerged on a 36,000-mile trek around the world.

The reactor of United States submarines lies to the rear of the control room. Water shielding and a thick layer of lead and polyethylene protect the crew from radiation. The reactor is on the lower level of the reactor compartment and, when in operation, is sealed off from the upper level where the crew functions. Many other safety devices

are part of the system. Pressurized water both cools the reactor and carries away the fission-produced heat through the primary system (or loop) to a heat exchanger. The water under pressure does not become steam, but, as it passes through the steam-generating tubes of the heat exchanger, it transmits the high heat to a secondary loop or steam-producing system; the pressurized water then circulates back through the reactor in a constant cycle. In the secondary loop (which is sealed off from the radioactive primary coolant water) the water boils and expands into steam and thence to the turbogenerators to propel the submarine. This steam then goes to the condensers that convert it to water to repeat the cycle.

New submarine designs

In 1959 the United States commissioned a radical new type of fleet submarine that combined nuclear power with a streamlined underwater hull design that had no protuberances except a slender "sail" (containing bridge, periscope tubes, etc.) with diving planes mounted on it. It was the world's fastest and most manoeuvrable submarine, vastly increasing the submarine's effectiveness and at the same time the difficulties of antisubmarine warfare. The United States next brought out attack submarines, boats designed primarily for antisubmarine warfare. Able to dive to unprecedented depths, these are exceedingly swift and manoeuvrable. They excell in stealth, concealment, and surprise. The boats have advanced sonar gear, four torpedo tubes approximately amidships, and the unique Subroc (submarine rocket) attack system—*i.e.*, an antisubmarine missile fired from torpedo tubes; after leaving the tube, the missile's solid-propellant motor ignites; directed thrust puts it on course; it then emerges from the water and flies as a rocket; at a preselected point, the warhead separates and dives into the water and continues on course to its target. This "flying torpedo" has a range of probably more than 20 miles (32 kilometres).

The United States has concentrated first on atomic power plants for submarines, as did other nations. Despite espionage, the Soviet Union lagged at first, experiencing technical problems; once these had been overcome, it made rapid progress. In the 1960s the Soviet Union constructed large, modern, automated building yards to produce nuclear-powered submarines—one huge yard alone has a capacity to build nuclear boats that is greater than all of the yards in the United States combined. The highest national priority was given to producing ever new and better types, mounting increasingly sophisticated weapons systems. The shipyards can build at least 20 nuclear submarines a year.

In 1971 the United States had almost 100 nuclear submarines of all types. The Soviet Union had about the same number but was building at a rate several times that of the United States, which launched only three nuclear boats in 1970. The Soviets were certain to gain marked superiority in numbers by 1974, perhaps by 1973. Besides the rapidly growing nuclear fleet, the Soviet Navy continues to introduce new designs for diesel-powered boats, with particular emphasis on guided and ballistic missile types.

Ballistic and guided missile submarines. *The United States.* The United States Navy began to fit guided missiles to submarines in the 1940s. USS "Cusk" fired a historic first guided missile from her deck on February 12, 1947. Several other conversions followed her. In 1959 the United States launched the nuclear-powered "Halibut" as the world's first submarine designed from the keel up to fire guided missiles.

While working with guided missiles, the United States Navy began the development of an even more significant naval weapon, the ballistic missile, the production of which represented an outstanding technological achievement. The task force overcame a series of apparent impossibilities to create a new type of solid-propellant ballistic missile and mate it with the nuclear submarine. Within four years, on December 30, 1959, the complex and powerful fleet-ballistic-missile submarine USS "George Washington" was commissioned, constituting a weapons system generally said to be without parallel for power, accuracy, concealment, range, and effectiveness.

On July 20, 1960, the submerged "George Washington" launched her first Polaris missile; it hurtled 1,000 miles (1,850 kilometres) downrange, precisely to target.

Polaris and Poseidon guided missiles

Subsequent versions of the missile extended the range in steps to 2,500 nautical miles (one nautical mile equals 1.15 miles or 1.85 kilometres). In 1970 the more advanced Poseidon, with similar range but twice as powerful and even more accurate, was perfected. It can carry either a single, huge warhead or about a dozen independently targeted ones.

USS "James Madison," the first to be altered to take the 30-ton Poseidon, has a submerged displacement of 8,250 tons (before conversion). An inertial navigation system positions the submarine accurately beneath the surface at all times. The missile itself carries a compact, self-contained, inertial guidance that takes over as it bursts out of the sea and steers it to target.

All of these boats have two crews, to alternate on the continuous patrols. Maintaining speeds of over 25 knots whenever and as long as desired, with no need to surface for air, navigation, or communications, they return to the surface only on entering port.

The Soviet Union. The Soviet Union began to adopt the Polaris submarine concept with conversions and newly constructed diesel boats, mounting two to three launching tubes, firing relatively short-range missiles. The first nuclear, fleet-ballistic-missile type with three launching tubes appeared in the 1960s. Missile ranges, beginning with about 400 miles (650 kilometres), progressively increased. By the end of the decade the Soviet Navy began commissioning what has been termed, in the North Atlantic Treaty Organization (NATO), the Yankee-type submarine. It appears to be a copy of the Polaris submarine, though its missile range is apparently on the order of 1,500 miles. A longer range missile is scheduled for use in the early 1970s.

The Soviet Union has developed several classes of boats designed or converted to fire cruise missiles. These are low-flying, self-propelled guided missiles with homing heads, effective within visual or radar range or, with midflight guidance as by an airplane, able to hit targets perhaps 400 miles from the boat. The missiles are fast, have variable flight profiles, including low-altitude terminal flight, and can deliver either a conventional or nuclear homing warhead. They carry sophisticated electronics with excellent countermeasure resistance. In the early 1970s the Soviet Navy had 70 or more submarines equipped with these missiles; about half are nuclear-powered and are capable of firing while submerged.

The Soviet Union's large fleet of conventional diesel submarines can carry torpedoes or mines either with conventional or nuclear explosives.

Other countries. France and Great Britain have also developed nuclear ballistic submarines in small numbers, and Britain has three nuclear attack submarines in commission, with others under construction. Communist China is believed to have the third largest submarine fleet, all diesel powered, a few vessels mounting ballistic missiles.

NAVAL AVIATION

Attack carriers. After World War II, the rapid development of jet aircraft required longer and stronger carrier decks. The angled deck and the steam catapult were developed by the British and quickly adopted by the United States to handle increasingly larger and faster planes.

Of the new carriers, the most powerful was "Enterprise," which displaced 85,600 tons with an overall length of 1,101 feet (336 metres) and extreme width of flight deck of 252 feet (77 metres). She carried over 70 of the new, large planes and her powerful reactors thrust her through the water at speeds much above 30 knots. Her greater manoeuvrability and endurance at high speed, without delays for refueling (except for planes), gave her outstanding advantages over other carriers.

The "Enterprise"

The modern carrier task force roughly follows the pattern of World War II, with one to three carriers accompanied by cruiser and destroyer types as available. Ships

and equipment are far more sophisticated. Topsides bristle with electronic gear, radars, communication antennas, and missile guidance and countermeasure equipment. Anti-aircraft armament consists of guided missiles, supplemented by a few rapid-fire, medium-calibre guns. The numerous small-calibre mounts of World War II have disappeared; future air attack is anticipated as coming from long-range missiles.

Several navies today employ antisubmarine carriers. Most are World War II types built by Great Britain or the United States, though France and the United Kingdom completed a handful during and after the Korean War. In general, all of these carry a mixture of helicopters and fixed-wing planes. Both carriers and aircraft have extensive electronic gear. Weapons include depth charges and homing torpedoes.

An important supplement may be the helicopter mated to destroyer types. Great Britain, Canada, and the United States participated with other NATO allies in development and test of suitable helicopters and equipment to handle and house helicopters on the rapidly pitching and rolling decks of smaller ships. Canada, for example, has solved most of the problems met even in heavy seas.

Amphibious carriers. As successful helicopters evolved, the U.S. Marines modified amphibious assault techniques to include vertical envelopment, landing by air to the rear of hostile beach defenses. The United States Navy converted some World War II carriers into helicopter carriers for amphibious assault, and their success led to the construction of "Iwo Jima," commissioned in 1961 as the first ship designed and constructed for helicopters. She and six sister ships can each carry a marine battalion landing team, with guns, vehicles, and equipment, and a helicopter squadron to lift them to the beach. As vertical-and-short-takeoff-and-landing planes go into service, this type may also take on functions of the escort carriers. Displacing 18,000 tons, full load, this type transports about 2,000 troops and helicopters to take them to the beach.

Soviet helicopter carriers

Carriers of the Soviet Union. The first carrier built for the Soviet Navy was a helicopter carrier; in 1967 "Moskva" was commissioned, followed by a sister ship, "Leningrad." "Moskva" and "Leningrad" combine a carrier deck for helicopters (situated aft) with the guided missile armament of a cruiser forward. At 18,000 to 19,000 tons, full load, they are suitable for antisubmarine or amphibious operations. Their appearance coincided with intensive work in the Soviet Union on vertical-and-short-takeoff-and-landing planes. The Soviet Union now flies at least three types of short-takeoff fighters with maximum speeds ranging from 2 to 2.8 times the speed of sound. Experimental models of vertical-takeoff planes are also being tested, in subsonic versions that represent important progress toward supersonic successors. A carrier, probably to handle these, was under construction in the mid 1970s.

Naval aircraft. Navies continue to employ aircraft for search, reconnaissance, patrol (especially against submarines), early warning (especially against air attack), interceptors (fighters), attack against ship and shore, amphibious lift and tactical air control, electronic countermeasures, antisubmarine warfare, and other duties including training and transport. The newest reconnaissance planes combine several means of detection, including low-light-level TV, sonar, radar, infrared, and magnetic detection devices. Detection of submarines, solution of the fire-control problem, and release of the missile comprise an automatic computer process (see AIRCRAFT, MILITARY).

Since World War II the helicopter has progressed from a small, light aircraft with feeble lift into a powerful craft that has wide use in a broad spectrum of naval as well as other operations. The small helicopter, with more powerful lift, is of course still in greatest demand, especially for antisubmarine duties.

CRUISERS AND SMALLER GUN–MISSILE WARSHIPS

Cruisers. Construction of this class with guns as its main armament appears to have ended. Except for the nuclear-powered, missile-armed "Long Beach," the United States has only a handful of cruisers in commission, laid down in World War II and converted to carry guided missiles. In the 1950s the Soviet Union built 14 gun-armed light cruisers (19,000 tons displacement) but thereafter changed to smaller types, mounting antiship or anti-aircraft missiles for the main battery and also fitted with strong antisubmarine capability. The largest of these are the 6,000-ton Kynda class, carrying two, surface-to-surface, quadruple, Shaddock missile mounts and the newer 7,000-ton Krestas, with two twin Shaddock launchers and considerably increased missile capability. Navies in the 1970s build surface warships not so much to fight one another as to combat submarines, air-craft, and guided missiles.

First nuclear-powered surface warship

The United States Navy's "Long Beach" was the first surface warship in the world to steam under atomic energy and remains the only cruiser so propelled. She can race through the water indefinitely at speeds probably exceeding 35 knots. Her armament consists of guided missiles, two-5 inch guns, and antisubmarine weapons.

Superdestroyers. While gun cruisers pass from the scene, destroyers have grown larger until some are, in fact, small cruisers. The United States calls her largest types frigates, a term most other navies tend to reserve for smaller escort or antisubmarine types.

The United States Belknap class, laid down in the early 1960s and the latest United States Navy operational nonnuclear type, displaces nearly 8,000 tons full load. Armament includes five- and three-inch guns and a twin missile launcher for anti-aircraft and antisubmarine missiles. Development of the antisubmarine missile concept as in an improved United States version and in the Australian Ikara missile, provides for sending target information from the firing ship's sonar or helicopter sonar into the automatic fire-control data-computing system that controls the missile throughout its flight to keep it directed on the target regardless of manoeuvres.

The United States has also built the only destroyers with nuclear power. The first, "Bainbridge," of 8,500 tons displacement, was commissioned in 1962. "Truxtun" was commissioned in 1967 and in the early 1970s two more were under construction.

Destroyers. The large frigates that merge the destroyer type with the small cruiser merely continue a process underway from the first torpedo-boat destroyer. Tasks multiplied for this high-speed, versatile vessel, whose appearance in the fleet coincided with the birth and swift growth of submarines and aircraft, and combatting these hazards became principal duties. The phenomenal growth in effectiveness of these two foes through nuclear power for submarines and jet engines for aircraft has coincided with a similar increase in the destroyer's powers to combat them.

Instead of many guns, destroyers bristle with an array of sensors and are packed with electronic gear for swift detection, identification, and computation of firing data. Magazines hold supersonic, homing, guided missiles—as well as projectiles for the minimum number of guns that must remain in the armament. The complex warship carrying all of this must also have high speed, excellent sea-keeping ability, and long endurance. Lightweight materials such as fibre glass and aluminum and the use of solid-state devices and other miniaturization steps have compensated for only part of the additions. The result is that the size of destroyers had to increase.

The newest in most fleets now range in displacement up to 5,000 tons. Soviet destroyers built in the 1960s displace 4,000 to 4,600 tons, as do most of those of the United States. Destroyers of France and The Netherlands on the ways displace about 1,000 tons more. The United States has begun to build the Spruance class, displacing more than 7,000 tons, driven by gas-turbine engines and carrying two helicopters.

United States terminology, distinguishing destroyer leaders (frigates) from conventional destroyers, largely implies only more command functions. Otherwise, the two have essentially the same armament, speed, and employment in fleet-escort duties. The United Kingdom and many other nations call both large types destroyers and

Growth of destroyer escorts

assign frigate classification to the slower, usually smaller, and lighter gunned ocean escort, sometimes called the junior destroyer. Destroyer escorts have also grown and have acquired characteristics in many cases close to those of destroyers. In the 1950s a number of nations built escort types of under 2,000 tons displacement. Yet the rapid advance in submarine performance, especially under nuclear propulsion, has caused these types to experience marked growth. This has been necessary to accommodate helicopters, the more complex sensors (radar and sonar; hull-mounted and variable-depth), weapons, and associated equipment more complex than ever before. At the end of the 1960s and in the 1970s, many displaced 4,000 tons or more. Canada, notably successful in building destroyer escorts of just under 3,000 tons, particularly in helicopter operations, now builds in the 4,000-ton range to accommodate advanced weapon systems. Great Britain has under construction gas-turbine patrol frigates estimated to displace under 3,000 tons and to run at a top speed of 40 knots. If successful in handling the most effective antisubmarine weapon systems, this type could bring marked improvement to surface-ship antisubmarine operations.

Other types. Corvettes, fast patrol ships, and other small types serve in various fleets for coastal and limited-radius deep-sea operations. These vessels may range in size from France's new corvette types of about 1,300 tons, and the Soviet Union's 1,200-ton Mirkas, to Japan's 500-ton submarine chasers. Several nations have also built small, fast, torpedo, gun, or missile boats, most of which do not have antisubmarine capability. Great Britain has two 114-ton fast patrol boats with over 50 knots maximum speed. The Soviet Union has 125 or more 35 to 40 knot surface-to-surface-missile boats displacing 100 or 200 tons according to class. One of these on loan to Egypt sank the Israeli destroyer "Elath" with Styx missiles at a range of 13–14 miles in 1967.

The Asheville 240-ton class of gunboats was commissioned in the late 1960s, with diesel and gas-turbine propulsion that gives top speed of over 40 knots. The United States also has two small 60-ton hydrofoil gunboats and two larger experimental vessels (one of 310 tons) in the 40–50 knot speed range, which may be forerunners of fast escort ships of the future.

SEABORNE MISSILES

Significant missile developments in the last generation fall into six categories: submarine-to-ship or shore target; antisubmarine; air-to-air; air-to-ship or shore; ship anti-aircraft; and surface-to-surface (ship-against-ship or shore).

Submarine-to-ship or shore target. The Polaris–Poseidon ballistic missiles and similar types brought revolutionary new power to sea forces.

Submarines, of course, continue to fire the ancestor of the guided-missile family, the torpedo. Submarines in most navies now carry acoustic-homing missiles, with conventional or nuclear warheads and with influence detonation at the target. This type will be employed particularly by nuclear-propelled attack submarines in hunting down other boats. With special modification, submarines can fire cruise guided missiles against surface targets in visual or radar range of the firing boat or of a midflight guidance assistant.

Soviet cruise missile

The navy of the Soviet Union pressed ahead with employment of the cruise missile. In the early 1970s it had 70 to 75 nuclear and conventional submarines with this type of missile. Called Shaddock by NATO, the missile can deliver either a conventional or nuclear warhead and has variable flight profiles, including low-altitude terminal flight, which, along with good countermeasure equipment, makes it a difficult missile to combat.

Antisubmarine. The United States did develop a submarine air-flight guided missile against its primary naval rival, the large Soviet submarine fleet. Subroc, previously described, is unique among guided missiles in that it starts underwater, flies through the air as a rocket, then dives into the water to become again a submerged missile as its inertial-guidance and homing equipment direct it.

Numerous surface-ship and air-launched antisubmarine missiles exist. These range from improved versions of the depth-charge and contact-hedgehog type of World War II to air-dropped nuclear depth bombs and acoustic-homing torpedoes launched by plane, ship, or submarine. The United States, Australia, and France, respectively, have the Asroc (antisubmarine rocket), Ikara, and Malafon that, with solid-propellant rockets, hurl an acoustic-homing torpedo or depth charge several miles through the air at very high speed toward the target. Sonar and other data feed into the fire-control computer that positions the launcher correctly in train and elevation. In some systems now and eventually in all, the missile track is changed in flight by radio–radar tracking and guidance to defeat target manoeuvres. Hence it is certain to be near the hostile submarine when it enters the water (slowed by parachute) and begins to home acoustically. To prevent hazards to its own forces, homing life terminates after a set length of time.

Air-to-air missile. As in other missile areas, the principal developers of naval air-to-air missiles have been the United States, the United Kingdom, France, and the Soviet Union. The United States Navy Sidewinder is probably the simplest and least expensive of guided missiles. The early 1970s version was a winged five-inch rocket nearly ten feet long with infrared homing (or alternately semi-active radar in which the firing plane's radar illuminates the target and the missile's receiver homes on the echo). It ranged over 3,000 yards at more than two and a half times the speed of sound. The much more complex and expensive Sparrow is a 12-foot long, eight-inch diameter, semi-active, radar-homing rocket propelling a 60-pound warhead for about seven miles, at somewhat greater speed than Sidewinder's. The British HSD Red Top (with infrared homing) and French Mafra (infrared to semi-active radar homer) have similar performance. In the early 1970s the United States had in the final development stages the 35–50-mile, radar-guided, very fast Phoenix missile to arm a new fighter plane.

The U.S. Sidewinder

Air-to-surface missiles. Of the several nations producing missiles for this function, the United States has the widest range, and most of its effective ones are Navy developed. Typical are Walleye, an unpowered, 1,100-pound, glide bomb with TV homing on target; Shrike, ranging ten miles or more and homing on radar emissions of target; Bullpup (improved version) carrying nuclear or conventional warhead, 50 percent heavier than Walleye, of greater range than Shrike, and radio-controlled to target by the pilot of the launching plane, with visual flares to aid in guidance. The United States Navy also has a later radar-emission-homing missile, Standard Missile (adapted from the anti-aircraft version), perhaps double the range of Shrike, with a more powerful warhead and improved homing features. Soon going into service in carrier-attack aircraft is TV-guided Condor, with a stand-off range of over 35 miles. France's small, short-range, wire-guided missiles AS.11 and AS.12 have wide use in maritime planes of NATO and other nations. Martel, a medium-size missile of over 30-mile range (in antiradar and TV-guidance versions) has been developed jointly by France and Great Britain. West Germany's Kormoran, of similar size and range, has inertial guidance with terminal-radar or infrared homing.

The Soviet Union has emphasized the production of large, long-range missiles—which, of course, are the most effective for employment against important warships or any ship without accompanying air protection. The Kipper and Kitchen (NATO names) missiles are over 30 feet long, with turbojet power plants, and are probably radio-controlled. These Soviet-produced missiles undoubtedly have active or semi-active radar homers (or infrared, or both) and extremely powerful conventional or nuclear warheads.

Ship anti-aircraft and antimissile. The United States Navy produced an ancestor of guided missiles in a robot plane in World War I, later developed into the drone target plane for ship-anti-aircraft-firing practice. Radar gave instantaneous, precise measurements, thus making it possible to guide the drone to intercept a target. The

navy went ahead on this concept in World War II, producing television-guided drones. The next step toward a workable guided missile was a suitable rocket or projectile-like vehicle to fly the explosive to the target. By 1947 naval laboratories and research contracts had produced the basic elements of the nation's first true guided missile: first, the "flying stovepipe" or ramjet rocket power; second, radar-beam-riding guidance; third, homing equipment for the final, precision approach on the target; and fourth, influence detonation.

From these and other achievements evolved guided missiles for every aspect of naval warfare. In the early 1950s, the anti-aircraft missile, Terrier, with its complex launcher-handling system, easily developed from naval experience with gun turrets, became the lead guided missile to become operational. The current generation of Terrier, with dual boosters supplementing the solid-propellant sustainer, flies about 20 miles at two and a half times the speed of sound. It is effective against low- as well as high-flying aircraft. Cousins of the Terrier, Tartar and Talos, also still fly in improved versions. Because planes speed up and fire faster missiles at longer ranges, navies continue to develop missiles. The United States Standard Missile, a semi-active homing missile, has a 30-mile (48-kilometre) range. An interim, short-range, point-defense, antimissile system, the Sea Sparrow (an adaptation of air-to-air Sparrow), also operates in a number of United States warships. Another new missile, Aegis (Shield), began operations in the early 1970s with the particular concept of combatting cruise missiles (as well as the vehicle launching them), whether fired from plane, ship, or submarine. It detects targets under most difficult weather and countermeasure conditions with a scanning radar that can search through 360° almost instantly, along with very rapid computing and fire-control equipment.

Aegis designed to combat cruise missiles

France has a medium-range missile and a short-range one for point defense. Great Britain's point-defense short-range Seacat, contact and influence-fuzed and usually mounted in four-round launchers, has wide use in British-Commonwealth and other navies. An improved missile, Sea Wolf, that can counter small, fast missiles as well as the firing vessel or plane, is designed to replace it. The Seaslug, now in its second version, gives British destroyers excellent capability against aircraft at about the same range as Terrier. Test firings have shown high effectiveness at altitudes up to 50,000 feet. As in the United States, advances in technology and in potency of attacker have brought an improved medium-range missile, the Sea Dart. Developed for new destroyers to meet the missile threat of the coming decade, it has improved anti-aircraft performance at very high and very low levels and is also an antiship weapon.

The Soviet Union's surface-to-air missiles, which proved effective in hostilities in Vietnam and in Egypt, operate from warships under the name Goa, a two-stage, solid-propellant, guided missile. Two versions have been developed for use ashore, one to combat planes flying at normal to high altitudes and a later model that is very effective against low-flying planes. Both models undoubtedly have been installed aboard ship.

Surface-to-surface missiles. In this type of guided missile, the Soviet Navy leads all of the others. Small ships can accommodate missiles with massive warheads because rocket firing does not create recoil. A rocket is inherently inaccurate, compared with a projectile, but the advances in guidance and homing after World War II gave the guided missile a theoretically much higher chance of hitting than the projectile, unless shot down or diverted by electronic countermeasures. The missile must perform better than a projectile to be of value, for a ship can carry only a few large ones.

Following earlier models, cruise missiles now on Soviet ships consist of the short-range Styx (about 20 miles) and the long-range Shaddock (perhaps several hundred miles with midflight guidance). The Styx is 20 feet (6 metres) long, Shaddock at least twice as large, with two jettisonable booster rockets. The 100-ton Komar patrol boats carry two Styx missiles, the 200-ton Osas carry four. Destroyers and missile cruisers carry the larger Shaddock. Ships armed with these missiles are primarily effective against ships (or coastal areas) without air protection.

Several other nations have surface-to-surface missiles, mostly very short range, though not with the emphasis of the Soviet Union. France has two types of small missiles developed from its air-to-surface AS.11 and AS.12 types. The larger have a range of about three and a half miles (five and a half kilometres) and the destructive power of a six-inch projectile. Italy has two Sea Killer models, the longer range effective to 11–12 miles. The foregoing, of course, are primarily useful as armament for small vessels. The two countries were jointly developing a medium-to-long-range antiship missile, Otomat, that it was hoped might be operational in 1973.

Sweden has a land- or destroyer-launched missile of over 100-mile range, weighing about one ton on launch. The U.S. is developing the Harpoon, a long-range, homing, antishipping missile for launch by ship or aircraft.

Several nations have new lightweight rapid firing guns. Their high rate of fire of simple and relatively inexpensive projectiles may make them competitive with the guided missile as missile killer.

AMPHIBIANS, MINECRAFT, AND AUXILIARIES

Amphibians. The United States has set the pace in amphibious developments since World War II. The Soviet Union has made much progress in amphibious build-up, including its new helicopter carriers and improved versions of old British–American landing vessels such as her 15-knot Alligator LST (tank landing ship). The United States, however, has brought out a whole family of new or vastly improved types—generally faster, larger, and far more effective. Examples are:

The World War II LST, of which the United States built 1,041, displaced about 4,000 tons full load and transported about 150 troops with equipment at 10–11 knots. The current, fourth-generation Newport class displaces over 8,000 tons, transports amphibious craft, tanks, and other combat vehicles, along with up to 500 men, to a hostile shore at over 20 knots. Instead of simply beaching, like her predecessor, she has an extendable ramp supported by huge projecting derrick extensions on each side of the bow. As the ship grounds, the ramp shoots forward hydraulically 112 feet (34 metres). While vehicles and troops land over the ramp, amphibious tractors in the tank deck move out the stern gates.

Capabilities of newest amphibious vessels

Besides the helicopter assault carriers earlier described, becoming operational in the middle 1970s is another completely new, giant type of amphibious assault ship displacing 40,000 tons and combining the functions of several types of landing ships. By carrying a balanced load of troops, combat vehicles, and cargo for a full battalion landing team, this vessel maintains their tactical integrity together. With a large dock well, under a flight deck extending the full length of the ship, it can land the team by helicopter and landing craft. Vertical-and-short-takeoff-and-landing planes will in time give it much of the necessary capability of close air support that World War II escort carriers provided.

Two other important ships are notable. The LSD (dock landing ship) that proved so useful in World War II has been improved in speed and troop-carrying capacity.

The LPD (amphibious transport dock) was developed from the LSD but provides more versatility. "Nashville," representative of the latest, displaces 17,000 tons and cruises at better than 20 knots.

Riverine warfare has close affinity to conventional amphibious operations. Since VE Day the Soviet Union has built numerous small ships and craft to maintain strong river and coastal forces. The powerful Soviet Danube River flotilla alone has about 100 craft including 120-ton armoured gunboats, large landing craft, fast artillery boats, assault boats, and hydrofoils. Missile patrol boats and smaller vessels of the deep-sea fleet can also participate in river operations.

As a result of the Vietnam War, the United States developed several new types of river craft, including 50-foot Swifts, adapted from oil-drilling support craft; river patrol boats developed from fibre glass yacht-design

hulls, with Styrofoam flotation gear to keep them afloat even when shot-riddled; monitors converted from medium landing craft, heavily armoured and armed with a 105-millimetre (4.1-inch) howitzer, 81-millimetre (three-inch) mortar, or 40-millimetre (1.5-inch) machine gun in the turret, along with machine guns; and assault-support boats, multipurpose craft, well armoured, heavily gunned, and capable of 15-knot speed, about twice that of the other armoured craft.

Minecraft. Like other weapons, mines of each decade are progressively more sophisticated. Acoustic, magnetic, and pressure-firing devices become harder to sweep. Nuclear mines have introduced the ultimate in explosive power. The Soviet Navy has the largest minelaying force, including fast surface minelayers with four launch tracks and below-deck capacity for 400 mines. Mine racks can be installed in any ship, naval or merchant. Most submarines can serve as minelayers, as can aircraft.

Since the Korean War, the United States has maintained a strong mine-warfare capability by development of mines and minecraft, including wooden-hull, nonmagnetic vessels for sweeping some of the advanced firing mechanisms in modern mines.

Auxiliaries. Many different types of ships and craft must serve to meet the everyday requirements of a fleet in normal operations. When crisis comes, the need mounts and is met in the initial stage by ships in reserve and by conversion of commercial vessels. Of notable potential value in this connection are the Soviet Union's new merchantmen, ocean-fishing and hydrographic ships, most of which are, in concept and design, fleet auxiliaries.

Fishing and hydrographic ships

In addition, the Soviet navy has constructed a series of destroyer, submarine, and other tenders, oilers, ammunition and other special-purpose ships, emphasizing at-sea logistic replenishment and service.

Of the United States Navy's approximately 100 types of ships, a good number are auxiliaries—tenders, tankers, refrigerator ships, ammunition ships, net, survey, oceanographic research, salvage, degaussing, and other types of essential and usually unsung ships. In key combat-type auxiliaries the United States has made significant progress during the last two decades, especially in design.

FUTURE DEVELOPMENTS

Though the future of navies is unclear, it is possible to speculate on the basis of current trends.

Manned-satellite space stations may come into use to perform or monitor naval missions. Unmanned satellites for communications, navigation, mapping, reconnaissance, weather information, and perhaps even antisubmarine warfare will increase in value.

Advances in vertical- and short-takeoff-and-landing planes will cause a proliferation in numbers and kinds of warships carrying high-performance supersonic planes. As a consequence, super aircraft carriers may have reached the maximum in size. Medium-to-small carriers will take over a number of the many functions of this indispensable type of warship. The Soviet Union, China, and other nations will begin or continue light-carrier programs. Helicopter programs in many navies are part of the increasingly widespread employment of aircraft from ships. Although slow, helicopters serve admirably for antisubmarine warfare, reconnaissance, enemy-missile warning, and midrange guidance of friendly missiles.

The principal menaces to ships will continue to be undersea weapons and long-range missiles from ships and planes. Antimissile systems and weapons will grow in complexity, speed of reaction, and effectiveness. Among these the gun may make a surprising comeback. With very high rates of fire, automatic operation, rocket-assisted projectiles, and, possibly, in-flight guidance, the shell from guns may supersede many, more expensive, guided missiles.

Hydrofoils and air-cushion or surface-effect war vessels will make striking advances in size and usefulness. They have particular promise for operations through submarine waters.

Nuclear-propulsion plants will become smaller and cheaper, though remaining costly for initial installation. More surface ships will enjoy their advantages. In nonnuclear plants, the trend will be toward the use of the gas turbine.

Automation, solid-state electronics, miniaturization, module replacement, and other steps will permit reduction in size of crews. These steps will also make possible small, fast warships to fill the gap left by the growth in size of destroyer types. Many navies have already built small, swift gunboats; a more versatile type, displacing between 1,000 and 2,000 tons, is likely to be an important development.

Antiballistic-missile ships will appear, capable of intercepting hostile missiles over the ocean. Ballistic missiles will probably go into some surface ships.

Submarines will run faster, quieter, deeper, and be armed with larger, longer-range missiles. Measures to combat them will include super-quiet attack submarines, helicopters on small ships, other airborne efforts, distant-detection devices, and advanced weapons.

BIBLIOGRAPHY. Useful general works on the strategic and tactical use of modern naval forces include SAMUEL ELIOT MORISON, *History of United States Naval Operations in World War II*, 15 vol. (1947–62); STEPHEN W. ROSKILL, *The War at Sea, 1939–1945*, 3 vol. in 4 (1954–61), which have no peer in covering the strategy, tactics, logistics, and types of naval vessels prior to and during World War II; Roskill's *Strategy of Sea Power* (1962); and the *United States Naval Institute Proceedings* (monthly) and its annual *Naval Review*. These give a useful look at the naval thinking of the World War II period and since. SIR HERBERT W. RICHMOND, *Sea Power in The Modern World* (1934), is valuable background for study of the pre-World War II period. PHILIP COWBURN, *The Warship in History* (1965); RICHARD A. HOUGH, *Fighting Ships* (1969); and DONALD G. MACINTYRE and B.W. BATHE, *Man-of-War* (1969), discuss and illustrate the general evolution of the warship from early to modern times. HANSON W. BALDWIN, *Strategy for Tomorrow* (1970), on the U.S. Navy; and E.M. ELLER, *Soviet Sea Challenge* (1971), cover the development of the two current leading navies and strategy. The origin and growth of the gun-armed warship are traced in DOUGLAS G. BROWNE, *The Floating Bulwark: The Story of the Fighting Ship, 1514–1942* (1963); and FREDERICK L. ROBERTSON, *The Evolution of Naval Armament* (1921), which devotes as much attention to ships as to weapons. MICHAEL A. LEWIS, *The Navy of Britain* (1948), while focussing on the Royal Navy, provides a lucid discussion not only of the development of ships and weapons but also of tactics of war at sea from the Middle Ages to World War II. PETER PADFIELD, *Guns at Sea* (in prep.), gives a solid and readable account of the history of naval guns and gunnery techniques. E.H.H. ARCHIBALD, *The Wooden Fighting Ship in the Royal Navy, A.D. 897–1860* (1970); and HOWARD I. CHAPELLE, *History of the American Sailing Navy: The Ships and Their Development* (1949), concentrate on the age of the wooden sailing warship. WILLIAM HOVGAARD, *Modern History of Warships* (1920); and the SOCIETY OF NAVAL ARCHITECTS AND MARINE ENGINEERS, *Historical Transactions, 1893–1943* (1945), provide concise summaries of developments in ships, armament, and machinery since the introduction of steam propulsion and armour. The multivolume *Dictionary of American Naval Fighting Ships* (1959–), contains not only concise histories of ships, with profuse illustrations, but also numerous appendixes on various types. Evolutionary treatments of major modern warship types include: OSCAR PARKES, *British Battleships* (1950), invaluable also to the study of foreign battleship development; SIR ARTHUR HEZLET, *Aircraft and Sea Power* (1970) and *The Submarine and Sea Power* (1967); and NORMAN POLMAR, *Aircraft Carriers* (1969). Four general descriptive handbooks of the world's warships are *Les Flottes de combat* (Paris, 1908–); *Jane's Fighting Ships* (London, 1898–); *Taschenbuch der Kriegsflotten* (Munich, 1900–); and *Weyer's Warships of the World* (United States, 1968–). These internationally known annuals give general descriptive and illustrative coverage to the navies of this century. More recent volumes also include information on naval aircraft. *Jane's Weapon Systems* (1969–), a new annual, includes descriptions and illustrations of naval guns, missiles, fire control systems, and the like. Eight editions of JAMES C. FAHEY, *The Ships and Aircraft of the U.S. Fleet* (1939–65 and 9th ed. by SAMUEL ELIOT MORISON and JOHN S. ROWE, 1971), contain photographs and much tabulated information on the United States Navy's materiel.

(E.McN.E./R.L.Sc./J.C.Re.)

Navigation

Navigation, once considered an art, has become a science and a technology. The word is a combination of the Latin *navis*, a ship, and *agere*, to drive. Navigation therefore originally meant "ship driving," which included steering and setting the sails. At the start of the air age, the word "avigation," in which *avis* (bird) replaced *navis*, was suggested, but the older word came to be applied to air and, later, to space navigation.

Explorers in a new region must first find the way. As others follow, a path is made and signposts go up. As traffic grows, a new problem arises, avoiding collision, and in all fields of navigation, emphasis has tended to shift from finding the way toward avoiding collision. In space man is still finding the way. On the roads, the dangers are from collision. At sea and in the air, collision avoidance is growing in importance. Ships, large and easily seen, but slow moving, require a long time to stop. When many are in one area, avoiding one may bring danger to or from others. The one-way street principle is, therefore, being introduced in certain waters where the traffic is heavy. Aircraft, on the other hand, travel so fast that a pilot may fail to see another aircraft in time, particularly if it is approaching head-on. The ground organization, Air Traffic Control, therefore has the task of directing aircraft along selected paths to maintain safe intervals. Civil air navigation is governed almost entirely by the need to follow the paths chosen by the air traffic controller.

Fuel consumption is another problem the navigator must deal with. Air and sea navigators keep track of fuel consumption, chiefly for economy's sake. In space flight, fuel is the major navigational consideration. The route to be followed and even the moment for lift-off must be chosen to make the smallest possible demands on fuel.

Timekeeping, a matter of vital importance in space navigation in connection with fuel consumption, is growing important in sea and air navigation for a different reason. Today it is recognized that a voyage or a flight is a link in a chain of transport that carries people and goods from one place on land to another place on land. Ships and aircraft, therefore, increasingly must fit schedules into road and rail timetables.

The four basics of navigation

Navigation, in short, has to do with the whole of a journey from start to finish and is concerned with four basics: (a) finding the way; (b) avoiding collision; (c) fuel consumption; and (d) timekeeping.

The classical definition of navigation is "the art of conducting a vessel on its ways." Today, a better definition might be: "navigation is the science of conducting a craft or creature on its ways."

Development of navigation

ANIMAL INSTINCT

Remembering the way. An animal finds its way home by remembering what it saw or smelled on the way out. This process is common to all creatures, including birds, insects, and fish. The eel, which travels great distances across the ocean, is believed to find the way back by remembering traces of smell in the ocean around. Birds, on the other hand, navigate almost entirely by sight and are extraordinarily adept at recognizing landmarks from the air.

Direction and distance. When a young bird migrates, it instinctively aligns itself in a certain direction, flies until it has no more energy, and then lands. To set out in the right direction, the bird instinctively uses the sun by day or, by night, reacts to the pattern of stars around the pole. To use the sun by day, the bird has to make allowance for the change in direction from rising through to setting. It, therefore, needs a timing instinct that can correct for the passage of the sun across the sky. Not only birds but insects and fish have such timing instincts. Existence of these instincts and migration patterns has been confirmed by extensive experimentation (see MIGRATION, ANIMAL).

When a bee, laden with honey, returns to the hive, it dances on an upright piece of honeycomb. The slowness of the dance shows how tired it is and, therefore, suggests how far it has travelled; the direction of the dance shows the direction from which it has come in relation to the sun. The other bees can then set out knowing instinctively the distance and direction of the honey-bearing area.

Navigating by the sun. A little wild Manx Shearwater, taken from its nest off the coast of Wales, carried across the sea, and released in Boston was back on its nest in twelve days. Though nobody knows exactly how such exploits are achieved, it can be shown that a bird's eye can watch the sun and the horizon at the same time and is so sensitive that it is able to detect the motion of the sun across the sky by comparison with the horizon.

By such methods, wild birds have travelled extraordinary distances. When released hundreds of miles from their homes in sunny weather, they can often set out for their nests within a few minutes. When 50 or so miles away, they are able to pick up landmarks that they have seen before and home in to their mates.

HUMAN ART

Maps and charts. In featureless country, primitive man cut notches in tree trunks and built distinctive piles of stone; in the early days of sailing, he lighted bonfires on headlands. As ships became larger, channels of deep water were marked with sticks and eventually with buoys, lighted so they could be seen at night.

At the same time, men were developing records as aids to navigation. The Polynesians tied sticks together in a pattern in which the intersections represented islands. Arab and European sailors of the Middle Ages developed charts, at first only to help find the way but later to record their voyages. By the 17th century these charts showed coastlines and lights and, at the end of the 18th, depths of water. In the early 1900s, the airman developed special charts to record his paths, but he also carried maps to help him identify objects on the ground.

The spherical surface of a globe cannot be portrayed on a sheet of flat paper without distortion. For navigation, this distortion has to be "orthomorphic"; that is, the stretching at any point must be equal in all directions, even though the amount of stretching and therefore the scale may vary from one point to another. Unless a map or a chart is orthomorphic, a protractor cannot be used to draw directions accurately anywhere on its surface (see also MAPS AND MAPPING).

Types of navigational charts

Three types of orthomorphic charts are used in navigation. A cylinder may in effect be placed around the equator and the surface north or south of the equator stretched orthomorphically until it fits the cylinder. This produces the Mercator projection. Alternatively, a band of latitudes may be stretched orthomorphically into a cone to produce the Lambert conformal projection, known sometimes as a conical orthomorphic. Finally, a flat plate may be placed at either pole of the globe and the globe stretched orthomorphically to produce a stereographic projection.

The Mercator chart has one great advantage. Since it is produced by flattening an orthomorphic cylinder, all of the lines of longitude are parallel. Thus, a vessel travelling in a constant direction will cross these lines at the same angle and the path will appear as a straight line. As the distance from the Equator increases, so does the orthomorphic stretching, so that, on a Mercator chart of the world, Greenland appears larger than the United States.

The shortest distance between two points on a Mercator chart is only seldom a straight line and more often is a curve known as a great circle. In order to follow this shortest route, navigators formerly calculated the points at which the great circle from a starting point to the port of arrival would cut the various lines of longitude and joined these points by a series of straight line "legs" on the Mercator chart (see HYDROGRAPHIC CHARTING). The introduction of radio, the waves of which in general reach a craft by the shortest route, has made the Mercator projection even less convenient. It is, nevertheless, still used in marine navigation though frequently replaced by the Lambert conformal in air navigation. In polar regions, however, the stereographic projection is adopted.

Compasses and logs. Primitive man used the sun and stars as guides. When the sky was overcast, he learned to determine directions by the weathering of rocks sheltered from the sun, the branches on the sunless side of trees, the anthills oriented toward the warmth of the morning, and other such signs. Early seamen observed migrant birds and sometimes carried birds on board, releasing them to find land.

The magnetic compass

At about the beginning of the Christian Era, the Chinese discovered that a piece of iron ore, when floated on water, always points north. Carried westward by the Arabs, this early magnetic compass was universally used by Columbus's time.

It was soon realized that the direction indicated by a compass needle often varies considerably from true north though the reason for this was not understood till quite recently. The north magnetic pole, or the place on the earth to which a compass needle points, is located some distance from the actual North Pole. Thus, a magnetic compass is usually in error. This error may be great or small depending upon one's location on the earth's surface. In addition, iron ore bodies in the ground may affect compasses. A practical solution to the problem was found in the 19th century. Magnetic compass points were measured at a number of places, with the sun or the stars used as references. Compass error, recorded as so many degrees to the east or to the west of true north, was compiled and transferred to charts. Now a navigator can look at a chart to find the amount of compass error at his location.

As iron and steel replaced wood in hulls, the ship itself acted as a magnet, distorting the compass reading in the direction in which the ship was pointed. This deviation was corrected by experimentally fitting pieces of iron and steel close to the compass in such a way that they offset the magnetism of the ship. Compass compensation was later applied to aircraft.

In days of sail, a ship's speed could be crudely measured by throwing overboard a small log tied to a line which was paid out as the log drifted astern. The line was knotted at intervals and the number of knots paid out while a sandglass was emptied gave the speed of the vessel in knots.

In a later, more sophisticated method, the log of wood was replaced by a trailing propeller, whose revolutions recorded the ship's speed. A method, developed originally for aircraft, measured the pressure set up in the open end of a forward-facing tube, known as a pilot tube. Like the propeller, this gives a speed measurement that can be translated into distance if the time of travel is known. Seamen, however, continued to refer to all such devices as logs and to record speed in knots, a knot being one nautical mile per hour. A nautical mile, originally 6,080 feet, is today 1,852 metres, or 6076.12 feet.

Dead reckoning. Starting from a known point, the mariner with his compass could draw a line on his chart to represent his course, along which he marked off the distance given by the log. The calculation was known as deduced reckoning, shortened later to dead reckoning. In addition to errors in the compass and in the log, dead reckoning suffered from errors due to the drift of the water. When ocean currents were first marked on charts of the open sea, and tidal streams appeared on coastal charts, navigators could make allowance for drift. Fortunately, the currents were seldom fast and, on long voyages, often tended to cancel each other out. The situation in the air, however, was quite different. Early airplanes flew at around 100 knots and the air that supported them was blown over the ground by the wind at speeds of up to 40 knots. It was therefore necessary to determine the velocity of the aircraft through the air and the velocity of the air over the ground in order to find the true velocity of the aircraft with respect to the ground. This was achieved by the triangle of velocities as shown in Figure 1. A line was drawn to show the direction in which the aircraft was heading, the length of the line representing the distance the aircraft would travel through still air in one hour, in other words, the true air speed. Such a line is called a vector; *i.e.*, it indicates both magnitude and direction. From the end of this line, a second line or vector

Calculating aircraft velocity

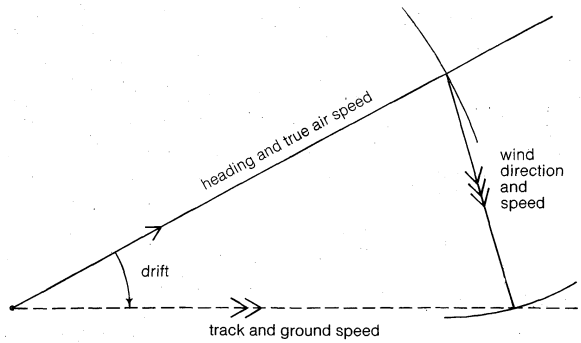


Figure 1: Triangle of velocities.

was drawn in the direction toward which the wind was blowing, the length of this second line being determined by the wind speed. Finally, a third line drawn from the starting point of the first line to the end of the second line showed the path that the aircraft was following over the ground, the distance representing the true ground speed. Stated mathematically, the true ground speed was the vector sum of the craft's air velocity and the wind velocity.

The angle between the heading of the aircraft and the track was known as the drift because it resulted from the drifting effect of the wind. Early aircraft were fitted with drift sights through which the airman visually aligned a grid with the moving ground below and so determined the drift. The plotting of triangles of velocity was simplified by using graphic instruments called computers before that term was appropriated for much more complex devices.

Paradoxically, higher aircraft speeds failed to eliminate the problem of wind drift because jet aircraft fly higher as well as faster and, above 20,000 feet (6,100 metres), very narrow belts of wind known as jet streams that travel at speeds of 100 or 200 knots occur under certain meteorological conditions.

Sea navigators did not follow the practice of air navigators and allow for ocean currents and tidal drifts in their initial calculations. Dead reckoning, long established as a navigational technique, continued to be used; an estimate for ocean current or tidal drift was added afterward. This practice continues today.

By dead reckoning, the navigator can quickly find a position that he can often check by a landmark. On the other hand, the errors that are inseparable from dead reckoning accumulate; when a position has been checked, the reckoning is therefore generally restarted from that position.

Dead reckoning enables the navigator to plot not only where he is but also where he will be provided he maintains his planned course and speed. It also makes it possible for him to plan his whole journey, including the time of arrival at his destination, before he sets out. Planning ahead is a part of all navigation; the preparation of a complete flight plan is mandatory before taking off in a civil aircraft. Space navigation is based even more completely on flight planning and the time of splashdown is calculated within minutes, many weeks before lift-off.

Direction finding. To avoid the shoals marked on his chart, a seaman needs to know his exact position. By means of a sight fitted to his compass, he can measure the direction of a headland or a light. This direction, or bearing, can be marked on a chart by a position line. A second line drawn from a second bearing will fix the position of the vessel as shown in Figure 2.

The invention of radio led to direction finding, an improvement in this navigational technique. A loop of wire perpendicular to the direction in which radio waves are travelling picks up a minimum signal. By tuning in a particular radio station and then rotating the loop for minimum signal pickup, the direction of the radio station can be plotted as a position line. Plotting a similar line for a second radio station enables a navigator to find his position.

Radio direction finding, known as DF, developed in two ways. First, radio transmitters, or "beacons," were

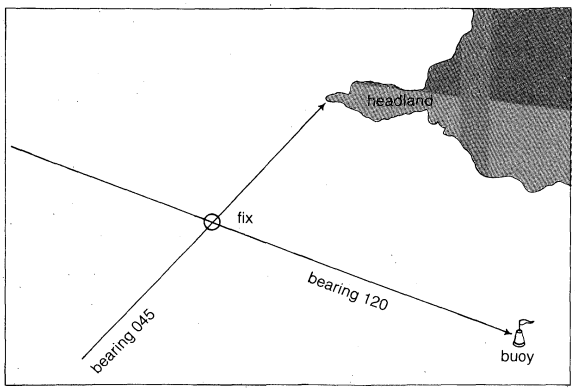


Figure 2: Determining position by intersection of compass bearings to known points.

Radio direction finding

sited, particularly along coasts, to enable ships or aircraft to fix their positions. Second, ground DF stations that could pick up radio signals sent out by a ship or an aircraft were built. If two or more stations could pick up the signals, they could fix the position of the craft. In spite of these developments, DF has remained an uncertain art, because reflections from layers of ionized air high in the sky (ionosphere) and refractions (bending of the radio waves by the atmosphere) distorted the direction of arrival of the waves.

Navigation by the sun and stars. The Polynesians used the stars effectively in navigating the South Pacific. They knew which stars passed directly over various islands and sailed their catamarans north or south until the proper star was overhead, then east or west, keeping their craft directly below the passage of the star, until they reached their destination.

The Polynesian had to wait until the star was at its zenith before he could tell whether it was passing overhead or not. Medieval Norse navigators, however, recognized that the Pole Star moved but little in the sky and that its altitude above the horizon was nearly constant for each destination. Still later it was realized that this altitude is approximately the same as the observer's latitude, as suggested in Figure 3. Thus, at the North Pole, latitude 90°, the Pole Star is roughly overhead, or at 90° altitude; at the equator, latitude 0°, the Pole Star is on the horizon at an altitude of 0°. Halfway between, the altitude and the latitude are both 45°.

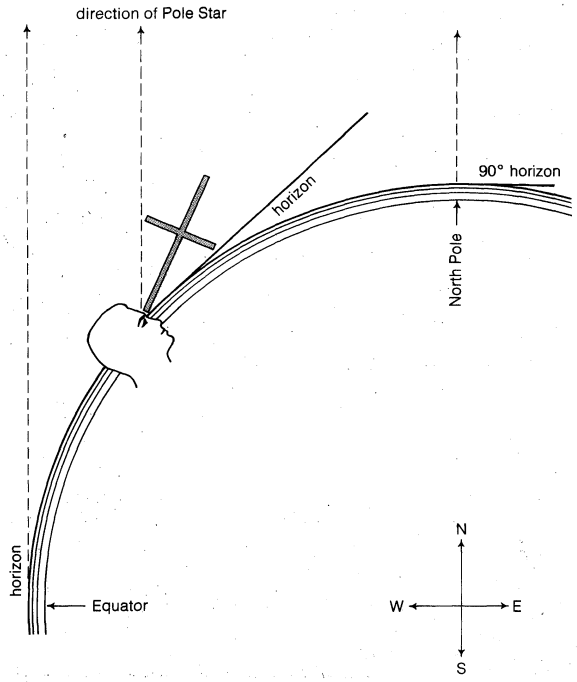


Figure 3: Altitude of Pole Star and latitude.

In Figure 3, the altitude of 45° is being measured by a simple instrument known as a cross staff, consisting of a cross piece that slides along a graduated haft or handle. One end of the haft was held against the eye and the cross piece moved until the horizon was at one end and the star at the other.

A method was subsequently developed for finding latitude by measuring the altitude of the sun at its highest point, at noon. Later on, the moon and the planets were used. Because these bodies move about the sky during the year, their heights compared to the vertical at the Equator were tabulated in almanacs, providing a reliable guide to latitude. But to determine longitude it was necessary to know the position above the earth that the stars had reached, which required an accurate means of measuring time. The water clocks of the ancients had been supplanted by pendulum clocks driven by weights, but pendulums were upset by the rolling of a ship and, besides, since gravity is not constant over the surface of the earth, the period of a pendulum varies. In the middle of the 17th century the offer of a large prize for an accurate means of measuring longitude hastened invention of the balance wheel and the hairspring (see CLOCKS, WATCHES, AND SUNDIALS) and led in turn to the development of the chronometer, a very accurate timepiece. Accurate timing was now possible. Chronometers became standard equipment on ships and, in the early 1900s, the development of radio made it possible to check a chronometer by time signals broadcast from a radio station.

Use of the sextant

Measurement of altitude was made much more precise by development of the sextant, an instrument consisting of a mirror that reflects the light from a star onto a silvered glass so arranged that an observer can simultaneously see both the star and the horizon. By shifting the angle between the mirror and the glass until the star appears to sit on the horizon, the observer produces an angle between the mirror and the glass that is just twice the altitude of the star. In a sextant, the angle can be as much as 60°. In a similar instrument called an octant, the angle is limited to 45°.

For air navigation, in which the horizon is seldom sharply defined, the sextant was equipped with an artificial horizon in the form of a liquid level, the star being aligned with the centre of the bubble. Because accelerations cause the bubble to swing, the sextant was provided with an automatic integrating or averaging device. Later developments included the periscopic sextant in which the star is viewed through a periscope protruding through the top of the fuselage.

Sea-navigation practice for many years was to take one altitude on a suitable body to find the latitude and a second altitude on another body to find the longitude and to calculate the position by basic trigonometry. In 1837 one Captain Sumner, finding himself close to a dangerous coast with only one star in sight, worked out a number of possible positions for his ship assuming various latitudes. These positions fell along a straight line. From this accidental discovery, systems were developed to translate an altitude taken on a star into a position line on a chart. A second observation on a second star produced a second position line. A third observation with a third position line was generally used as a check.

Other aids. The Viking sailor took soundings with a lead weight on a line, hauling in the line and measuring it by the span of his arms. The action reminded his companions of an embrace, and for centuries depth was measured in 6-foot (1.8-metre) intervals called fathoms from the Old Norse word *fathmr* meaning an embrace. The lead was often given a hollow bottom filled with tallow to pick up a sample of the seabed for comparison with the chart as a check on the position of the ship.

Distance from a cliff could be measured by timing the echo from a shout or the beat of a drum. In World War I, echo sounders were developed to time the interval between sending out a sound from the ship's keel and receiving an echo from the bottom, a system also used to detect submarines. Meanwhile, the airman, like the mountaineer, was using a form of barometer (atmospheric-pressure indicator) to get some idea of his height. The

height indicated by the barometer depended on the pressure at sea level and on temperature, and the errors were apt to be considerable. A height estimated by this method is accordingly known as an altitude, thus emphasizing that true height is not being recorded.

Up to the end of World War II, the shipping lanes were not unduly crowded. With a few exceptions, ships were not large and were reasonably manoeuvrable so that each master could avoid collision by using his own judgment. At night ships carried "navigation lights," which enabled other seamen to tell which way a ship was going. Similar lights were carried by aircraft but, to prevent these lights being mistaken for stars, they were made to wink. In general, the navigation light on the port side of a craft, the left-hand side looking forward, is red and the navigation light on the starboard side of a craft, the right-hand side looking forward, shows green.

NAVIGATION AS A SCIENCE

The three stages of navigation. Sailing or flying aimlessly is not navigating. To navigate, an objective is necessary. There are three stages of navigation. The first stage is the route along which the craft moves. The route has to do with position either on the surface of the sea or below or in the air, on land, or in space. To follow the route, the second stage is employed; the craft must move with a certain velocity, which has two elements: direction, which has already been defined as the course, and speed. Velocity must be built up at the start, slowed down at the finish, and adjusted at various stages during the voyage. These changes of velocity are accomplished by accelerations applied along the length of the craft by altering the motive power to adjust speed or across the length of the craft by steering. Accelerations are therefore concerned with the third stage of navigation, driving the craft.

The process of navigation. As a craft proceeds on its way, the navigator determines his position. This enables him to check his progress compared to the predetermined route and, if he finds an error, amend his route, select a new course, possibly change speed, and drive his craft accordingly. Progress along the amended route is then checked again by finding further positions, recorrecting course and speed, and driving the craft accordingly. The continuous flow of information and action involved is known as a "closed loop."

Closed loops

Navigation loops. Closed loops are seldom simple but consist of a series of loops one inside the other like the skins of an onion. Inside the outer position-finding loop there is an inner course and speed loop. In this loop, the course to be steered and the speed to be maintained are checked by compass and speedometer or log. Within the inner course and speed loop, there is also an innermost driving loop that relies on instruments to measure accelerations or, alternatively, uses the human being as an instrument. Finding the way thus involves three loops. Of these, the outer position-finding loop is used only at intervals, the course and speed loop is corrected more frequently, and the innermost driving loop is continually in action. A motor vehicle driver, by analogy, looks only occasionally at signposts but always keeps one hand on the steering wheel.

The aids that are employed in the outer position-finding loop cannot effectively be adapted to measure course and speed. For example, in the case of an aid that tells the distance to the destination, a minor irregularity, which momentarily diminishes this distance, would be translated as a complete reversal of speed, a particularly undesirable manoeuvre.

The reverse process is, however, quite possible. It has already been shown how, by dead reckoning, course, speed, and time can be used to find position. Similarly, devices used for driving may be adapted to give useful course and speed information and even position. To avoid collision, an outer position loop is used that relies on separation; for example, by one-way streets. There is also a device called an inner course and speed loop that enables vehicles to avoid more imminent collision with each other.

Modern navigation techniques

DRIVING

Steering a craft. That steering is a complex process can be seen in the problem of steering a ship down the centre of a canal. Presumably, if the ship veers to the left, turning the rudder to the right by an amount depending on how far the ship was off to the left would return the vessel to the centre line.

On arrival at the centre line, however, the ship will be headed to the right and the rudder will point dead ahead so the ship will veer off to the right. Though this can be corrected by turning the rudder left, the same thing will happen again, and the ship will weave back and forth across the centre line and never settle down on it.

If the rudder is turned by an amount gauged by how fast the ship is departing from the centre line, the ship will settle down in a straight line travelling along one side of the canal. If the system depending on distance and the system depending on speed or rate of change of distance are mixed, most of the problems are solved. In the initial stages both systems combine to correct the ship; as it turns toward the centre line, the rudder effect due to the rate of change of distance tends to steer the ship to the left as if anticipating arrival at the centre line.

Rudder effect

If the canal is straight, the ship settles down on the centre line with the rudder straight out astern. If the canal turns, however, the ship continues ahead at first, then drifts off to one side of the centre line. The rudder turns to offset the movement away from the centre line, the amount of turning increasing until eventually the craft settles down and follows the curve of the canal. The distance off the centre line is taken up with making the ship turn round the curve. Therefore, the ship travels round the outside of the bend. To correct for this "hang-off," the rudder is turned by an additional amount until the error is eliminated.

Thus it is evident that steering a craft is not simple. The human navigator uses his brain to apply the necessary corrections. Computers can do the same job; modern control systems in craft may be designed so that they can learn as they go along. Such equipment is described as "self-adaptive."

Stabilization. The faster a bicycle's wheels spin, the steadier the balance; the spinning wheels tend to keep spinning in the same direction. Since a ship or aircraft cannot be stabilized by fitting a huge wheel inside, a small model of a spinning wheel, a gyroscope, is used instead. This gyroscope is mounted on gimbals, devices that permit it to remain level even if its support (the craft) is tipped by rolling or pitching. Special electrical devices known as pick-offs are mounted on the gimbals so that, when the craft is not aligned with the gyroscope, electrical signals will indicate the amount of tilt or direction change (see also GYROSCOPE).

In a ship stabilizer, signals from the pick-offs on the gimbal control motors that drive fins fitted to either side of the hull. When the ship rolls to port, the gyrowheel stays upright but the pick-off sends a message to the port motor to drive the port fin so that the rear end is forced downward and the water flowing past the hull presses it upward. Simultaneously, a pick-off sends a message to the starboard motor to drive the starboard fin so that the rear end is forced upward and the water flowing past the hull presses it downward. Thus the pressure of the water on the fins tends to correct the roll.

Stabilizers are fitted to passenger liners. Similar devices, known as autostabilizers, are fitted to aircraft. Autostabilizers correct for roll by means of ailerons fitted in the wings that act exactly like the fins on a ship. Pick-offs on the other gimbals of the gyroscope detect also any changes in the fore-and-aft attitude, known as changes in pitch, which are corrected by signals sent to motors that drive the elevators up or down.

Gyroscope drift. Even the best gyroscopes tend to "drift"; *i.e.*, lose alignment. If not corrected, this drift causes a stabilizing gyroscope to tilt; and the ship in turn will lean at an angle instead of remaining upright. The gyroscope is therefore corrected by the force of gravity. The simplest method is to fit a small weight to the lower

part of the mounting of the wheel. The weight must be heavy enough to correct the tendency of the gyroscope to drift, and yet not so heavy that, when the ship rolls, it causes the gyroscope to swing to and fro like a pendulum.

In actual practice, the weight may not be mounted on the gyroscope but an accelerometer may be used instead, in the form of a spirit level that, when tilted by the rolling of the ship, will signal the gyroscope. These signals cause the gyroscope to react only very slowly so the effects of each individual roll are negligible. Over a long period, however, signals from the accelerometer ensure that the gyroscope remains upright in spite of its drift.

Autohelmsman and autopilot. Stabilizers and autostabilizers keep the craft level, greatly easing the task of the human helmsman or pilot. Autohelmsmen and autopilots are automatic devices that drive the craft, taking instructions for this purpose from steering and positional aids. Perhaps the simplest kind of autopilot is the steering vane fitted to single-handed ocean racers. The lone yachtsman sets the vane according to the wind and it then keeps the boat on a steady course as long as the wind comes from a constant direction. Perhaps the most complex form is the automatic landing autopilot, which consists of three separate autopilots arranged so that, if one should fail to function satisfactorily, the other two will override it and keep the aircraft on its correct path.

Autohelmsmen and autopilots use gyroscopes to measure any deviation of the craft from its desired path. In certain instances, accelerometers are also employed. In aircraft, autopilots are sometimes supported by autothrottles that keep the aircraft at a constant speed selected by the pilot. In addition, aircraft are fitted with directors. These instruments tell the pilot which way to move his controls and, in effect, are autopilots that use a human being as the source of control.

Instruments. Driving instruments measure and record engine conditions, fuel status, and the position or attitude of the craft in relation to the earth. Engine instruments usually measure the rotation of the engine or of the turbine that blows air into the engine. They also measure the pressure inside the engine either directly or by the temperature it generates. Fuel flow is also often displayed on an instrument, as a useful guide to engine performance. On a long-term basis, fuel in tanks is also recorded.

Artificial horizon

In aircraft, an artificial horizon is generally necessary. When flying visually, the pilot may use the natural horizon, but since this is often obscured by clouds, particularly when flying high, the pilot uses an instrument consisting of a model of an aircraft behind which is fitted a representation of the real horizon. This artificial horizon is kept parallel to the real horizon by means of a gyroscope; a linkage to the gyroscope causing the horizon to move down or up, exactly as a real horizon seems to, when the aircraft pitches nose up or down.

Aircraft also carry air-speed indicators that indicate the pressure set up by the forward motion of the aircraft. They therefore indicate the negative pressure set up on top of the wings due to the forward motion; that is, the lift of the wings. When this pressure falls below a minimum value, the pilot knows that the lift may collapse and the aircraft will stall; that is, fall out of control.

An aircraft and occasionally a ship, such as a powered barge that must continually negotiate bends in rivers, may carry a rate-of-turn indicator in which a gyroscope detects how fast the craft is turning. Most craft also carry devices to alert the driver to emergencies; these vary from the simple warning lights used in automobiles to complex electronic pictures and statements of malfunctions printed automatically on cathode-ray tubes.

Inertial navigators. By means of very precise accelerometers, it is possible to measure the direction of the force of gravity with extreme accuracy. Gyroscopes can also be set to spin in the direction of the force of gravity at the point where the journey started. The difference between these two directions shows the distance and direction in which the craft has travelled. Aircraft inertial navigation "platform" is the name given to such an accelerometer and gyroscope package. Inertial navigators also require computers to translate the indications of the platforms into headings, speeds, and positions and to correct the gyroscopes to allow for the rotation of the earth and other effects.

Inertial navigators used at sea are even more accurate than those used in aircraft. Submarines have been navigated automatically by these systems, travelling under the polar ice cap with errors of less than a mile a week. Inertial navigators are also used to guide long-range intercontinental ballistic missiles. These navigators steer the ballistic missile for the few minutes the motors are burning. At the end of that time, they ensure that the missile is on the desired trajectory. After that, the motors are shut off and no more guidance is necessary.

A space vehicle is effectively weightless when its propulsion motors are not operating. Therefore, inertial navigators cannot be used to find position except during launch. An inertial platform is used, however, to keep the space craft on a steady alignment during a brief period of acceleration, or "burn." Before acceleration it is aligned with the stars. This steady alignment is essential. Otherwise, the motor could accelerate the space craft in the wrong direction. The accelerometers on the platform can, of course, be used to measure the change of speed resulting from the brief acceleration.

STAYING ON COURSE

The modern compass. A magnetic compass is so seriously affected by the metal in a ship or an aircraft that it is mounted as far away as possible from ferrous material, necessitating that the signals from the compass be led into the ship's bridge or the pilot's cockpit. A distant reading compass of this kind does not use a magnetic needle mounted on a pivot but instead has fixed detectors that can sense the magnetic direction without rotating.

In a high-speed marine craft or in an aircraft, turning the craft causes the magnetic detector to swing in the direction of the turn, thus causing an error. Fixing the detector to the hull would also give erroneous readings when the ship or aircraft rolls. Similar errors occur when speed increases or decreases. Thus signals from the magnetic detector are passed to a gyroscope that acts like the gyroscope in a stabilizer when controlled by an accelerometer. The irregularities in magnetic direction are smoothed out and the gyroscope maintains correct alignment. The compass indicator is driven by the gyroscope. Such a device is known as a gyromagnetic compass.

Gyromagnetic compass

On turns, magnetic-compass errors may be very high, and even with a gyrocompass, they may be substantial. In general, however, a gyromagnetic compass will indicate the direction of the earth's magnetic field to within less than a degree. It is unfortunate that, over much of the world, this direction of the earth's magnetic field is only known to an accuracy of 2° or 3°.

Special compasses have been devised to measure direction by means of the sun and the stars. Originally little more than simple sighting devices, these "astro-compasses" required books of tables similar to those used for astronavigation. More recent, automatic devices, driven by clockwork, allow for passage of the sun across the sky. In spacecraft, alignment is not affected by the rotation of the earth and a relatively simple telescope can be used.

Gyrocompass. If a gyroscope is put on the rotating earth with its axis at right angles to the earth's axis as shown in Figure 4(a), it is tilted as the earth rotates and therefore turns until it is aligned with the earth's axis as shown in Figure 4(b). Thereafter it is not tilted by the earth's rotation unless it drifts off a little to one side. It is then automatically tilted back to the correct alignment. Such a device is called a gyrocompass.

The gyrocompass can be used in ships that travel along steady courses. It is affected, however, by turns which it cannot distinguish from rotation of the earth. Although on a steady course the error may be less than a degree, in a turn even at a moderate speed, the error can be two or three degrees. Hence the gyrocompass is not used in high-speed marine craft nor in aircraft. An inertial platform can act like a gyrocompass, however, and can be made to align itself in the correct direction before the start of a voyage or a flight.

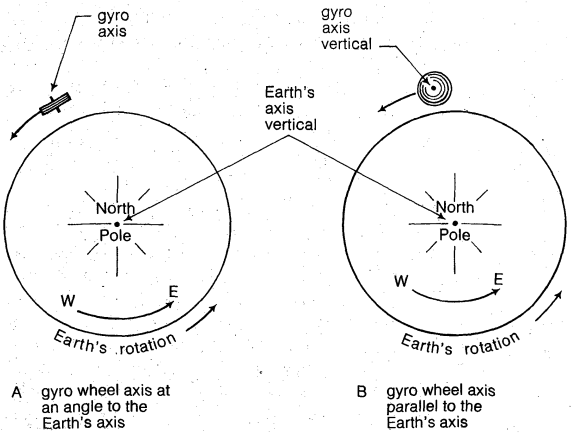


Figure 4: Principle of the gyrocompass.

Speed measurement. The classical methods of measuring speed through the water have already been described. In ships, the modern log may take the form of an electromagnetic detector that senses the speed of the water flowing past it. In many ships, as in racing cars, engine revolutions are used as an indication of speed. Inertial navigators also provide accurate measurements of speed.

The latest method used for measuring speed is known as Doppler. When a racing car goes by, the high pitched whine of the approaching engine falls to a low pitched growl as it disappears into the distance. By measuring the change of engine note, it is possible to measure the speed of the racing car. This principle is used in Doppler radio.

Doppler radio

Radio waves of a certain frequency are transmitted by a craft and, when they strike a moving object, are reflected either with a higher or a lower frequency according to whether the craft is approaching the object or receding from it. By this means an aircraft can measure the speed at which it is passing over objects on the ground and, by means of other transmissions at right angles, it can also measure the sideways speed over the ground. Similar Doppler systems have been used to measure the speed of hovercraft and land vehicles over the ground.

Collision avoidance. Figure 1 illustrates the calculation of an airplane's true velocity. A similar procedure can be used to calculate the course an airplane must avoid to prevent collision with another aircraft. In Figure 1 the wind is replaced by the course and speed of the craft drawn in the opposite direction. What was track and ground speed in Figure 1 becomes the line of sight to the craft to be intercepted and the speed at which the two planes are approaching each other. If both planes maintain the speeds and directions indicated in Figure 1, a collision will occur.

Recent interception theory is based on a development of the way in which an animal hunts its prey. It runs toward it and, as the prey tries to escape, it changes direction so that it is always running straight toward it. Thus the hunter changes course at the same rate as its line of sight to its prey changes direction. Missiles that have to intercept ships or aircraft do a little better than this; they alter course faster than the direction of the line of sight changes.

Modern collision-avoidance theory depends increasingly on the fact that, if course is altered in a direction opposite to that in which the line of sight to another craft is changing, the miss distance will be increased. Thus if a ship is apparently travelling across the bow to the left, the miss distance will be increased if the course is altered to the right. Similarly, if the other ship is on the same course but moving ahead, the miss distance will be increased by slowing down or, when it is well ahead, by turning toward it.

The real difficulty arises when the line of sight to the other craft is not changing direction. This means that either the other craft is diverging or converging directly toward the first. Traditional "rules of the road" at sea require two ships meeting head-on to both turn right. The turn has to be sharp to be effective and to make intentions clear.

Aircraft are too small and fast for visual avoidance. Air collisions have occurred in daytime in absolutely clear weather. Collision avoidance depends on systematic separation of flight paths. Radio systems are now being developed, however, to aid evasive action when, by some accident, the separation proves to be inadequate.

Dead reckoning by computer. In modern craft, dead reckoning is being left more and more to computers. Military land vehicles use computers fed by compasses and signals taken from the wheels. Ships depend on the gyrocompass and the log, and aircraft on the gyromagnetic compass and Doppler. Recently, computers have been devised that generate printed positional information. Generally, inertial navigation provides the dead-reckoning information only, though compass and Doppler data may be included with inertial outputs. Information from radio-fixing systems may be added to the dead reckoning; each item of information is automatically compared with the others.

Instead of using a chart, position may be continuously displayed on a moving map. If the route is known beforehand, the map may be in the form of a strip driven along by rollers with a pointer moving to and fro to indicate exact position. When greater flexibility of route is essential, a series of maps covering a very wide area may be printed on film and projected on a ground glass screen on the bridge or in the cockpit.

The most important use of dead reckoning is by Air Traffic Control. Not only does every aircraft have to file a complete flight plan showing courses and timings but, during the flight, air traffic controllers keep what is virtually a dead-reckoning check on each aircraft. The flight plans filed by each aircraft before take-off are split up into sections and "flight strips" are prepared showing height, speed, and timings of each aircraft in a section. The air traffic controller handling that section can thus see which aircraft will be arriving and in what sequence. These flight strips are continually updated by reports sent by radio from the aircraft and also by changes in height and course ordered by the air traffic controller to maintain adequate separation. In modern systems, automatic flight strips are displayed on cathode-ray tubes.

Flight strips

POSITION FINDING

Sound, radio, light waves. To find the position of a craft in relation to a ground feature, a link is needed between the craft and the earth. In inertial navigation, the link is the direction of gravity. In astronavigation the horizon is used. In all other systems, the link is waves. Sound waves travel through water at about 5,000 feet (1,520 metres) per second and through air at about 1,100 feet (335 metres) per second. Electromagnetic waves in the form of radio or light waves travel about 160,000 knots (nearly eight times round the world) in a second. The difference between radio and light waves is length; radio wavelength varies from many kilometres to less than a centimetre; light waves are so short that a thousand of them stretch only one millimetre. Because wavelength varies slightly according to whether the wave is passing through space, air, ground, or water, radio waves are normally described by frequency. Radio waves longer than one kilometre are termed low frequency and radio waves shorter than ten metres as high frequency.

At heights varying from tens to hundreds of kilometres above the earth, layers of ionized particles (ionosphere) are formed by the ultraviolet light from the sun, the lower levels dying away during the night. Long waves (low frequency) usually bounce off the ionosphere and also tend to follow the curvature of the earth, producing strong ground waves. As a result, they can be reliably transmitted over long distances. Short waves that strike the ionosphere at a small angle are also reflected, but those that strike at a large angle (nearly perpendicular, close to 90°) will pass through. At these frequencies, the ground wave dies out rapidly so that there may be a dead space before the first of the ionized layer reflections, known as "sky-waves," hits the ground. This dead space is known as the "skip distance" (see also RADIO).

As radio waves become shorter, they behave more like light. Centimetric waves travel mostly in straight lines

and penetrate the ionosphere easily. They bounce off objects including large rain drops and, as the length decreases, are more and more readily absorbed by mist and fog. Longer waves, on the other hand, though affected by lightning and static discharges, sweep undisturbed through water particles.

Radio-bearing systems. The loop antenna used for direction finding is almost as old as radio itself. It has been developed into an automatic direction finder (ADF) in which a motor turns the loop, the device originally being known as a "radio compass" because it could guide the navigator to a radio destination. ADF instruments are often linked to compass systems so they display not merely the direction of the radio station compared to the heading of the craft but the actual direction as plotted on a chart. In ground stations, two fixed-loop antennas at right angles have long been used; the direction of origin of a transmission may be displayed instantaneously on a cathode-ray tube. The device can also indicate the direction of thunderstorms.

Automatic direction finder

Direction-finding stations for aircraft appeared between the two world wars, consisting of loop antennas in pairs at right angles, arranged so one antenna radiated radio waves in a pattern of dots and the other radiated waves in a pattern of dashes. The dashes fitted into the intervals between the dots. Midway between the directions of maximum dots and maximum dashes, dots and dashes merge in the receiver into a steady note; this is called the "beam." Only a slight deviation of the pilot's headphones away from this beam produces a change in the steady note, and he determines the direction by which he is "off the beam" by whether the dots or dashes predominate. The pilot flies in one of four directions to and from the beacon, which is known as a "four course beacon" or a "radio range."

The limited range of the radio signals from this device limited the directions in which an aircraft could travel from one beacon to another, and it was therefore replaced by VOR (very-high-frequency omnidirectional radio range), the beacons of which operate on a completely different principle. One antenna sends out a steady wave in all directions. A second antenna rotates and sends out a narrow beam of waves such that, when on north, the beam coincides with the steady wave, when on east it diverges by 90°, when on south by 180° and so on. A receiver in the aircraft measures the divergence or phase difference and records the direction of the VOR beacon that is displayed together with the heading of the aircraft on a radio magnetic indicator (RMI). Today, beam antennas may be rotated electronically.

From the radio range, with its so-called beams, true beam systems have developed. In these the loops are replaced by antennas that concentrate the radio waves into narrow beams a few degrees wide, while the dots and dashes are replaced by more sophisticated patterns or modulations. In the instrument landing system (ILS), used to help aircraft approach and land on an airfield, the two beams use very short waves to keep antenna size small (see also ANTENNAS AND WAVE GUIDES). The beams point almost in the same direction and, once the aircraft has entered the beams, a receiver can measure the angular displacement from the centre line and display this displacement on an instrument or use it to guide the aircraft along a line toward the point of landing. In addition to the steering beams, which make up the localizer element of the ILS, there are two similar but ever narrower beams transmitted in the vertical plane that guide the aircraft down the correct glide slope to the point of touchdown.

Distance measurement. The echo sounder, already mentioned, transmits a short sharp "ping" from the keel of the ship and measures the time required for the echo to return from the bottom of the sea. A radar altimeter similarly measures the vertical distance from an aircraft to the ground below by transmitting a short sharp burst or pulse of radio waves. The radio altimeter is, however, more common. It transmits continuous waves and utilizes a device known as a phasemeter, which measures the extent to which the pattern received from ground reflections diverges from, or is out of phase with, the pattern transmitted by the antenna. Naturally, the height of which a radio altimeter can be used is limited to that at which the reflected patterns are in phase with the transmitted patterns. Within this limit, the radio altimeter is accurate and is used as an aid in landing aircraft automatically.

The radio altimeter

Distance measurement equipment (DME) uses pulses that are transmitted from the aircraft, received by a DME beacon on the ground, and transmitted back to the craft. The time taken for the pulse to make the round trip is measured by a very precise electronic clock; distance is displayed automatically on counters similar to the odometer in an automobile. DME beacons are located with VOR beacons to provide simultaneous bearing and distance information to an approaching aircraft, thus giving its position. DME is not used by ships because its range is not adequate.

Hyperbolic systems. If the two guns M and S in Figure 5 are fired simultaneously, each consecutive ring in the figure will represent a distance of 1,100 feet (340 metres) travelled by the sound in one second. Thus a craft anywhere on AB, the perpendicular bisector of MS, would

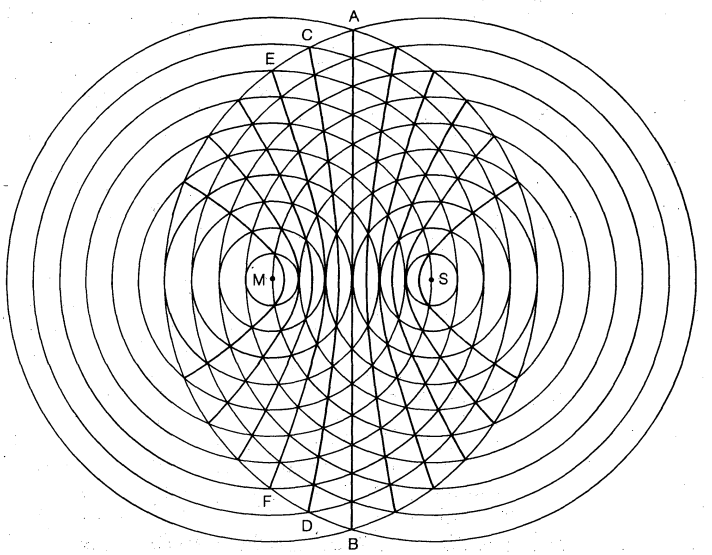


Figure 5: Hyperbolic position lines (see text).

hear both guns simultaneously. A craft 1,100 feet closer to the gun M will hear the gun S one second later and will lie on the hyperbola CD. A craft 2,200 feet closer to gun M will hear gun S two seconds later and will lie on the hyperbola EF. Hence, by timing the interval between hearing gun M and gun S, it is possible to find on which hyperbola the observer is located if the interval were an exact second. Other hyperbolas drawn proportionally represent intervals of fractions of a second.

In navigation the firing of guns is replaced by radio transmissions. The sheaf (family) of hyperbolas as shown in Figure 5 may be printed on a chart. A second sheaf of hyperbolas, from a second pair of stations, can be printed on the same chart; two position lines are plotted and a fix obtained. In radio systems, one of the two stations in a pair (the master) controls the other (the slave). In some systems, the stations are arranged in "chains" with two or three slaves distributed about the master and two or three sheafs of hyperbolas printed on the appropriate chart.

Loran, long range navigation, is a radio-navigation system that permits a ship to locate its position accurately by timing the arrival of pulses from synchronized shore transmitters. Either a master transmitter and two slaves keyed to it are necessary or two synchronized masters, each with a keyed slave. The difference in time of the arrival of pulses from a master and slave shows that the ship is somewhere on a curved line (hyperbola) running between the two and curving back around the nearer. The second pair of transmitters (or second slave) is necessary

The Loran system

to supply another, intersecting curve; the point of intersection is the ship's location.

In Loran systems, radio waves that are reflected from the ionosphere cannot be used. Such reflected waves, or sky waves, can cause confusion, arriving after the correct pulse and sometimes being much stronger.

Decca and other hyperbolic systems. Decca is a hyperbolic system similar to Loran but it operates at a longer wavelength and uses continuous transmission instead of pulses. The difference in time of arrival of the waves is determined by the difference in phase between the incoming signals; otherwise operation is similar to Loran. This arrangement provides a remarkably accurate and reliable system covering a range of 100–300 miles (160–480 kilometres) from the master station and widely used by ships and aircraft particularly for navigation close to the destination.

Omega is a new long wave hyperbolic system which probably has the longest range of all radio navigational aids. Like Decca, it is a hyperbolic system that uses continuous waves and, like Decca, it has a system of special transmissions for lane identification.

During World War II, Germany produced the Sonne beacon that has been developed into the Consol and Consolan systems, in which the masters and slaves are so close together that the hyperbolas become practically straight lines radiating out from the centre. By changing the phase of transmissions of the slaves, the hyperbolas can be made to swing to and fro. Morse signals are superimposed so that a craft can tell its location in a particular lane or sector.

The Decca system adapted for computer. Decca covers a wide area and is inherently a flexible navigation system. Before Decca's invention, however, aviation policy had been committed to signpost navigation using VOR beacons and subsequently VOR/DME. Consequently, the Decca system was adapted to work with a computer that would make the area system work as if there were beacons placed anywhere the navigator wished. Thus Decca could work as a signpost or "point" navigation system.

The "point" navigation philosophy has since run into difficulties. As air traffic builds up, more and more routes are needed; thus more and more signposts must be set up. VOR/DME computers, however, have been developed so an aircraft can follow any track irrespective of the position of the VOR/DME. Thus the computer has restored to the "point" navigation system the flexibility originally possessed only by the area navigation system, a flexibility increasingly needed to ensure adequate separation between aircraft.

The computer has already become part of the inertial navigator; it is used also in automatic dead reckoning. Originally, computers used in navigation systems were analogue, and calculations of a relatively simple nature involving inputs from various electrical sources were continuously performed. Today's trend is toward the use of digital computers for performing the necessary calculations.

Use of the digital computer

The digital computer works so fast that for navigational purposes it can be considered virtually instantaneous and can therefore provide continuous information for control purposes. It has a memory to store information for use when needed. It is built from electronic modules that are mass-produced at low cost. It has only one disadvantage. Conversion of analogue information into digital form is costly. Hence, although far superior to the analogue computer, it is less economical whenever a large number of electrical signals must be combined in relatively simple ways without any need for memory. Such situations still apply to control systems in many craft.

Navigation satellites. Satellites may be designed to function as navigational aids. It is possible to place a satellite in a stationary orbit, in which it circles the earth at an altitude of about 19,300 nautical miles. It then circles the earth once a day and, if positioned above the Equator to revolve in the same direction as the earth rotates, appears to be stationary in the sky. Stationary satellites can receive signals from various craft and retransmit these signals to the ground. They may thus act as relays to help measure the position of a craft by distance-measuring systems or by hyperbolic methods; they can then retransmit the position back to the craft.

Satellites that travel in orbits a few hundred miles above the surface of the earth circle it in a matter of an hour or two, with 80 minutes a typical period. Such satellites may be placed in orbits at an angle across the earth so that they cover a wide band of latitudes, moving a little further to the west at each orbit as viewed from the earth below. Such satellites may be used, as in the Transit system, to find the position of the craft by Doppler. The satellite transmits a steady radio frequency and a receiver in the craft detects the change of frequency. The speed at which the note changes establishes how far away the satellite is, just as one can detect how close a racing car is by the abruptness with which the engine note drops from a whine to a growl as it passes. The satellite also transmits continual details of its orbit so the craft can calculate its position at the moment the satellite passes by. Though the calculation may be made by arithmetic, a computer is often employed. Obviously a fix is obtainable only once every 80 minutes or so and, when the satellite passes almost directly overhead, it may not be possible to decide whether it has passed to the left or to the right of the craft. It may be necessary to wait another 80 minutes to obtain a useful fix.

RADAR

Operation. Radio waves with wavelengths in the centimetre range can be beamed by a reflector like light in an automobile head lamp to make up a radar system. The narrowness of the beam depends on the shortness of the waves and on the width of the reflector. For ships and aircraft, radio waves of a very few centimetres in length are commonly used to keep the reflector width down to a few feet. Ground radars can have much bigger reflectors and wavelengths of ten centimetres or more are common. A radar antenna mounted on a ship is wide and shallow to produce a beam that is narrow in the horizontal plane and wide in the vertical plane. Narrowness in the vertical plane could cause the radio waves to miss the target when the ship rolls. As the radar antenna rotates, the transmitter sends out a series of very short pulses every degree or so. When the pulse strikes an object, it is reflected back to the radar aerial and hence passed to the radar receiver. The pulse is displayed on a cathode-ray tube. Electronic lines are drawn from the centre outward, each starting out as the pulse starts out from the transmitter. When an echo returns, the image on the tube brightens up. Thus a spot of light appears on the cathode-ray tube at a distance from the centre proportional to the time the pulse takes to go out and back and in a direction the same as that in which the pulse was transmitted. Hence, on the cathode-ray tube a faint ray of light rotates around the screen like a searchlight, following the rotation of the antenna, and paints in the positions of any reflecting objects as if they were on a map. The face of the cathode-ray tube is coated with long-persistence phosphor that continues to glow for several seconds after it is excited, thus allowing the viewer time to study and analyze the image.

The strength of the return signals will vary depending on the reflectivity of the surface of the reflecting object and on its distance. There will be little reflection from water unless it is very rough. From cliffs, ships at sea, and buildings, there will be strong reflections due to the vertical surfaces. From the ground, there will be only scattered reflections, generally stronger in wooded country. Nevertheless, because the shortest radar waves are so much longer than light waves, the picture painted by radar is imperfect and requires careful interpretation.

Applications. Although radar can produce a map, its main use is in avoiding collision. Using radar, the navigator of a ship can see other vessels irrespective of light conditions. He can therefore proceed slowly even in thick fog, whereas without radar he would have to heave to or anchor. Since the radar picture is imperfect, collision avoidance requires great skill in interpretation. In particular, it is not easy to see the direction in which other vessels are travelling. Because of this, the seaman has to

Use of radar to avoid collision

make particularly bold alterations of course in order that other ships may know his intentions.

With radar, an air-traffic controller can watch the progress of aircraft in his area. Special height-finding radars, whose aerials are tall and narrow instead of broad and shallow, are also used. Radar is also used to watch aircraft approach and land, one radar following the aircraft in the vertical plane and the other following it in the horizontal plane. If necessary, the aircraft can be "talked down" (told exactly how to land) by the radar operator on the ground.

During World War II, craft carried radar responders that gave out coded signals when hit by radar beams, permitting identification of the craft as friend or foe. Similar systems are being developed for civil aviation in which the aircraft automatically identifies itself and gives its altitude when struck by an Air Traffic Control radar beam.

Special systems. For military purposes, infrared waves may be used. Midway in frequency between radio and light waves, these are, in fact, heat waves. An infrared detector can distinguish between objects of different temperatures and can identify a ship on the water or a man hidden in undergrowth. Infrared needs only a very small antenna and therefore has been used to guide small missiles launched from aircraft to strike at enemy planes. Infrared, however, is distracted by heat, from the sun and from friendly jet engines.

Lasers (*q.v.*) can also be adapted to produce radar-like devices of extreme precision. In addition, there is increased interest in simple television cameras; when equipped with light intensifiers, they can "see" in almost pitch darkness. Finally, for underwater detection of either submarines or shoals of fish, sonar radars are being developed.

Automation. The computer has already been mentioned in connection with point and area navigation systems, the use of satellites, and the mixing of various aids with inertial navigators. Automation is needed because craft speeds are increasing to meet demands for quick delivery of passengers and cargo and to lower operating costs. A jet aircraft with a crew of fewer than ten can carry more passengers across the Atlantic than an ocean liner with a crew of hundreds, because its greater speed enables it to make many more trips a year.

Automation can also reduce crew sizes by taking over tasks that machines can do more reliably and more consistently. On the other hand, because engineers must design machines years before they are used, they cannot foresee some situations that arise. The first landing of a man on the moon involved events not anticipated by the designers of the landing module. Had men not been in the spacecraft, it would have crashed. Therefore the human navigator will doubtless be required in many types of craft well into the foreseeable future. Although the machine can manipulate, it cannot manage except insofar as it is preprogrammed to do so.

The navigator of the future will become more and more a manager of systems. He may also be required to undertake certain manipulative tasks simply because under certain circumstances it is less costly and equally efficient to use a human.

BIBLIOGRAPHY. Good texts on general navigation include: E.W. ANDERSON, *The Principles of Navigation* (1966); N. BOWDITCH, *American Practical Navigator*, U.S. Hydrographic Office, rev. ed. (1962); W. LEY, *Rockets, Missiles and Men in Space* (1968); *Navigation Dictionary*, U.S. Hydrographic Office (1956); A.E. PUCKETT and S. RAMO (eds.), *Guided Missile Engineering* (1958); and P.V.H. WEEMS and C.V. LEE, *Marine Navigation*, 2nd ed. (1958).

Works on instinct and art include: M. BLEWITT, *Celestial Navigation for Yachtsmen*, 4th ed. (1964, reprinted 1967); C.H. COTTER, *The Principles and Practice of Radio Direction Finding* (1961); W. DENNE, *Magnetic Compass Deviation and Correction* (1951); C.H. DEETZ and O.S. ADAMS, *Elements of Map Projection*, U.S. Coast and Geodetic Survey, 5th rev. ed. (1945); H. GATTY, *Nature is Your Guide* (1958); D. GREENHOOD, *Mapping* (1964); H.L. HITCHINS and W.E. MAY, *From Lodestone to Gyro Compass* (1952); G.V.T. MATTHEWS, *Bird Navigation*, 2nd ed. (1968); and L.B. OLSON, *Olson's Small Boat Seamenship* (1956).

Works treating control factors, including driving, steering, and speed are: R.A.B. ARDLEY, *Harbour Pilotage and the Handling and Mooring of Ships* (1944); W.R. ASHBY, *Introduction to Cybernetics*, rev. ed. (1969); C.S.W. DRAPER, W. WRIGLEY, and J. HOVORKA, *Inertial Guidance* (1960); E.R. KING and J.V. NOEL, *Ship Handling* (1954); A. PORTER, *Introduction to Servo Mechanisms* (1950); C.J. SAVANT, *Control Systems Design*, 2nd ed. (1964); VAN VALKENBURGH, NOOGER AND NEVILLE INCORPORATED, *Basic Synchros and Servo mechanisms*, 2 vol. (1955); and N. WIENER, *Cybernetics*, 2nd ed. (1961).

Works on guidance, including position finding and radar are: L. BRILLOUIN, *Science and Information Theory*, 2nd ed. (1962); R.F. HANSFORD (ed.), *Radio Aids to Civil Aviation* (1960); J.W. HORTON, *Fundamentals of Sonar*, U.S. Naval Institute (1959); INTERNATIONAL HYDROGRAPHIC BUREAU, *Radio Aids to Marine Navigation and Hydrography* (1956); E.G. BOWEN (ed.), *A Textbook of Radar*, 2nd ed. (1954); L.N. RIDENOUR (ed.), *Radar System Engineering* (1965); U.S. DEPT. OF COMMERCE, MARITIME ADMINISTRATION, *Radar Instruction Handbook* (1969); and F.J. WYLIE (ed.), *The Use of Radar at Sea*, 4th rev. ed. (1968).

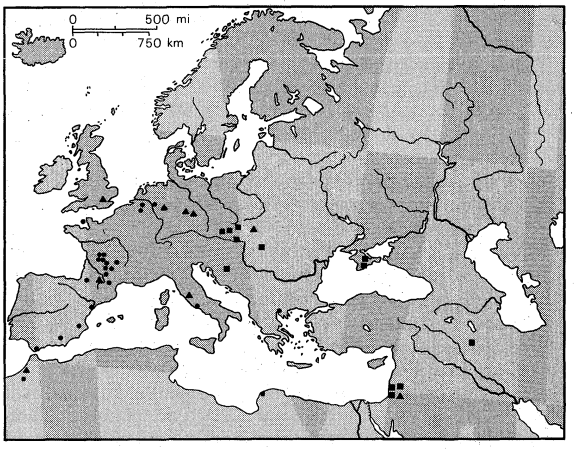
(E.W.A.)

Neanderthal Man

Human populations referred to as Neanderthal man inhabited much of Europe and the lands surrounding the Mediterranean during the extended time interval of the earlier upper Pleistocene. The name derives from the gorgelike Neander Valley above the small stream of Düssel, a tributary of the Rhine, seven miles east of Düsseldorf, Germany. There in the Feldhofer Cave in 1856 workers encountered portions of a human skeleton; 14 pieces were eventually salvaged by Johann Carl Fuhlrott, a professor at the Realschule in Elberfeld. Following the description of the remains (by German anthropologist Hermann Schaaffhausen), a lively controversy arose as to whether their unusual morphology was normal, and hence indicative of an archaic variety of man, or whether it was merely a case of pathology in modern man. The former view was duly proved correct after the discovery in 1886 of two similar skeletons in the Cave of Spy (Belgium) in association with implements of chipped stone and animal bones representative of a subarctic fauna including extinct species. In over a century of research on human evolution and the prehistory of man, much additional information has been gained concerning the Neanderthal peoples, their distribution in space and time, their origin and disappearance, and their ways of life.

Recognition of Neanderthal man

Adapted from E. Vlcek, *Neandertaler der Tschechoslowakei* (1969); Czechoslovak Academy of Sciences and Hermann Bohlaus Nachf., Vienna



▲ Hominid remains of Riss Glacial Age and of Last Interglacial (Eemian) Age
• Western European Neanderthals of Last Glacial Age
■ East European and Southwest Asian Neanderthals, some with affinity to anatomically modern man

Figure 1: Distribution of pre-Neanderthal and Neanderthal finds in Europe, southwestern Asia, and northern Africa.

Ancestors of Neanderthal man. The forerunners of Neanderthal peoples are hardly known. Perhaps they are to be found among human populations in Europe dating back to the Riss Glacial Stage, 100,000 to 150,000 years ago. To date, however, only a half-dozen localities in France have yielded relevant fragmentary skull parts of that geologic age. These specimens diverge in dental and

skull morphology from an older human species, *Homo erectus*, and to the degree that they do so they approach more closely the structure characteristic of the modern human species, *Homo sapiens*. The skeletal parts available are too limited to permit an interpretation of their shape and structure as especially Neanderthal-like. Better preserved and more complete human crania of older (Great Interglacial) age are not particularly Neanderthalian, either. Some students of human evolution reasonably regard them as early representatives of *Homo sapiens*. The time and the place of Neanderthal origins, thus, still remains unknown.

Early Neanderthals

Neanderthals of the Third Interglacial Period. The first recognizably Neanderthal peoples occur during the Last Interglacial Age in Europe. Their jaws, teeth, and skull parts are known from some half-dozen sites in Czechoslovakia, Germany, France, and Italy. Among other features, they were characterized by moderately small cranial capacity; a short, narrow, and well-arched skull vault; large but somewhat separate and curved supraorbital ridges; moderately large cheekbones with some hollowing beneath them, and gracile and straight or only slightly curved limb bones, generally lacking enlarged ends. The general pattern of the anatomy of the face and even of the skull vault of these early Neanderthal peoples was not greatly unlike that of modern *Homo sapiens;* this was also true of the bones of the limbs, although these are still hardly known. There was nonetheless a complex of skeletal features that distinguished them from modern human populations.

Neanderthals of the Fourth Glacial Period. Far more numerous and complete skeletal remains have been recovered of Neanderthal peoples from the succeeding initial phases of the Last (Würm) Glacial Stage (between about 35,000 and 85,000 years ago). This is probably due in large part to the marked tendency toward utilization of cave fronts and rock shelters below cliffs for habitation sites, as well as the widespread practice of deliberate burial of the dead. Such peoples were distributed, evidently as small semi-isolated populations of hunting and gathering bands with a Mousterian stone industry, from western Asia throughout much of Europe, except the far north, as well as along the Mediterranean littoral of northern Africa.

The cultural adjustments of these extinct people can be inferred, and then only very imperfectly, from careful, well-controlled excavations of their habitation places. Hundreds of such occurrences are known, mostly cave fronts and rock shelters, but sometimes also open-air encampments. Many have been dug in some way, but relatively few have received really thorough scientific study. Some sites are quite small and were only visited by prehistoric men, probably briefly, on one or a few occasions. Other sites are of quite substantial size, and evince repeated episodes of human occupation, sometimes nearly continuously, sometimes interrupted by intervals of abandonment. The magnitude and intensity of occupation can be deduced from the extent, thickness, and overall quantity of occupational residues. In some instances sites were not merely occupied but were modified through stone emplacements, cobble pavings, and excavated depressions and holes; some of these were doubtless related to shelter-like structures. Persistent and extensive employment of fire is evidenced by reddened areas, hearth traces, charcoal and ashes, and even heat-fractured stone. Animal bone, often in abundance, testifies to efficiency and success in hunting and probably also in snaring and trapping. There is frequently some diversity in the representation of herbivorous and carnivorous species; but there may also be a predominance of one or several species—wild horse, reindeer, aurochs, bison, fallow deer or gazelle, red deer, goat or ibex, for example. These differences evidently reflect a combination of ancient environmental conditions, local availability of particular species, and dietary preferences. Remains of birds, even fish, are also known and in infrequent instances also invertebrates (*e.g.*, clams). The nature and extent of plant food utilization is scarcely known because of the very poor condition of their preservation; however, some traces may be preserved in charcoal, seeds, and pollen. Stones appropriate to the cracking and grinding of plant seeds have also been found at these sites.

The technological capabilities of Neanderthal peoples are best inferred from the products of their stoneworking; although bone was also fashioned into tools and weapons. Less durable materials, such as wood or cordage, are scarcely ever preserved; only a single wooden lance, apparently employed to kill a straight-tusked elephant, has been recovered (in Germany) from this time range. The stone technology reveals special procedures for the preparation of blanks (rough pieces of wood) for the eventual fashioning of chipped stone tools and other artifacts. The stone tool kit is quite diversified and there is a substantial number of "standardized" tool forms or types and subtypes. These are frequently found to occur in different sets or combinations, and some archaeologists consider that such recurring clusters of tool types and subtypes may reflect locally distinctive styles of life.

Stone-working technology

Deliberate burials are not uncommon in Neanderthal occupation sites. Both individual and group burials are known, which have yielded remains of adults and juveniles. Their emplacement, and accompanying artifactual materials, evidently reflect some concern for the deceased. In other instances human bones are incomplete and seem to have been so deliberately broken as to indicate cannibalistic practices. Some unusual ritual-like arrangements of animal remains, including bear and goat, are also known, and these are sometimes associated with human skeletons. In general, however, little can be confidently inferred about Neanderthal beliefs and rituals. Artistic or decorative proclivities are reflected in the frequent occurrence at some sites of natural colorants (manganese oxides, ochre) with evidences of their use, and more rarely of cut and incised bone.

The most numerous remains of these peoples, sometimes termed classic Neanderthals (Figure 2), have been recovered from the western and southwestern parts of Europe. Over this area, including Belgium, France, westernmost Germany, Spain, and Italy, lived short, stout, powerfully built Neanderthal peoples possessing a number of distinctive skeletal characteristics that have been thought to set them apart from their contemporaries elsewhere. This was true of the braincase (very large cranial capacity; long, low, and wide skull vault, flattened behind; heavy, bi-arched supraorbital ridges), of the face (projecting and large; round orbits; small cheekbones with no hollowing below them due to expanded sinuses; quite large teeth and palate), of the trunk (rounded, broad chest; short vertebrae with large muscular processes; long, slender collarbone), and of the limbs (heavy, curved thigh and forearm bones; large feet and hands but short fingers and toes; arm and leg bones with enlarged ends; long heel bone). The popular conception that these people were slouched in posture and walked with a shuffling, bent-knee gait seems to have been due in large part to faulty reconstruction of the skull base and to

By courtesy of the Musee de l'Homme, Paris

Figure 2: Side view of Neanderthal skull of classic type, found at La Ferrassie, France.

Physical characteristics of Neanderthals

misinterpretation of certain features of the limb bones of one of the Neanderthal skeletons discovered early in the 20th century.

Such were the peoples to whom the specific name *Homo neanderthalensis* was applied in 1864 after the original discovery of the type skeleton in the Feldhofer Cave. (Actually two other such skulls had been found earlier in the 19th century but went unrecognized until after the Neanderthal discovery proper.) It is now apparent that these peoples represented merely the western periphery of a particular range of variation within a quite widespread human species. Since such peoples were in fact fundamentally like *Homo sapiens*, the use of *neanderthalensis* as a distinct species name can be most misleading. It is more reasonable to consider such people as no more than an extinct subspecies: *Homo sapiens neanderthalensis*.

Skeletal remains of other Neanderthal peoples are known from central Europe (Czechoslovakia, Hungary), southeastern Europe (Yugoslavia, Greece), Crimean Russia, southwestern Asia (Israel, Iraq), and from as far east as central Asia (Uzbekistan). Others are known from cave sites in Cyrenaica and Morocco. The skeletal anatomy of individuals from eastern Europe and western Asia is unquestionably Neanderthalian; but there are also divergences in morphology, particularly in those from western Asia (see Figure 3), from the characteristic European condition. This seeming progression in the variation of morphological features between geographic populations increasingly more distant one from another may merely reflect a gradient (cline), as often is observed within species.

By courtesy of the Peabody Museum of Harvard University

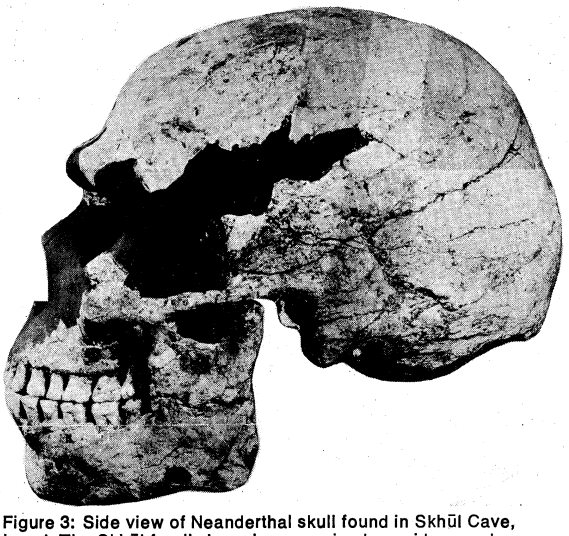


Figure 3: Side view of Neanderthal skull found in Skhūl Cave, Israel. The Skhūl fossils have less massive brow ridges and shorter, less massive crania than the classic Neanderthal type.

Causes of morphological distinctiveness. Populations referred to as classic Neanderthals have sometimes been thought to have been relatively isolated from other such peoples in the southwesterly parts of Europe as a consequence of the increasing severity of the climate during the initial phases of the Last Glacial Age. Such isolation could have restricted gene flow between populations; increased inbreeding; and perhaps tended to bring about shifts, even drift, in gene frequencies, because of accidents of sampling from the original population. Such changes would encourage a progressive increase of distinctiveness in the morphology of the isolated Neanderthal populations from others elsewhere, not so restricted in breeding possibilities.

Increased selection pressure under increasingly harsh environmental stresses could also have been effective in bringing about morphological change and also, perhaps, in reducing population size below a critical point beyond which eventual partial or total extinction could have occurred. This simple model is inadequate to account for the progressive variation now evident among Neanderthal populations. It fails to consider the several interstadial phases of climatic amelioration both prior and subsequent to the principal prolonged cold of that period. More extensive and thorough archaeological investigations in recent years have also revealed far more elaborate and sophisticated cultural adjustments of Neanderthal peoples than were known previously.

Neanderthal man and Cro-Magnon man. The factors responsible for the disappearance of the Neanderthal peoples are an important problem to which there is unfortunately still no clear solution. In fact quite different factors may have operated in different parts of their range of distribution. In Europe nearly all cave and rock-shelter sites where occupation was repeated reveal a sterile horizon between the last Mousterian or related cultures (presumably of Neanderthal peoples) and succeeding occupations of early upper Paleolithic type (either Chatelperronian or early Aurignacian). The latter occupations appear always to be associated with Cro-Magnon peoples, a modern variety of *Homo sapiens* that is characteristically European in skeletal anatomy (see CRO-MAGNON MAN). In western Europe these earliest upper Paleolithic occupations date to *c.* 32,000 BC. In central Europe (Poland, Hungary), however, related cultural manifestations appear to have been present locally by *c.* 38,000 BC (and this seems also to have been the case in southwestern Asia and in Cyrenaican Libya). The youngest radiometrically dated Mousterian occupation in southwestern France is *c.* 33,000 BC, but even later occurrences are inferred on geological grounds in southeastern France. A climatic amelioration (Hengelo Interstadial), resulting in warmer temperatures and substantial reforestation, occurred between *c.* 37,000 and 35,000 BC, when these changes in material culture and in population were taking place. There are still, however, many ambiguities in the results of such radiometric dating as well as in the enumeration and age assessment of episodes of climatic amelioration in Europe.

Disappearance of Neanderthal man

It is clear, at least in Europe, that several populations were involved and that the Cro-Magnons succeeded the Neanderthals; but it still cannot be demonstrated conclusively whether these different human populations overlapped in time in the same territory, the indigenous Neanderthals being perhaps ultimately eliminated or swamped by the immigrant anatomically modern peoples, or whether the Neanderthals were generally no longer present, except perhaps in small local groups.

Asian and African peoples with Mousterian culture. In some cave sites in southwestern Asia and in northwestern Africa there are human skeletal remains that verge on an anatomically modern morphology, yet the cultural association is Mousterian. The cave of as-Skhūl on Mount Carmel (Israel) and Jebel Irhoud (Morocco) are the principal examples. Most recently, the cave of Jabal Qafzah, in Galilee, has afforded a substantial series of burials of individuals still more modern in appearance, again in a Mousterian cultural context. Some workers believe that such populations, who were perhaps already coexisting with and were eventually to replace Neanderthals in southwestern Asia during these initial phases of the Last Glacial, indeed represented the forerunners of the Cro-Magnon populations who later replaced the Neanderthals of Europe.

Non-Mousterian contemporaneous populations. Specifically Neanderthal peoples are not known to have existed in either Africa south of the Sahara or in eastern or southeastern Asia. Broadly contemporaneous peoples in those areas, however, have been sometimes referred to as Neanderthaloid, since a number of features of their skull morphology simulate the Neanderthal condition. Human skeletal remains that approach the morphology of modern *Homo sapiens* in significant respects are now known from several localities in eastern Africa from a range of time that coincided with that of the Neanderthals of more temperate latitudes.

BIBLIOGRAPHY. F. BORDES, "Mousterian Cultures in France," *Science*, 134:803–810 (1961), a general discussion of the nature of Mousterian stone technology, and the particulars of the Mousterian and its variations in France; C.L. BRACE, "Refocusing on the Neanderthal Problem," *Am.*

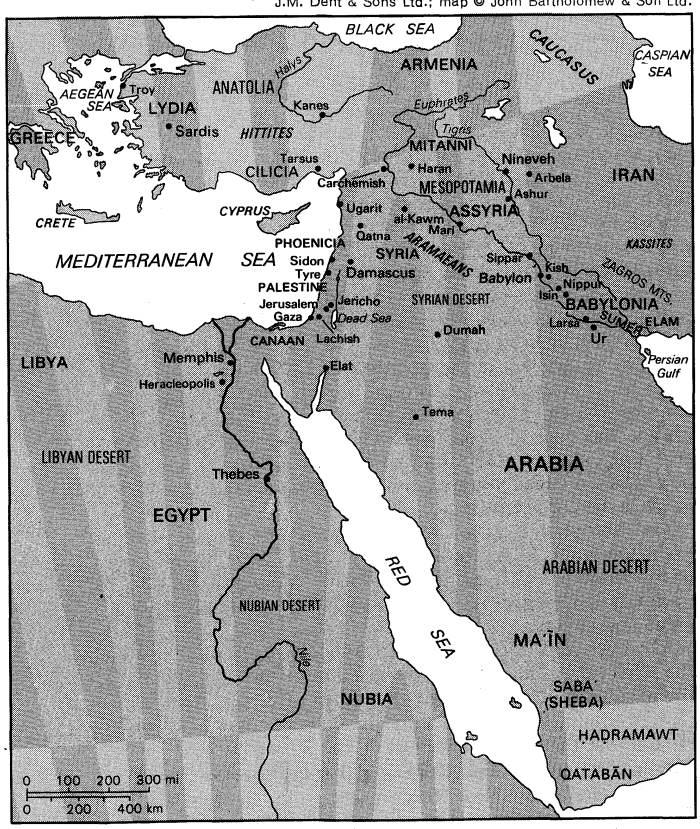
Anthrop., 64:729–741 (1962), a critical re-evaluation of various viewpoints and approaches toward the study of Neanderthal skeletal remains and their evolutionary significance; M.H. DAY, *Guide to Fossil Man* (1965), a general, illustrated handbook of the skeletal remains of the various fossil Hominidae; F.C. HOWELL, "The Evolutionary Significance of Variation and Varieties of Neanderthal Man," *Q. Rev. Biol.*, 32:330–347 (1957), a critical appraisal of the temporal and spatial variation of Neanderthal populations, their relationships, and evolutionary significance; E. PATTE, *Les Neanderthaliens: anatomie, physiologie, comparaisons* (1955), a detailed comparative study of the skeletal anatomy of Neanderthal peoples; J. PIVETEAU and F. BORDES (eds.), *L'origine de l'homme moderne* (1970), a series of major papers on the Neanderthal problem, viewed biologically, culturally, and paleoenvironmentally; E. VLCEK, *Neandertaler der Tschechoslowakei* (1969), a detailed study of Neanderthal skeletal remains from Czechoslovakia with comparative observations on Neanderthals from elsewhere in Europe and the Middle East.

(F.C.H.)

Near East, Ancient

In this article, the "Near East" (approximately equivalent to the "Middle East") is defined as the area of Egypt, Nubia, Palestine and Syria, Armenia, Asia Minor or Anatolia, and Mesopotamia, together with the outlying areas of North Africa, the Aegean, Iran or Persia, and the Indus Valley. Since many of these countries are referred to in biblical history, they are often called "Bible Lands." The term Fertile Crescent is often applied to Mesopotamia and Syria–Palestine. The article will deal with the period between the emergence of agriculture about 10,000 years ago and the conquest of virtually the whole ancient Near East by Alexander the Great in the 4th century BC. Iran and South Arabia, which continued to follow basically pre-Hellenistic cultural patterns until their conquest by the Islāmic Arabs, will be surveyed until the 7th century AD.

Adapted from J.O. Thompson, *Everyman's Historical Atlas* (1963), J.M. Dent & Sons Ltd.; map © John Bartholomew & Son Ltd.



The ancient Near East.

Most scholars agree that the first sedentary (non-nomadic) human societies developed in the Near East and that the sedentary and urban way of life spread from the Near East to other parts of Asia and Africa and to Europe. Carbon-14 (radioactive carbon) and other methods of dating by the physical sciences have established with remarkable precision the general chronological framework of late prehistory and early history in the Near East and surrounding areas.

The emergence of sedentary cultures (*c.* 9000–*c.* 3000 BC). The high antiquity of civilization in the near East is largely due to the existence of convenient land bridges and easy sea lanes passable in summer or winter, in dry or wet seasons. Movement of large numbers of people north of the Caspian Sea was virtually impossible in winter, owing to the severity of the climate; central Eurasia was often too dry in summer. Land passage between Asia and Africa was in early times limited to narrow strips of land in the Isthmus of Suez. Large-scale desert travel was limited to special routes in Iran and in North Africa, both east and west of the Nile Valley.

Origins of agriculture

Another reason for the early significance of this area in world history is the fact that the water supply and the climate were ideal for the introduction of agriculture. Several species of grain grew wild, and there were marshes and tributary streams that could easily be drained or dammed in order to sow wild wheat and barley. The seed had only to be strewn over a sufficiently moist surface to ensure some kind of crop under normal conditions. It is therefore not surprising that there is evidence of simple agriculture as far back as the 8th or 9th millennium BC, especially in Palestine, where more excavating has been done in early sites than in any other country of the Near East. Many bone sickle handles and flint sickle edges dating from between *c.* 9000 and 7000 BC have been found in Palestinian sites.

Early irrigation cultures. In Mesopotamia and Iran remains of this period appear in caves on the lower slopes of the Zagros Mountains between western Iran and Iraq. The date of the systematic introduction of irrigation on a large scale in Mesopotamia is somewhat doubtful because most of the early sites of irrigation culture were covered long ago by accumulation of alluvial soil brought down by the spring floods of the Tigris and Euphrates rivers. Archaeologists once thought that all irrigation originated in the foothills of the Zagros and that the earliest true farmers lived in the plains of Iran. But recent excavations and surface explorations have proved that irrigation around the upper Tigris and Euphrates, as well as their tributaries, dates from the early 6th millennium BC (*e.g.*, at el-Kowm on the Upper Euphrates). Small-scale irrigation was practiced in Palestine (*e.g.*, at Jericho) in the 7th millennium BC.

In northern and eastern Mesopotamia, main streams were soon partly diverted during moderate river floods into canals running more or less parallel to the rivers, which could thus be used to irrigate an extensive area. Such deflector dam irrigation avoided the self-destructive weaknesses of large storage dams, in particular the danger of depositing great masses of refractory mud in the storage basin behind the dam. In the north and east considerable urban installations developed at sites such as Nineveh no later than the 5th millennium BC, when southern Mesopotamia was still mostly swampland like the early Egyptian delta. The Euphrates had a much smaller flow of water than the nearby Tigris. The latter was much swifter, however, so that it was potentially more important for irrigation, even though much harder to tame.

The Egyptian Nile had a much more predictable water flow than the Mesopotamian rivers because it flowed through hundreds of miles of swamp, where unusually high annual floods spread out, interfering with navigation but averting the danger of the occasional destructive inundations of Mesopotamia.

Mesopotamia and Egypt to c. 1600 BC. The oldest known urban and literate culture in the world was developed by the Sumerians in Mesopotamia beginning in the late 4th millennium BC. About 2300 BC a Semitic leader, Sargon I, conquered all of Babylonia and founded the first dynasty of Akkad (Akkadu), which held power for about a century and a half. Sargon and his successors

The Sumerians

were the first known rulers in southwest Asia to gain control of the Fertile Crescent as well as of adjacent territories. They sent trading expeditions to central Anatolia, Iran, and as far as India and Egypt. After the fall of the dynasty of Akkad there was a Sumerian revival under the 3rd dynasty of Ur (Ur III: [21st–20th centuries]), followed by another influx of Semites. These people founded the first dynasty of Babylon (19th–16th centuries), whose most important king was Hammurabi. In the 17th century new ethnic groups appeared in both Babylonia and Syria–Palestine: Kassites from the Zagros Mountains, Hurrians from what is now Armenia, and Indo-Europeans from Central Asia. This period marked the end of the formative phase of Mesopotamian civilization.

Shortly after 3000 BC the numerous small states that had arisen in the Nile Valley during the 4th millennium were united under the 1st dynasty of Egypt. At this time the Egyptians had already developed a system of writing. Between *c.* 2686 and *c.* 2160 BC, their country was united under a powerful monarchy (the Old Kingdom) served by a complex bureaucracy.

Toward the end of the 3rd millennium there was a period of disunity, followed by reunification under the 12th dynasty (1991–1786).

During these two centuries Egyptian control was established over Nubia, Libya, Palestine, and southern Syria. Soon after 1800 BC the Egyptian Empire fell apart, and *c.* 1700 Egypt was overwhelmed by the Asian "Hyksos," who ruled the country for a century and a half.

New states and peoples. Before the close of the 16th century BC the native 18th dynasty rose in Egypt; it expelled the Hyksos and founded the New Kingdom. The New Kingdom rulers moved back into Syria–Palestine and came into conflict first with the Hurrian state of Mitanni and later with the Anatolian Hittites, who were expanding into Syria from the north in the 14th century BC. The Amarna Letters (diplomatic correspondence written in Babylonian script and language, and discovered in Egypt by archaeologists) are an important source of information on this period. In Mesopotamia the dominant powers were Kassite Babylonia and Assyria (which emerged from subjection to Mitanni in the early 14th century BC). Relations between states were governed by elaborate treaties, which were constantly being broken. After the fall of Mitanni (*c.* 1350) the Hittites and Babylonians both directed their hostility against Assyria. Kassite Babylonia was subjugated by Assyria *c.* 1230. This, followed by the fall of the Hittite Empire (*c.* 1200), ended what has been called the first "International Age" in the civilized world.

The invasions of the 13th century BC

The latter part of the 13th century BC saw the irruption of new peoples into the Aegean, Anatolia, and the Fertile Crescent; their appearance coincided with the Trojan War, the collapse of the Hittite Empire, and the destruction of many coastal cities of Greece, Cyprus, and Syria–Palestine. Best known of the new settlers from the west are the Phrygians, who occupied most of the old Hittite heartland, and the Philistines, who moved into Palestine.

At the same time, in Transjordan and western Palestine, the Hebrews founded a tribal confederation that was changed into a monarchy by Saul and David (*c.* 1020–960 BC).

In the east, the Iranian tribes, led by the Medes, were pouring into Iran from Turkistan. From the south and west came the Semitic Aramaeans. The Aramaeans and Medes were to transform the old face of the ancient Near East.

The Assyrian state suffered an eclipse in the 11th century BC, when the Aramaeans and related tribes occupied most of its territory. It was not until the late 10th century that the Assyrians began to recover, but by 850 they had conquered much of western Media and southern Armenia as well as Babylonia and Syria. In the following centuries, until just before 630, the empire was greatly expanded. It was also highly organized administratively; its language became Aramaic.

The Canaanite Phoenicians on the Syrian coast re-established their trading communities after the Philistine and Aramaean invasions; in the 10th and 9th centuries they moved out into the Mediterranean, establishing colonies in North Africa and as far west as Spain. Their influence in the western Mediterranean declined after the 6th century. Their Carthaginian colony then took over Phoenician trade in the western and central Mediterranean.

Farther east, the Medes and Chaldeans destroyed the Assyrian Empire at the end of the 7th century. The Chaldean dynasty in Babylonia carried on Assyrian traditions of administration and encouraged commerce; under Nebuchadrezzar II (604–562 BC), their Neo-Babylonian Empire became the most powerful political entity of its time. Its rule extended from the Taurus Mountains in Anatolia to eastern Arabia and deep into southern Iran. This short-lived state made a tremendous impression on contemporaries, especially on the Jews, whose state was destroyed and who were carried into Babylonian Captivity, and on the Greeks, to whom the glory of Babylon became legendary.

The Achaemenian Empire and its successors. In the 6th century the Iranian Persians under Cyrus the Great conquered their Median cousins and established the Achaemenian state (549). This was followed by the conquest of Lydia (546) and the Babylonian Empire (539). Aramaic became the official language of the Persian Empire, and its official religion was Zoroastrianism. Cyrus' enlightened policy put an end to the Assyro-Babylonian practice of deporting conquered peoples and trying to destroy all local nationalisms.

At its height the Achaemenian Empire ruled the whole of the Near East; Greek resistance prevented it from expanding successfully into Europe.

The Macedonian conquest

The Parthian and Sāsānian empires. In 334 BC Alexander of Macedon invaded Anatolia and nine years later completed the conquest of the Persian realm. His vast empire was broken up into Macedonian "successor states" after his death. The Seleucid kings of Syria controlled most of Anatolia, Mesopotamia, and Iran. About 250 BC the still seminomadic Parthians emerged from a small area southeast of the Caspian Sea. Establishing control over Iran, they declared their independence of the Seleucid Empire and in the 2nd century BC expanded westward into Mesopotamia. In the 3rd century AD the semi-Hellenized Parthians were replaced by the Persian Sāsānians. The Sāsānians ruled Iran from AD 224 to 642; they extended its boundaries, reinvigorated its administration and cultural life, and challenged Roman power in the Near East. In 636 the Sāsānian Empire was conquered by the Muslim Arabs, bringing to an end the last phase of ancient Near Eastern civilization.

Pre-Islāmic Arabia. Arabia was drawn into the orbit of western Asiatic civilization toward the end of the 3rd millennium BC; caravan trade between south Arabia and the Fertile Crescent began about the middle of the 2nd millennium BC. The domestication of the camel around the 12th century BC made desert travel easier and gave rise to a flourishing society in South Arabia, centred around the state of Saba (Sheba). In eastern Arabia the island of Dilmun (Bahrain) had become a thriving entrepôt between Mesopotamia, South Arabia, and India as early as the 24th century BC.

The discovery by the Mediterranean peoples of the monsoon winds in the Indian Ocean made possible flourishing Roman and Byzantine seaborne trade between the northern Red Sea ports and South Arabia, extending to India and beyond.

In the 5th and 6th centuries AD, successive invasions of the Christian Ethiopians and the counterintervention of the Sāsānian kings disrupted the states of South Arabia. The resulting economic decline made the rapid Muslim conquest of the area an easy task in the 7th century.

The importance of Near Eastern higher cultures for world history. *Religion.* Near Eastern religious thought had a strong influence on the ancient Greeks. The cosmogonies of Egypt, Babylonia, Phoenicia, and Anatolia were transmitted in part to the West and formed the basis of much of the cosmogonies of Hesiod and the Orphics before 600 BC, as well as the background for the cosmogonies of Thales and Anaximander in the 6th century BC.

There is some influence from the Near East in Pythagorean and Platonic thinking, but it is often hard to define and still harder to prove in detail. From the early 3rd century BC on, the Near East began to influence Greek thought increasingly. Babylonian astrology influenced Stoic philosophy, and some Jewish influence on Stoic ethics is likely as well.

Late Egyptian religious speculations were transmitted to the Greeks, especially through the "Physiologists" and Hermetic writers who flourished in Hellenistic Egypt beginning in the 2nd century BC. Astrology and alchemy were transmitted to our time in substantially the forms they received in Hellenistic and Roman Egypt.

Jewish influence

With the partial Hellenization of Judaism and its Christian offshoot in the 1st century AD, Jewish influence on the West rapidly became dominant. Most of it came through the Pharisees, but such Jewish sects as the Essenes and Baptists were directly involved.

Even such a heterodox Jewish sect as the Samaritans exerted disproportionately great influence. In the first years of the nascent Christian Church a Samaritan diviner named Simon, later called "Magus," founded a new faith known as Gnosticism. The Gnostics spread rapidly over both the Roman and Iranian worlds, and by the end of the 3rd century AD they had subdivided into a multitude of different sects that covered a wide spectrum of possible combinations between Judaism, Christianity, Zoroastrianism, and Greco-Roman paganism.

In four centuries Christianity conquered the entire Roman Empire and many outlying regions, thanks to the intensity of its faith and the tenacity with which Christians held to their views, following Jewish models, through the bitterest persecutions.

In the East, Zoroastrianism maintained its hold over the Iranians and neighbouring peoples until the Islāmic conquest, when it was replaced by Islām, itself an offshoot of the monotheistic Judeo-Christian stem.

With the Hebrew Bible (which became the Christian Old Testament) almost the whole of traditional Israelite and early Jewish religion passed into Christianity, which was to a great extent an extension of Judaism. What was lost by the surrender of a large part of Jewish ritual law, with its stress on purity, was compensated by the triumph of monotheism over Greco-Roman polytheism and the transformation of ethical and spiritual ideas in the Greco-Roman world. Judaism itself survived to coexist with Christianity and Islām through all the intervening centuries to our own day.

Science and technology, law and logic. During the 3,000 years of urbanized life in Mesopotamia and Egypt tremendous strides were made in various branches of science and technology. The greatest advances were made in Mesopotamia—very possibly because of its constant shift of population and openness to foreign influence, in contrast to the relative isolation of Egypt and the consequent stability of its population. The Egyptians excelled in such applied sciences as medicine, engineering, and surveying; in Mesopotamia greater progress was made in astronomy and mathematics. The development of astronomy seems to have been greatly accelerated by that of astrology, which took the lead among the quasi-sciences involved in divination. The Egyptians remained far behind the Babylonians in developing astronomy, while Babylonian medicine, because of its chiefly magical character, was less advanced than that of Egypt. In engineering and architecture Egyptians took an early lead, owing largely to the stress they laid on the construction of such elaborate monuments as vast pyramids and temples of granite and sandstone. On the other hand, the Babylonians led in the development of such practical arts as irrigation.

Mesopotamian astronomy and mathematics

Both sciences and pseudosciences spread from Egypt and Mesopotamia to Phoenicia and Anatolia. The Phoenicians in particular transmitted much of this knowledge to the various lands of the Mediterranean, especially to the Greeks. The direction taken by these influences can be followed from Egypt to Syria, Phoenicia, and Cyprus, thanks to a combination of excavated art forms that prove the direction of movement, as well as to Greek tradition, which lays great stress on what the early Greek philosophers learned from Egypt. Mesopotamian influence can be traced especially through the partial borrowing of Babylonian science and divination by the Hittites and later by the transmission of information through Phoenicia. The Egyptians and Mesopotamians wrote no theoretical treatises; information had to be transmitted piecemeal through personal contacts.

The westward transmission of Babylonian mathematics was associated with that of law. All early Babylonian mathematics was transmitted in case form, introduced by a condition followed by its solution. This model appears first in late Sumerian law: from law it was extended to scientific problems, and the form remained the same until well into the 1st millennium BC. This was true also of such derived systems as Hittite law and the so-called covenant law of Israel, as well as the earliest Greek codes (Draco and Gortyn), all of which are formulated similarly: condition (protasis), secondary condition or conditions, and conclusion (apodosis). The Babylonian Code of Hammurabi provides an example: "If a man accuses another man and brings a charge of murder against him but does not prove it, the accuser shall be executed."

Logical thought

With the spread of case law in such conditional formulation, it was eventually discovered that formulation as generalized propositions or prohibitions was a simpler and more logical means of setting up a coherent code of laws than the case-law formulation (*e.g.*, in the Hebrew Decalogue). The next step was the formulation of generalized geometrical propositions. This was first done by the Ionian Greek Thales (about 600 BC), whose listing of mathematical propositions in this generalized form instead of as conditional sentences was quite naturally described later as the "discovery" of mathematical theorems. With Thales logical reasoning made a giant step forward from the age of empirical modes of thought.

The Judeo-Christian concept of ethics and morals in law often prevailed in the Roman law of Christian times. Roman forms of law were ultimately adopted in almost the entire Western world, and through their universal sway many biblical approaches to legal problems became dominant.

The place of the alphabet in world history. Of all the accomplishments of the ancient Near East, the invention of the alphabet is probably the greatest. While pre-alphabetic systems of writing in the Old World became steadily more phonetic, they were still exceedingly cumbersome, and the syllabic systems that gradually replaced them remained complex and difficult. In the early Hyksos period (17th century BC) the Northwestern Semites living in Egypt adapted hieroglyphic characters—in at least two slightly differing forms of letters—to their own purposes. Thus was developed the earliest known purely consonantal alphabet, imitated in northern Syria, with the addition of two letters to designate vowels used with the glottal catch.

This alphabet spread rapidly and was in quite common use among the Northwestern Semites (Canaanites, Hebrews, Aramaeans, and especially the Phoenicians) soon after its invention. By the 9th century BC the Phoenicians were using it in the western Mediterranean, and the Greeks and Phrygians adopted it in the 8th. The alphabet contributed vastly to the Greek cultural and literary revolution in the immediately following period. From the Greeks it was transmitted to other Western peoples. Since language must always remain the chief mode of communication for *Homo sapiens*, its union with hearing and vision in a uniquely simple phonetic structure has probably revolutionized civilization more than any other invention in history.

BIBLIOGRAPHY. Four general works are outstanding in quality; all but one are international in scope and authorship. First must be rated the *Cambridge Ancient History*, 4 vol. (1923–26), 2nd ed. of vol. 1–2 in over 60 fascicles (1961–), 3rd ed., vol. 1, pt. 1 (1970). Second comes *Historia Mundi*, esp. vol. 1–4 (1952–56), in which the individual contributions are much longer, which makes it more readable, although it is not so useful for the serious student. Third is the latest group of volumes, compact in size, of the *Fischer Weltgeschichte*,

vol. 2–4 by leading French scholars on *Die altorientalischen Reiche* (1965–67). Still very useful is EDUARD MEYER, *Geschichte des Altertums*, rev. ed. by H.E. STIER; 2 vol. (1931–39). All the following books dealing in English with single countries or cultures may be highly recommended: SIR ALAN GARDINER, *Egypt of the Pharaohs* (1961); S.N. KRAMER, *The Sumerians* (1963); A. LEO OPPENHEIM, *Ancient Mesopotamia* (1964); KURT BITTEL, *Hattusha: The Capital of the Hittites* (1970); J.A. BRINKMAN, *A Political History of Post-Kassite Babylonia, 1158–722* B.C. (1968); JOHN BRIGHT, *A History of Israel* (1959); and R.N. FRYE, *The Heritage of Persia* (1963).

(W.F.A.)

Near Eastern Religions, Ancient

The religions of the ancient Near East (extending geographically from Iran to Egypt and from Anatolia and the Aegean Sea to the Arabian Peninsula, and temporally from *c.* 3000 to 330 BC, when Alexander the Great conquered much of the area) are the antecedents of the major Western religions; *i.e.*, Judaism, Christianity, and Islām. The religious beliefs, attitudes, and practices developed in the ancient Near East have had an enduring effect and influence on Western civilization. Among the many peoples whose religions contributed to this influence were the Egyptians, Hebrews, Moabites, Phoenicians, Ugaritians, Arabians, Aramaeans, Sumerians, Babylonians, Assyrians, Hittites, and Iranians.

NATURE AND SIGNIFICANCE

The ancient Near East cultural-historical area. The ancient Near East constituted an "ecumene." The term ecumene comes from the Greek word *oikoumenē*, which means the inhabited world and designates a distinct cultural-historical community. The material effects of the commercial and cultural interconnections that permeated the component regions of the ancient Near Eastern ecumene are richly supplied by archaeological excavations, which provide evidence of the spread of architectural, ceramic, metallurgical, and other products of ancient Near Eastern man's industry. Manufacturing and services tended to be monopolized by professional guilds, including religious personnel specializing in sacrifices, oracles, divination, and other kinds of priestcraft. The mobility of such guilds throughout the entire area helps to explain the spread of specific religious ideas and techniques over great distances. Just as guild potters spread ceramic forms and methods, so also guild priests spread their religious concepts and practices, from the Indian Ocean to the Aegean Sea, and from the Nile River to Central Asia. The Greek poet Homer, in the *Odyssey*, noted the mobility of guildsmen, mentioning religious personnel as well as architects, physicians, and minstrels. Guild priests called *kohanim* were found at ancient Ugarit on the Mediterranean coast of northern Syria as well as in Israel; and Mycenaean Greek (Late Bronze Age) methods of sacrifice are similar to the Hebraic methods, which are preserved in many countries to this day in the traditional techniques of Jewish ritual slaughter.

The spread of religious ideas and practices

Common elements in ancient Near Eastern religion. *Egypt and Mesopotamia.* In Mesopotamia, Egypt, and Canaan, each a sub-ecumene of the Near Eastern ecumene, the basic unit was the city-state with its own local religious cult. Each sub-ecumene had its universal gods; but in addition, each city-state within the sub-ecumene had its own local pantheon. In Egypt, from one end of the Nile Valley to the other, the sun god Re was revered as the universal god of life and Osiris as the universal god of the dead. But to those who dwelt in the Fayyum, or the Crocodile Nome (a political unit), Sebek, the crocodile god, had a special place. Similarly, in the Hare Nome, the hare god was worshipped. Each nome had its own cultic centre and its own priesthood that supervised the festivals, offerings, and other rites. An Egyptian who changed his residence from one nome to another put himself under the protection of a new local pantheon and its priesthood. His new religious calendar would prescribe a set of festivals differing from that of his former residence. He would have to acquaint himself with new customs and taboos. In the Hare Nome he would be forbidden to kill or eat rabbits; in the Crocodile Nome a parallel taboo would apply to crocodiles. Because of the widespread concept of such zoomorphic (animal-form) gods it therefore took little fundamental readjustment on the part of an individual who moved from one nome to another. Such was not the case, however, if an Egyptian moved to Mesopotamia. A ziggurat (tower temple) in Babylonia was functionally different from a pyramid (royal tomb) in Egypt; the former had to do with worshipping a living god, whereas the latter was the tomb of a dead king. Any nome in Egypt fostered the cult of Osiris and Isis (cult of the dead), which offered the prospect of a happy physical afterlife. Mesopotamia had nothing of the sort. Anywhere in Egypt the religious calendar was affected by the annual rising and subsiding of the same Nile. The situation was very different in Mesopotamia, where the twin rivers (Tigris and Euphrates) conveyed the concept that flooding was catastrophic—nothing like the beneficent inundation of the Nile.

Variety of local practices in Egypt

In Mesopotamia a man might move from Cuthah (in central Iraq), the cultic centre of the underworld deities Nergal and Ereshkigal, to Ur (in southern Iraq), the cultic centre of the lunar (moon) deities Nannar and Ningal. Each centre in Mesopotamia had its own cultic calendar, but the pantheon and the religious ideas were the same from one end of the sub-ecumene to the other.

Canaan. By the same token, the city-states of Canaan (on the eastern Mediterranean coast) adhered to a single world view that embraced a distinctive set of religious concepts and institutions. When the Moabite Ruth (in the Old Testament Book of Ruth) decided that she would leave her native land and attach herself permanently to Judah, the land of her mother-in-law, Naomi, she pronounced the formula for effecting the change: "where you go I will go, and where you lodge I will lodge; your people shall be my people, and your God my God; where you die I will die, and there will I be buried" (Ruth 1:16–17). This was not a conversion in the modern sense. Instead, it was a permanent change of residence within the same sub-ecumene, implying that one joins the people of his new land, in life and in death, under the aegis of the gods of that land and people. Ruth did not have to learn new concepts; she merely left the god Chemosh's Moabite sector of Canaan for the god Yahweh's Judaen sector of Canaan. No major change of language or psychology was required, any more than would have been the case for an Egyptian moving from Memphis to Heliopolis, or for a Babylonian moving from Mari to Sippar.

Political, social, and religious organization. All urban societies in the ancient Near East were stratified. At the top was the ruling class, controlling warfare, administration, and religion. This class was generally established by conquerors who reduced the vanquished natives to the status of helots (to use the Greek Spartan terminology) or "hewers of wood and drawers of water" (as the Bible expresses it in Deut. 29:11; Josh. 9:21,23,27). The conquerors received land grants in perpetuity in exchange for continued service to the state. In theory the land was owned by the god. The Sumerian inscriptions of Gudea (*c.* 21st century BC) depict the local god Ningirsu as the owner of the city-state of Lagash. The human ruler, or *ensi*, was the god's agent, governing in accordance with the god's decrees as revealed through dreams interpreted by qualified religious personnel. In the Bible, Lev. 25:23 expresses the theory plainly; God says "The land shall not be sold in perpetuity, for the land is mine; for you are strangers and sojourners with me."

On the steppe, nomadic tribesmen, who raised cattle by extensive grazing, were independent or semi-independent of the urban states. Yet some sort of accommodation had to be maintained between the desert and the agricultural and urban peoples, if for no other reason to keep the trade routes open.

Mobility of religious personnel

Guild mobility was such that priests or prophets might be invited by one ethnolinguistic group from another over considerable distances. In the biblical book of Numbers (22:2–24:25) there is an account of the Moabites (living in the hills east of the Dead Sea) summoning Balaam, a seer from Aram on the Euphrates, to come and execrate their Israelite foes because of Balaam's international fame

in the black art of cursing. The biblical account does not question that Yahweh inspired Balaam's pronouncements (23:16 ff.), which to the disappointment of his employers turned out to be blessings instead of curses. Denominationalism, in the modern sense, was alien to the ancient Near East. Priests were specialists in the techniques of finding out the will of the gods, of influencing the gods in favour of the priests' clients, and of pleasing the gods through food, drink, and song so that gods would be well disposed to their devotees and bring them good fortune. Theological indoctrination was at most secondary, and generally unimportant to the ancient priest and layman.

The monarchical rulers were naturally tempted to monopolize the powers of society by controlling the religious, military, economic, and other forces that made up the state. At Ugarit, for example, the king consolidated his control by placing family and favourites in the positions of high priests and generals and by giving them large land grants in exchange for loyal service. The Hebrews, on the other hand, aimed at a balance of power in their commonwealth. Though the legitimate king had to be of the line of Jesse, a member of the Judaen tribe, the priesthood was restricted to the tribe of Levi; and within that tribe, the high priesthood was initially limited to the line of Zadok (the high priest of King Solomon in the First Temple in the 10th century BC). The lay citizenry also played a role in important affairs of state. When the northern tribes wanted assurances that Rehoboam would not continue the harsh program of his father, Solomon, it was "all Israel" (I Kings 12:1, 12:3) or "all the assembly of Israel" that issued the ultimatum. When Rehoboam rejected their demands, it was "the people" who repudiated him and seceded.

Monotheistic trends. Biblical monotheism, as a historic force, is rooted in the international Amarna Age (15th–14th centuries BC). Egypt was the aristocratic centre of the Amarna Age, but included in it also were Mesopotamia, Canaan, Anatolia, and the area surrounding the Aegean Sea. Traders and diplomats from Egypt, Canaan, and Babylonia could not avoid the conclusion that Re (the Egyptian god of the universe), the Canaanite "God the Most High, maker of heaven and earth" (Gen. 14:19), and Marduk (the Babylonian creator of the universe) were all essentially the same supreme god, and that their names were merely translations of each other. When Alexander the Great (356–323 BC) claimed to be the son of Zeus for Greek acceptability, but the son of Amon for Egyptian acceptability, he was not talking out of both sides of his mouth. He was merely translating into action the phrase of the Greek historian Herodotus (2:42), "Amon is the Egyptian name for Zeus." By the same token Cyrus the Great was not hypocritical in calling the god of the heavens "Ahura Mazdā" for his Iranian subjects, "Marduk" for his Babylonian subjects, and "Yahweh" for his Jewish subjects.

Concept of an international supreme deity

Association of religion with the arts and sciences. Religion was associated with both the arts and sciences, though in the literature of the area it is hard to disentangle the secular from the sacred. Hymns, at one level, and omen or ritual texts, at another level, are clearly religious. But it would be difficult to categorize the Gilgamesh epic of Mesopotamia or the Homeric epics of Greece as definitely secular or religious. They deal with human events or worldly problems, but the gods are constantly on hand to affect the destiny of men. The same may be said for two Ugaritic epics, the epic of Kret and the epic of Daniel and Aqhat, of the Late Bronze Age. This also holds for the patriarchal narratives in the biblical Book of Genesis about Abraham, Isaac, and Jacob, in which God and his messengers play the same kind of role in human affairs as the gods in the Homeric or Ugaritic epics.

Religion had close ties with science as well as with literature and art. Astronomy, mathematics, and time reckoning are sciences in which the ancient Near East made great strides at an early date, long before 3000 BC. Heavenly bodies were at the same time both deities and personified numbers. The star Venus was the star that the Assyrians and Babylonians called Ishtar, which was at the same time both the goddess Ishtar and the deified number 15. The moon was not only earth's satellite but also the lunar deity Sin and the deified number 30. The most perfect number was one, for by advancing from zero to one men believed they proceeded from nonexistence to existence. Moreover, all other whole numbers were regarded as multiples of one, representative of the Creator, the Prime Mover, of the universe. The Egyptians called Re "the one One"; the Babylonians identified the divine "One" with Anu, the god of heaven. When the Hebrew prophet Zechariah (14:9) proclaimed "on that day the Lord will be one and his name one," he indicated that the Hebrews, like their neighbours, reckoned with sacred numbers and saw in the number one a symbol of the creator. Biblical monotheism, therefore, has more than one dimension, including not only the monotheistic principle that there is one God and none beside him, but also the mathematical principle of the primacy of "one" and its deification as the Prime Mover. Since even the earliest books of the Bible arose long after the development of the sciences in the ancient Near East, the Bible or its religious features cannot be regarded as primitive.

Significance of the number one

The importance and influence of Near Eastern culture and religion. Judaism, Christianity, and Islām are all products of ancient Near Eastern culture. The area is thus of primary importance in the history of Western religions. Unlike the paganisms of the ancient Near East, Judaism developed the concept of canonical (standardized and authoritative) Scripture that gave it continuity. No cuneiform (Mesopotamian wedge-shaped script) or hieroglyphic (Egyptian pictorial script) religious text ever gained such canonical exclusiveness as the Bible. The paganisms were also limited by idolatry. Various conquerors captured or destroyed the idols of the vanquished, whose morale was thereby crushed, because their gods were no longer present, or alive, to help them. The Hebraic insistence on an incorporeal God, with no tangible image, established bonds that were not material and therefore could not be broken. Moreover, the prophets of Israel developed a blueprint of history that prepared the Jewish people for the troubles that lay ahead, but with the eschatological promise that God had a happy ending in store for the saving remnant who deserved well, in a golden age under the messianic King.

The optimism of the Hebrews was quite different from the attitudes of other nations. The Mesopotamians, for example, were inclined to regard the gods as evil. In the 11th Tablet of the Gilgamesh epic all the gods except Ea were regarded as malevolent. The Egyptians cherished a cult of the dead whereby the individual could, under proper conditions, achieve a happy immortality of an earthly nature, but there was nothing built into the religion to perpetuate the Egyptian way of life, once Egypt was exposed to Christianity and Islām. Egyptian religion and culture eventually became bankrupt and vanished. The Mycenaean Greeks did not even look forward to a satisfactory afterlife at a personal level. When the Homeric hero Odysseus tries to comfort the spirit of Achilles (who had died in the Trojan War) in the underworld, Achilles replies: "Do not seek to speak to me soothingly of death, O glorious Odysseus. I would prefer so that I might live on earth, to work as the hireling of another, of some portionless man whose livelihood was not great, rather than be king over all the dead that have perished" (*Odyssey* 11:488–491). For the Greeks the poorest of men among the living was considered better off than the most fortunate in the afterlife.

Concepts of the afterlife

There is also a qualitative factor that helps explain why, of all the ancient Near Eastern religions, Judaism survived and laid the foundation for Christianity and Islām. Situated at the centre of the ancient Near East and open to influences from all sides, Judaism enjoyed the opportunity of absorbing a maximum of influences from its talented neighbours that it proceeded to build upon creatively. Sumero-Akkadian poetry at its best is impressive, but it falls short of the sublimity of the psalms of David or the oracles of Israel's prophets. Egypt gave the world its first secular prose literature: the novelette. But a comparison of the Joseph story in Genesis with a similar Egyptian story of an attempted seduction shows that the Hebrews

learned their lessons so well from the Egyptians that they eclipsed their masters. The same holds for religion. The ancient Near East had, long before the appearance of the Hebrews, a rich record of ethical, moral, and religious expression. But it remained for the Hebrews to make "social justice"—the standard whereby, in any conflict of interests, the poor or the weak is always right versus the rich or the strong—the supreme requirement laid down by God, to require international peace as a permanent condition for a world worth living in, and to proclaim that universal and total disarmament was the only way of achieving that goal (in Isa. 2:4; Mic. 4:3).

The standard of social justice

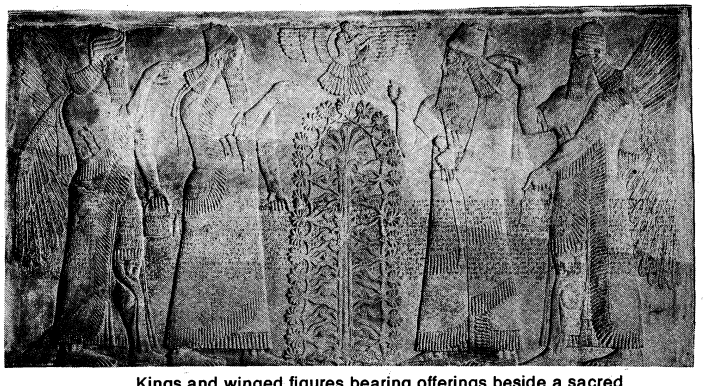
Materialism, no less than war, was a blight on the ancient Near East because of the great emphasis placed on trade. Several Ugaritic texts represent the god Baal not only as coveting, but revered all the more by his Canaanite devotees because he was mighty enough to get what he coveted. Reacting against this aspect of their milieu, the Hebrews came up with the Tenth Commandment (against covetousness).

SOURCES OF KNOWLEDGE

The "archaeological revolution." The decipherment of Mesopotamian and Egyptian literatures in the 19th century opened new vistas of ancient Near Eastern history. Hitherto, scholarly knowledge had been limited to the contents of classical Hebrew, Greek, and Latin literatures. Explorations and excavations in the Near East yielded not only texts but an abundance of ancient objects of art, artifacts of daily life, and architecture, and thus have revolutionized scholarly knowledge of the ancient Near East, including its religions. A ziggurat excavated at Babylon illustrates the form of the biblical Tower of Babel. The prototype of the biblical story of the Deluge has turned up in the Gilgamesh epic. A fragment (dating from about 1400 BC) of that Babylonian epic has been found at Megiddo in Israel, showing that the Mesopotamian version was current in Palestine before the Hebrews, under Joshua, conquered the land around 1200 BC. A previously little-known people, the Hittites, are, due to archaeological discoveries, now recognized as a major power of antiquity with a rich legacy of religious texts, especially rituals.

The earliest and certainly the most fundamental ancient Near Eastern civilization—the Sumerian—had vanished without a reference in the literatures of the world. Now, Sumerology is an important field of investigation. Biblical studies have been greatly revolutionized by the tablets (1400–1200 BC) found from 1929 onward at Ugarit. It has become extremely difficult for anyone to keep abreast of the continually growing body of material, and very few scholars today feel secure enough to venture beyond limited areas.

By courtesy of the trustees of the British Museum



Kings and winged figures bearing offerings beside a sacred tree. Assyrian marble relief, reign of Ashurnasirpal II, *c.* 880 BC. In the British Museum.

Literary sources of knowledge of ancient Near Eastern religion. Classical literature remains an important source for ancient Near Eastern religion. The Roman historian Livy (59 or 64 BC–AD 17) wrote many descriptions of religious rites of the ancient Near East. The Roman poet Virgil's (70–19 BC) *Aeneid* and *Eclogues* reflect Egyptian, Semitic, and Anatolian, as well as Greek, antecedents. The Greek biographer Plutarch's (AD 40/50–120/125) *De Iside et Osiride* ("Concerning Isis and Osiris") is still the best description of the Egyptian myth of Isis and Osiris and of the cult of the dead. The Greek satirist Lucian's (*c.* AD 120/125–190) *De Dea Syria* ("Concerning the Syrian Goddess") is of enduring value for an understanding of Canaanite religion. The writings of Herodotus, the 5th-century-BC Greek historian, remain an indispensable source for the cultural history and religion of the ancient Near East. And due to the discovery of texts from Ugarit, the Homeric epic of the Greeks is now firmly linked to Near Eastern literature.

The Hebrew Bible is still the most important single source for knowledge of the ancient Near East, reflecting life from Egypt to Iran, and from the Bronze Age beginnings to the Hellenistic Age. There is very little in the Old Testament that does not follow the types of religious literatures in the older Near East: psalms, hymns, laws, rituals, prophecy, wisdom literature, and other types. Sometimes parts of the Bible are related in detail to specific outside sources. The Egyptian *Wisdom of Amenemope*, first published in modern times in 1923, parallels Prov. 22:17–24:22 so closely that it effectively opened up the comparative study of ancient Near Eastern wisdom literature.

The Old Testament as a source of knowledge

WORLD VIEWS AND BASIC RELIGIOUS THOUGHT

The sacrality of nature and natural forces. All of the ancient Near Eastern people saw the agency of the gods in every aspect of life and nature. Everything on earth was regarded as a reflex of its prototype in the divine or sacred sphere, such as in the biblical description of the creation of man "in the image of God"; God was viewed as the primary reality and man the reflection thereof. In Egypt Thoth was the scribe in the pantheon. Mortal scribes were viewed as the human reflections of Thoth, and "the beak of the Ibis [*i.e.*, Thoth] is the finger of the Scribe" (*Wisdom of Amenemope*, ch. XV, 17:7).

The ancient Near Eastern people believed that the universe resulted from the injecting of order (cosmos) into chaotic primordial beings or matter, followed by divine acts of creation. Gen. 1:1–3 says: "When God set about creating the heavens and the earth [*i.e.*, the universe], the earth was in a state of chaos and darkness was upon the face of the deep; and the Spirit of God was flying over the face of the waters. And God said, 'Let there be light!'; and there was light." Thus darkness (*i.e.*, evil) was pre-existent. Moreover, the deep (*tehom* in Hebrew) is the same as the primordial dragon called Tiamat (cognate to the Hebrew *tehom*) in the Babylonian epic of creation. The first act of creation is God's evoking light (*i.e.*, the forces of good) by fiat. Accordingly, God is not responsible for the forces of evil, which were there before he embarked on the creative process. Proceeding by fiat he separated the water-containing earth from the water-containing heaven, confined the earth's waters to the bodies of water (leaving the rest as dry land), created the various species of vegetation, the heavenly bodies, the animal kingdom, and finally man who is to rule over the earth. All this takes six days, after which God rests on the seventh, so that the sabbath crowns the epic of creation and imposes the obligation to observe the sabbath in keeping with the principle of *imitatio Dei* (the imitation of God).

Creation of the universe

The Babylonian creation epic (*Enuma elish*, "When on High") states that at first there existed only the male (Apsu) and female (Tiamat) gods of the deep. They raised a family of gods that were so unruly that Apsu resolved to destroy them. Rebellion and chaos ensued. Among the deities was Marduk, the god of Babylon. Since the main version of the epic of creation is the Babylonian, Marduk occupies the role of Creator. (In the Assyrian version, Ashur is important.) Tiamat, who had embarked on a course of destruction, was slain by Marduk, who cut her in two and used her carcass to create the universe. Out of half her body he fashioned the sky containing the heavenly bodies to mark the periods of time. The epic culminates in the glorification of Marduk and the establishment of his

order. The *Enuma elish* was read on the Akitu, or New Year festival, at Babylon, to re-establish order, in accordance with sympathetic transference principles, by reciting Marduk's creation. The function of the Akitu is thus to rejuvenate society for the new year.

Views of man and society. The lack of hard and fast barriers between gods and men left room for hybridizing. The aristocracy, in particular, claimed some divine form of ancestry. Gilgamesh, a mortal king who ruled Uruk in Mesopotamia, was, according to the Gilgamesh epic, born of the goddess Ninsun, even as among the Greeks Achilles was accepted as the son of the goddess Thetis. Sometimes kings claimed to have two divine parents. King Kret, whose epic was found at Ugarit, claimed to be the son of El, the head of the pantheon, and of Asherah, El's wife. Every Egyptian pharaoh was hailed as "the son of Re" (the sun god). This does not, however, imply the absence of a human father. The concept was one of paternity at two levels; qualitative superiority emanated from the notion of divine paternity, but one's position in society came from the human husband of one's mother. Odysseus "the Zeus-begotten son of Laertes" (*Iliad* 10:144) was a hero because Zeus presumably impregnated his mother; but he was also king of Ithaca because his mother's husband was King Laertes of Ithaca. In this regard the birth and station of Christ differ only in that Mary was a virgin when she was divinely impregnated. Though the divine component of Christ is due to his divine paternity, his position as king of the Jews comes not from his heavenly Father but from Mary's husband, Joseph, who was descended from King David (Matt. 1).

In the ancient Near Eastern world view, gods could become mortal, and men could become gods. Utnapishtim, the hero of the Babylonian Flood story, was deified together with his wife by the fiat of the great god Enlil: "Hitherto Utnapishtim has been but human; henceforth Utnapishtim and his wife shall be like us gods" (Gilgamesh epic 11:193–194). In the Hebrew Bible, God so loved Enoch (Gen. 5:24) and Elijah (II Kings 2:11) that he carried them aloft to heaven as immortals. But these were special cases, and in antiquity they set no precedent for common folk. Kings enjoyed deification regularly in Egypt, though in some other traditions only upon dying. The Hittite monarch Hattusilis III refers to his father's death as "when my father Mursilis became a god" (*Apology of Hattusilis*, line 22).

Purpose of man

From the ancient Near Eastern point of view, man was created to serve the gods, and he does so in the hope that the gods appreciate it and will reward him for it. The gods need food and drink and depend on men to supply them. After the Flood the biblical Noah won God's goodwill, for "the Lord smelled the pleasing odor" (Gen. 8:21) of the tasty flesh and fowl offered up to him. Noah was following a long tradition, for Utnapishtim (Gilgamesh epic 11:155–161) had, after the Flood, offered sacrifices and libations to the gods who "crowded like flies" as they "smelled the sweet savor." Though gods depend on man, man also depends on the gods, and therefore service to the deities must be maintained for the welfare of the state, even as the family and the individual must do what the gods expect of them for domestic and personal welfare.

Everything on earth reflects a divine prototype, and all human affairs are divinely ordered and scrutinized. Gods may even build the cities destined to be their cultic centres and in which they are to reside, at least part of the time. The Greek god Poseidon built the walls of Troy, according to the *Iliad* (21:446–447). At Ugarit, Baal's temple was designed and built by Kothar-wa-Hasis, the god of arts and crafts. The Israelite King David gave his son Solomon plans for the Temple drawn up by Yahweh's (the Lord's) own hand (I Chron. 28:19).

National policy went hand in hand with theology. Ashur was the god of Assyria; the kings of Assyria were in theory his chief executive officers. Thus Sennacherib, king of Assyria, in undertaking a military campaign, recorded that it was not on his own initiative but in conformity with Ashur's will: "In my second campaign, Ashur my Lord impelled me." When the Hebrews and Transjordanians had a border dispute, Jephthah told the latter: "Will you not possess what Chemosh your god gives you to possess? And all that the Lord our God has dispossessed before us, we will possess" (Judg. 11:24). There was no such thing as secular policy in the ancient Near East.

Since the king was the human agent of the god, he was exalted above other men. In Israel, the king was chosen by God to rule his people. God's representative was a priest or prophet who consecrated the king by anointing his head with oil. But the king of Israel was not divine, neither while on the throne nor after death.

The divinity of kings evoked certain fictions. By sucking the breasts of goddesses, crown princes imbibed a source of divinity. The baby pharaoh sucking the breasts of Isis (who was perhaps in real life represented by her high priestess) is a common motif in Egyptian art. In Mesopotamia, it was not the usual practice for kings to claim divinity, but now and then it cropped up. Naram-Sin (23rd century BC) prefixed the sign for divinity before his name and was officially a god. The same usage is attested among kings of the 3rd Dynasty of Ur (*c.* 2100 BC).

Views of basic values and ends of human life. The good life was one lived in accord with the regulations of one's god. In the realm of ethics and morals there was more international uniformity than there was in taboo and ritual. Honesty and kindness were universally recognized as good, theft and murder as bad. Wisdom literature tended to stress the same virtues and to condemn the same vices, regardless of the region and cult. It remained for the prophets of Israel to single out uncompromising virtue as the overriding consideration in the good life required by God. The most important factor in that system was "social justice," whereby the weak was always protected in conflicts of interest with the strong. This had an important place in what may be called "international religion"; *i.e.*, that governing relations between men from different areas belonging to different cults. That level of religion, called "fear of the gods," is tested when the strong man confronts the weak. The strong man who injures the weak lacks the fear of the gods; the strong man who helps the weak has the fear of the gods. This was religion transcending all the regional cults, and it came into play when strangers abroad were at the mercy of the local inhabitants. Odysseus in a foreign land wanted to know if the people there feared the gods or were lawless so that no stranger was safe (*Odyssey* 9:176). Abraham, too, was concerned in Philistia lest the inhabitants might kill him because there was no "Fear of God(s)" (Gen. 20:11). Men of all nations and all cults knew that only among god-fearing men was there decency or safety.

Ethical emphases

There was another common trend in international religion. No matter how polytheistic a cult may have been, it left a place for the god shared by all men. Theos, "God" (not merely "a god"), is in Homer; *pa netjer*, "the God," occurs in Egyptian exactly like Elohim, "(the) God," in Hebrew. Nebuchadrezzar, the 6th-century-BC Babylonian king, made Zedekiah, the Judaean king, swear by Elohim (II Chron. 36:13), the God of the universe for Babylonians and Hebrews alike. Similarly, when the Hebrews spoke of truth uttered by Pharaoh Necho, which fell on the deaf ears of the Judaean King Josiah, the text (II Chron. 35:21) states that Elohim, "God," had spoken through the mouth of the pharaoh.

In Egyptian religion (followed by Judaism, Christianity, and Islām), the concept of a happy afterlife depending on one's ethical and moral record in this world was developed. Vignettes in the various Egyptian books of the dead show the deceased's heart being weighed against the feather of truth in the balances before the scribe god Thoth, who records the text. When the Bible speaks of God as "who tests the heart and the kidneys" (Ps. 7:9; Jer. 11:20 and 20:12) it refers to the same concept.

Myths as the basic mode of religious thought. Myths were developed to account for the cosmos. How did the gods bring heavens, earth, plants, beasts, and men into existence? What is the divine origin of human institutions and of the ecumene? What divine process is responsible for prosperity or failure? To explain such basic questions etiological (origin or causal) myths were developed. For

Purpose of myths

example, the attraction between man and woman (and the consequent institution of marriage) is explained by the myth that primeval man was one creature, subsequently divided into two parts, male and female, which are attracted to one another to regain their pristine unity. Aristophanes expresses this theory of sexual attraction in Plato's *Symposium*. Genesis has the same theory in the familiar myth that a rib, taken out of Adam, was fashioned into Eve; and precisely because woman was taken out of man, man forsakes his father and mother to cleave unto his wife so that they become one flesh (Gen. 2:23–24).

Myths are often invoked in magic (which, unlike religion, aims at compelling, instead of imploring, the gods). To banish evil from the life of a client, the magician may invoke the cosmic myth whereby the forces of good triumph over the forces of evil. Evil is depicted as a seven-headed monster on a seal of the Akkad period (3rd millennium BC) in Mesopotamia, whose heads are being successively killed by good anthropomorphic (human-form) beings. At Ugarit, in mythological poems of the Late Bronze Age, the good gods Baal and Anath slay the wicked Leviathan of the Seven Heads, providing the precedent for the victory of good over evil. The Hebrews also nurtured this myth whereby God slays the many-headed Leviathan (Ps. 74:14) and will do so again at the end of days, to quell evil and establish good for all eternity (Isa. 27:1).

Magical concepts and attitudes. The loftier trends of ancient Near Eastern religion did not as a rule threaten to eliminate magic. White, or protective, magic was never seriously discouraged. Black, or destructive, magic was frowned on by organized society, regardless of whether the official religion was monotheistic or polytheistic, because black magic makes its victims unfit for functioning productively in society. Section II of the Babylonian King Hammurabi's (Hammurapi's) code punishes witchcraft (as well as false accusations of witchcraft) with the death penalty. Moreover, all organized religion tended to oppose magic that circumvented the official clergy. King Saul of Israel had characteristically banned sorcery, driving it underground. Yet when he wanted guidance from the dead prophet Samuel, Saul consulted the Witch of Endor who was practicing her art illegally (I Sam. 28:6–25). She was able to call up the spirit of the prophet from the underworld, which, incidentally, illustrates one of the reasons why society opposes spiritualism. The witch, by claiming to bring the greatest authorities of the past onto the current scene, threatens the authority of the establishment.

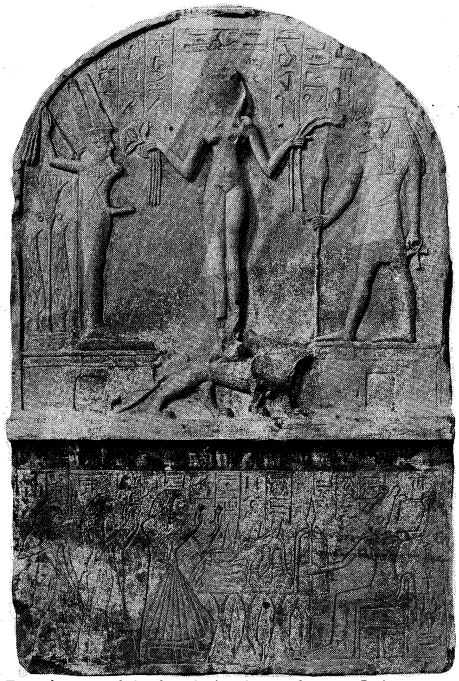
RELIGIOUS PRACTICES AND INSTITUTIONS

Nature: the framework of ideas and practices. Fertility of agriculture, of edible animals, and of the human population was a paramount factor in the life and religion of the ancient Near East. The forms that the fertility rites assumed varied from region to region, depending on climate and geography. Rain and dew were all-important in Canaan, but of little significance in Egypt. In both areas water was crucial, but the source of the life-giving water was entirely different. The agricultural year varied in the two regions. In Egypt the year was divided into three seasons: inundation, sowing, and harvest. In Canaan there were two seasons: the winter characterized by rainfall and the summer characterized by dew. The year was punctuated by different agricultural activities, as is indicated in the Gezer Calendar in which all 12 months are accounted for as times of profitable agricultural activity, with harvests in the rainless summer as well as in the green winter. Anxiety was caused by the uncertainty of rain in the rainy season and of dew in its season. All of the regions of the ancient Near East schematized the blessing of good years and the threat of bad years in terms of seven-year cycles. A Mesopotamian text illustrating this is the Gilgamesh epic (8:101–113), in which the slaying of the hero Gilgamesh would initiate seven lean years. At Ugarit the slaying of the hero Aqhat evokes a curse depriving the land of rain and dew for seven (or, climactically, eight) years. The seven lean and seven fat years in the biblical story of Joseph in Egypt reflect the same system. In Egypt, of course, rain and dew are out of the picture; instead, generous Nile risings mean prosperity; inadequate risings in the season of inundation spells misery. A text of the Ptolematic period (4th–1st centuries BC), purporting to record events of the Pyramid age, tells of seven lean years in the reign of Djoser (3rd dynasty; *i.e.*, *c.* 2686–*c* 2613 BC). The pharoah appealed to the gods, who responded by restoring an abundant flow of the Nile.

Seasonal patterns

The population desired the normal pattern of times and seasons, so that "seedtime and harvest, cold and heat, summer and winter, day and night, shall not cease" (Gen. 8:22). But since the seasonal pattern is not dependable, the need for order evoked a system of cycles: notably the sabbatical cycle. A Ugaritic liturgical text especially designed for this phenomenon aims at terminating a sabbatical cycle of privation and ushering in one of fertility by celebrating the birth and triumphal entrance of the Seven Good Gods, whose advent brings an abundance of food and wine.

By courtesy of the trustees of the British Museum



Egyptian sepulchral stela by the craftsman Qaha, 19th dynasty. (Upper portion) The Syrian fertility goddess Qadesh stands atop a lion in the presence of the Egyptian fertility god Min (left) and the Syrian god of lightning and thunder Resheph (right). (Lower portion) The Syrian goddess Anath (seated right) and worshippers. In the British Museum.

Fertility rites

It was only natural that fertility rites should include sexual myths acted out dramatically. The Ugaritic text just alluded to describes El, the head of the pantheon, copulating with two human women. This has echoes in Hosea and Ezekiel where God, as in the Canaanite literary tradition, is referred to as having figuratively a love affair with two women, symbolizing Judah and Israel. The Hebrews, however, eventually eliminated sex from their official theology and religious practices. Up to the time of King Josiah's reform (621 BC) there was a women's cult of Asherah (under *qedeshim*—or consecrated for fertility practices—auspices, according to II Kings 23:7) in the Jerusalem Temple, alongside the male cult of Yahweh. Asherah's devotees considered her the chief wife of Yahweh, even as she was the wife of El, head of the Canaanite pantheon, for in the Bible, El is identified with Yahweh. But Josiah eliminated the cult of Asherah, and official Judaism has since then left no place for other gods, which meant the elimination of every goddess. Popular religion, to be sure, persisted in the female fertility principle until the destruction of the Temple in 586 BC. In Judaean excavations Astarte figurines were found in private homes down to that time. Further purification of the

Hebrew religion, intensified by the catastrophe of 586, put an end to the practice of pagan fertility rites, including the use of goddess figurines. Without goddesses there could be no sexual activity in the pantheon, and thus Judaism has developed without any divine mother figure.

The ancient Near East made a place for homosexuality and bestiality in its myths and rites. In the Asherah cult the *qedeshim* priests had a reputation for homosexual practices, even as the *qedeshot* priestesses for prostitution. Israel eventually banned both the *qedeshim* and *qedeshot*, while in Ugarit the *qedeshim* and *kohanim* were priestly guilds in equally good standing. Baal is portrayed in Ugaritic mythology as impregnating a heifer to sire the young bull god. The biblical book of Leviticus (18:22–27) bans homosexuality and bestiality expressly because the Canaanite population had been practicing those rites, which the Hebrews rejected as abominations.

Ritual and myth

According to ancient views, the myth came first, and the rite imitated or re-enacted it. This, however, is not necessarily the order in which religion develops. Rites can be very tenacious; and when the origin of a rite had been forgotten, a myth was often invented to explain it.

Types of religious organization and authority. Religion occurs at different levels of society: personal, familial, local, national, and international. At the personal and international extremes there is need for but little organization. And yet in religion as the ancient Near Easteners saw it, there was a progression from one stage to the next. In the early myths of Genesis, God and Noah have direct personal relations. This leads to a covenant between God and all who went out of the ark: birds and beasts as well as mankind (Gen. 9:9–10). Through the sons of Noah and their descendants, who form the nations of the world (Gen. 10), there is a theoretical progress to international religion. This scheme of the relations between God and mankind, from the personal to the universal level, mirrors the historical record of religion. Judaism (followed by Christianity and Islām) traces "the Religion" back to Abraham, who had personal and direct relations with God, as was customary in the ancient Near Eastern milieu. Abraham's intimacy with God is similar to the intimacy between Odysseus and the Greek goddess Athena. The next step is a covenant between a particular deity and a particular man, binding the two together in a contractual relationship for all eternity from generation to generation. Such covenants were not rare; the Hittite King Hattusilis III made such a covenant with Ishtar. Abraham's covenant is unique simply because it was the only one destined to last in history.

The descendants of able men who established a dynasty or tradition would worship the God of their father, or fathers, and adhere to the original covenant. Gen. 31 portrays Jacob and Laban swearing by their respective ancestral gods: Jacob by the god(s) of Abraham and Laban by the god(s) of Nahor. Once a group expanded into a federation of clans or tribes, religious organization became necessary. A central shrine (such as the one at Shiloh in Israel) for amphictyonic (religious confederational) pilgrimage festivals required a professional priesthood and other religious personnel to take care of sacrifices, give oracular guidance, interpret dreams and omens, as well as to provide instruction. In an amphictyony of 12 tribes, each tribe could render federal service for religious and secular purposes, one month each year. A special tribe (such as the Levites in Israel, or the Magians in Iran) could be dedicated full time to cultic duties. A greater degree of centralization and organization of the cult would generally follow from the establishment of a powerful state. The cult of Marduk of Babylon spread in importance and influence because Babylon became the capital of a powerful kingdom in the time of Hammurabi (18th century BC) and of a mighty empire during the reign of Nebuchadrezzar (604–562 BC). The Egyptian cult of Amon-Re not only became powerful but took on the form of a universal religion as a result of the military and political triumphs of the rulers of Thebes, particularly during the reign of Thutmose III (1504–1450 BC). Under Solomon, Jerusalem became the centre of a great commercial empire. The Temple of Solomon and its God, the God of Israel, were catapulted into international prominence quite different from the national status that marked the extent of Hebraic religion previously. The new internationalism of Israel's involvements paved the way for the universality of the views of the prophets. The God of Israel was subsequently concerned with all mankind and not merely with one people in one small land. This ultimately meant the transformation of biblical religion from the cult of a people to a more subtle, spiritual movement that required different organization and different personnel. The priesthood became defunct with the destruction of Herod's Temple and the cessation of sacrifices in AD 70. The new religious leaders (rabbis) were rather teachers and spiritual guides united by dedication to the same scripture. The spread of the devotees over the face of the earth meant that they were now divided into regional groups, under different sovereigns, and the individual Jewish communities were organized independently, each with its own house of worship.

Professional priesthoods

There were various devices for holding an ethnic-religious group together even though it was fragmentized into scattered communities. Laws of purity, especially those pertaining to diet, kept different groups apart. Each respected the other's rules, but the fact that each group had different taboos kept them from breaking bread together and mingling socially. They could do business with each other in the marketplace, but they could not fraternize in each other's homes. Above all, laws of purity were deterrents to intermarriage, the major factor that breaks up religious communities and encourages homogenization.

Laws of purity

Folk religion. Alongside the official religion, folk or popular religion had a place. The devotees in ancient temples and houses of worship were mainly males. The women were more concerned with fertility. Love charms, magical rites to end barrenness, and amulets to help in childbirth and keep the children alive were important in the minds of the women; and their popular religion was very different from the official cults of their husbands.

Funerary cults. From the earliest times funerary rites were an important aspect of religion. All cults of the dead assume that there is some sort of continuity of the deceased's personality. Moreover, there was always the possibility of contact between the living and the dead. The community was conceived in terms of the deities, the living, and the dead—all in contact with each other. For this reason burial grounds were an important part of the community's installations. The dead might be buried under the floor of the home, in a family vault, in nearby caves, or in a necropolis—but in some way, the connection between the living and their dead must be perpetuated.

CONCLUSION

The soil of the Near East, which is the cradle of the most dynamic religions in the world today, continues to yield so many artifacts, texts, and monuments bearing on religion that the study of ancient Near Eastern religion is likely to remain a dynamic field for many years to come.

BIBLIOGRAPHY. An excellent unified cultural history of the ancient Near East is EDUARD MEYER, *Geschichte des Alterthums*, 2nd ed., 3 vol. (1907–37). More up-to-date, but very uneven, is the new edition of *Cambridge Ancient History*, now appearing in fascicles. A concise treatment is C.H. GORDON, *The Ancient Near East*, 3rd rev. ed. (1965). Biblical religion has an enormous bibliography. Among the many treatments is W.O.E. OESTERLEY and T.H. ROBINSON, *Hebrew Religion: Its Origin and Development* (1930). ROLAND DE VAUX, *Les Institutions de l'Ancien Testament*, 2 vol. (1958–60; Eng. trans., *Ancient Israel: Its Life and Institutions*, 1961), discusses Israelite religion and provides a copious bibliography. The best set of translations, with a bibliography for each text, is J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts*, 2nd ed. (1955), updated by *The Ancient Near East: Supplementary Texts and Pictures Relating to the Old Testament* (1969). A handy, abridged version of Pritchard's works is available in a paperback edition, *The Ancient Near East: An Anthology of Texts and Pictures* (1958). Two excellent works providing a background for the whole subject are S. MOSCATI, *The Face of the Ancient Orient* (1962); and H. and H.A. FRANKFORT, *The Intellectual Adventure of Ancient Man* (1946).

(C.H.G.)

Nebraska

One of the west central states of the United States, Nebraska, during its early history in the first half of the 19th century, served primarily as a water and land route to the rich trapping country to the north and west. With the coming of railroads after the Civil War and the consequent immigrations, however, the excellent soils of Nebraska were broken by the plow, and its grasslands gave rise to a range cattle industry. As a result, the state has been, almost since its admission to the Union in 1867 as the 37th state, one of the major producers of food commodities for the nation.

An overview of the state

Rivers were important to Nebraska's geography and settlement. The Missouri River, a major highway to the trans-Mississippi West in the early 19th century, forms the eastern boundary with Iowa and Missouri, and about a fourth of the northern boundary with South Dakota. The name Nebraska is derived from an Indian word meaning "flat water," a reference to the Platte River, which served as a magnet for urban clusters as it crossed the state. It is formed by the confluence of the North and South Platte rivers, both of which rise in Colorado on the southwest, though the North Platte swings northward through Wyoming, on the west, before entering Nebraska. The southern boundary with Kansas was established when the two territories were created by the Kansas–Nebraska Act in 1854, legislation that heightened the sectional hostilities that exploded into the Civil War. In 1970 a majority of the nearly 1,500,000 Nebraskans counted in the census lived close to the Missouri and Platte, leaving much of the state's 77,227 square miles (200,017 square kilometres) relatively lightly populated.

Agriculture is basic to Nebraska's economy, but only one-fifth of its labour force is employed directly in farming or ranching. Economic conditions long have had a direct bearing on the state's political life, including a brief period of protest through the agrarian-oriented Populist Party in the 1890s. Although Nebraska traditionally has been considered a stronghold of Republicanism, the changing fortunes of its people make this generalization dubious. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; and GREAT PLAINS.)

THE HISTORY OF NEBRASKA

Various prehistoric peoples inhabited Nebraska as early as 8000 BC. In modern times semisedentary Indian tribes, most notably the Poncas, Omahas, Otoes, and Pawnees, inhabited eastern and central Nebraska. The west was the domain of the nomadic Brulé and Oglala Sioux, but Arapahoes, Comanches, and Cheyennes, likewise nomads, were often present.

Exploration and settlement. The region that was to become Nebraska was included in the Louisiana Purchase, the vast tract of midcontinent America acquired from France in 1803. The French had lost interest in the trans-Mississippi West after less than three years' rule, and the Spanish before them had done little to develop the vast, ill-defined area.

In 1804 the Lewis and Clark Expedition visited the Nebraska shore of the Missouri River and conducted the first systematic exploration of the area. Shortly thereafter, a vigorous fur trade developed along the Missouri, but Nebraska was primarily a highway to richer fur areas to the north and west. In the 1840s the Platte Valley became another highway, as thousands of Americans moved westward over the Oregon, Mormon, and California trails.

Roles of speculators and railroads

Much interest soon developed in Nebraska and in the Platte Valley as a potential railroad route to the Pacific. Moreover, frontier land speculators in western Missouri and Iowa anticipated great financial gains if the Nebraska country, part of the large Indian domaine between the Missouri River and the Rocky Mountains, were opened for settlement. With the adoption of the Kansas–Nebraska Act in 1854, the federal government extended political organization to the trans-Missouri region. As originally established, Nebraska Territory comprised 351,558 square miles, but by 1863 the organization of the Colorado, Dakota, and Idaho (including future Montana and Wyoming) territories had reduced Nebraska almost to its present dimensions.

Much of the economy of the early Nebraska settlements along the Missouri River focussed upon land speculation. Agriculture began to develop within a few years, however, and some river towns became important transfer points for freight and passengers going west. The completion of the Union Pacific Railroad in 1869 and the further railroad construction that followed made possible the development of the interior of Nebraska.

Statehood. After Nebraska's admission to the Union, in 1867, and despite a depression and grasshopper plague, its population grew from about 120,000 to more than 1,000,000 by 1890. The Indian resistance on the frontier was broken, and settlement extended westward into the panhandle of the state. In the 1880s Omaha became an important industrial and meat-packing centre, and Lincoln grew to prominence as the state capital and seat of the University of Nebraska.

Impact of economic cycles

In the 1890s, Nebraska's farmers, afflicted with poor crop prices, high transportation costs, and depression, expressed their protest through the Populist Party. Although the Populist movement was short-lived, it invigorated the political life of the state. Yet prosperity returned by 1900, and Nebraska enjoyed two decades of economic advancement. Through the 1920s, Nebraska's agriculture again was beset with mediocre marketing conditions, and with the advent of the Great Depression of the 1930s the state's entire economy suffered its greatest adversity, a condition that was alleviated by massive federal assistance. World War II brought an economic recovery, and since then Nebraska has enjoyed a steady, though unspectacular, economic growth.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. Nebraska takes in parts of two physiographic regions, the till plains of the Central Lowland (in the eastern third) and the Great Plains.

The surface features. The elevation rises from a minimum of 840 feet above sea level in the southeast to a maximum of 5,424 feet near the Colorado–Wyoming boundaries. Much of the land is gently rolling prairie, although the river valleys, much of south central Nebraska, and a large portion of the panhandle district are flatlands. The Sand Hills country of north central and northwestern Nebraska is a vast, treeless, grass-covered region that comprises almost one-fourth of the state.

Soils. All of Nebraska's six basic soils are excellent for agriculture. The prairie soils of the southeast are outstanding for the production of corn, while the rich humus soils of central and northeastern Nebraska are, given adequate moisture, highly suitable for general farming. South of the Platte and west of the prairie-soil area, the soil is best suited to small-grain production. Winter wheat adapts to the soil and marginal precipitation of western Nebraska. The unique, wind-deposited soil of the Sand Hills, because of limited precipitation and the danger of erosion, is suited solely to cattle grazing. The many small lakes and luxuriant grasses of the Sand Hills have made the area one of the best rangelands in North America. The alluvial soils of the Missouri and Platte river valleys and the valleys of smaller streams are outstanding for raising corn and other crops.

Waters. Nebraska is entirely within the Missouri River drainage system; the Platte, the major Nebraska tributary, joins the Missouri south of Omaha. Although the Platte is shallow and unnavigable, its water is vital to the state's irrigation. The Elkhorn River enters the Platte west of Omaha, while the Loup River, formed by three tributaries flowing out of the Sand Hills, also discharges into the Platte. The Republican and Big Blue rivers flow through southern Nebraska, emptying into the Missouri in Kansas via the Kansas (Kaw) River. The Niobrara, a swift-moving stream that rises in the high country just east of the Wyoming border, flows across extreme northern Nebraska. Nebraska also has a vast supply of underground water that has made possible the extensive development of well irrigation.

Drainage to the Missouri River

Climate. Hot winds from the southwest often push summer temperatures in Nebraska above 90° F (32° C) and sometimes above 100° F (38° C). Average July temperatures range from 74° F (23° C) in the panhandle to 78° F (26° C) in the southeast. In the winter, northwest winds often bring in Arctic high-pressure masses from Canada, and temperatures commonly fall well below 0° F (−18° C). Low-pressure systems, moving out of the southwestern states, sometimes bring great blizzards to the state and present a danger to travellers and stock raisers. Average January temperatures vary from 25° F (−4° C) in the panhandle to about 21° F (−6° C) in the northeast. The average growing season is 168 days in the southeast, 133 days in the panhandle. The average annual precipitation varies from 33 inches (840 millimetres) in the southeast to less than 16 inches (400 millimetres) in the extreme west. Since a minimum of 20 inches (500 millimetres) is usually considered necessary for normal crop production, approximately one-half of Nebraska may be considered semi-arid. Irrigation is carried on extensively in eastern and central Nebraska, whereas the practice is essential to certain types of agriculture in the west.

Vegetation and animal life. A wide variety of grasses originally covered Nebraska's prairies, and the slopes of the river valleys were well covered with deciduous trees. At present, cottonwood, elm, and some oak and walnut are found along the bluffs of eastern Nebraska, while conifers grow in the Wild Cat and Pine Ridge highlands and the Niobrara Valley. The Nebraska National Forest in west central Nebraska resulted from a human effort to bring trees to the barren plains. Until their near-extermination at the hands of white men, bison roamed widely over the Nebraska plains. Some of these animals remain in their natural habitat on the Fort Niobrara National Wildlife Refuge, near Valentine. Antelope and deer are also native to the state, as are prairie dogs, coyotes, jackrabbits, skunks, and squirrels. Migratory birds, especially the pheasant, are common.

Patterns of settlement. Following the specifications of the Ordinance of 1785, the land was surveyed into townships six miles square and containing 36 sections; the section, comprising 640 acres (260 hectares), was the basic unit of land. To the present day Nebraska has retained this orderly gridlike survey system, which has its American roots in New England. Most of the towns and villages serving the surrounding countryside were located close to rivers, streams, and timber. Several developed as railroad terminals, but changing patterns of transportation decreed growth for some communities and stagnation or oblivion for many others.

Economics, geography, and politics created certain sectional distinctions within Nebraska. The placing of the capital of Nebraska Territory at Omaha so enraged the people south of the Platte that they sought annexation to Kansas. The animus against Omaha did not die, and in 1867 the state capital was moved to Lincoln, south of the Platte. The polarity between Omaha and "out-state" has continued, with the life-style of Omaha reflecting the boisterousness of urban commercialism, a contrast to the more sedate and university-oriented atmosphere of Lincoln and the generally conservative face of the smaller communities to the west.

Diversity of life styles

THE PEOPLE OF NEBRASKA

Ethnic groups. In addition to the native white Americans who came to Nebraska, large numbers of persons of European birth settled in the state in the late 19th century. From 1866 to 1879, the Nebraska State Board of Immigration employed an agent in Europe to recruit settlers. The Burlington and Union Pacific railroads made much greater efforts than did the state government in this direction. The largest immigrant group was the Germans, who in 1890 numbered 72,000, and the Scandinavian countries (particularly Sweden), Bohemia, and the British Isles contributed significantly to the settlement of Nebraska.

As a result of this European immigration, Nebraska was not molded in the traditional Anglo-American form that had characterized much of the nation's earlier development. Large numbers of Roman Catholics from Bohemia, Germany, and Ireland and Lutherans from Germany and Scandinavia gave diversity to the religious life of Nebraska. For many years Nebraska had a number of German-language newspapers, although the linguistic identity of the Germans and other non-English-speaking groups gradually faded away.

Negroes came to Nebraska early in the history of the state. Most settled in Omaha, which by 1900 had a black population of over 3,400 persons, a figure that by 1970 had risen tenfold. They concentrated in Omaha's near North Side, an area that increasingly became characterized by the usual social and economic problems common to the ghettos of other large cities. During the 1960s this traditional core of the black community has declined markedly in population as many Negroes moved to immediately adjacent neighbourhoods. At the same time, the usual decrease in the white population occurred in these areas. A number of racial disturbances took place during the decade, though none of them brought about comprehensive steps toward either repression or amelioration of the urban conditions.

The black community

Demography. Between 1960 and 1970 Nebraska's birth rate dropped by nearly one-third to about 17 per 1,000 population, and the death rate rose slightly. In addition, the state continued to lose more people to other states than it gained by immigration. Since the 1880s, a decade of tremendous growth during which most of the best farmlands were claimed, Nebraska has grown in population at a pace generally well below the national average.

The most striking trend in Nebraska's demography—one that is neither new nor unique to the state—has been the steady decline of the population of the rural areas and the marked growth of the cities and their suburbs. During the 1960s, 67 of the 93 counties in the state lost population. An overall 5 percent increase occurred because the growth of the urban areas more than offset the rural declines. As a result of the growth of metropolitan Omaha and Lincoln, the population centre of the state has moved steadily eastward, in 1970 lying only about 40 miles west of the Missouri.

Patterns of migration

Urban growth has been stimulated further by the mechanization of agriculture, which brought about the working of more land by fewer persons, decreases in the number of farms, and increases in average farm sizes. Similarly, most small towns, long reliant upon the local farm trade, continued to lose population, a condition undoubtedly hastened by a modern highway system that enlarged the trade areas of the cities. As employment opportunities continued to diminish in rural Nebraska, the 1960s saw an increase of nearly 30 percent in the number of work opportunities in manufacturing. The exodus from the farms and small towns may continue for some time, and Nebraska's urban areas will probably absorb much of the moving tide of Nebraskans.

THE STATE'S ECONOMY

Nebraska stood 23rd among the states in per capita personal income in 1970. The state's average of $3,700 was slightly below the national figure but above the average of the West Central states. Property income, manufacturing, wholesale and retail trades, farm proprietor's income, and state, and local governments were, in that order, the major sources of income. The total annual income of Nebraskans rose more than 50 percent from 1960 to 1968, but inflation accounted for a relatively large proportion of this.

Agriculture. Despite the fact that farm income showed the slowest growth rate among the major components of the Nebraska economy, the state remains a leading agricultural producer. In 1970 the principal crops, in order of value, were corn (maize), hay, wheat, and sorghum grain. Nebraska ranked fifth in the nation in the production of corn for feed, fifth in wheat, third in sorghum, and sixth in sugar-beet yield. In 1971 Nebraska ranked third in the number of cattle on its farms and sixth in the number of hogs and pigs.

National rankings in farm production

Manufacturing and processing. The processing of agricultural commodities is the most important of the state's industrial activities in terms of value added by manufacture. Other leading industrial activities are machinery, chemicals and allied products, printing and publishing, primary- and fabricated-metal industries, and transportation equipment.

Mineral and energy resources. The value of Nebraska's mineral output declined sharply during the 1960s. Crude petroleum accounts for almost one-half of the value of the state's mineral extraction. Nebraska also produces some natural gas and gas products, as well as significant amounts of cement, lime, pumice, sand, gravel, stone, and clay, but additional quantities of natural gas must be imported to serve the commercial, industrial, and residential needs of the state. All electrical utilities are publicly owned, and consumer rates are among the lowest in the nation.

Insurance. Nebraska, Omaha in particular, is known as one of the centres of the American insurance industry. During the 1960s the industry showed great dynamism within the state. The number of companies doubled, to 40, while the companies' policies in force rose from about $5,000,000,000 to nearly $14,000,000,000.

Public financing and economic management. In addition to local property taxes, Nebraska has state and local sales taxes, a state income tax based upon the federal income tax, and a gasoline tax, which is an important source of revenue. The state generally has been conservative in labour matters, and in 1968 the state ranked 40th, nationally, in percentage of unionized nonagricultural workers. Nebraska has a right-to-work law that forbids compulsory union membership; in recent years the time lost through work stoppages has been below the national average.

As has been the case throughout its history, Nebraska's future development will be heavily dependent upon private investment from outside its borders. A state Department of Economic Development seeks to bring new industry to Nebraska. As the problems of urban life in the great economic and industrial centres of the nation become more acute, the incentive for decentralization may work to the benefit of Nebraska.

Transportation. Nebraska is located on some of the most important arteries linking the East and West. Within the state, traffic flows toward Omaha, Lincoln, and Sioux City, Iowa, in the east and toward the cities in the Platte Valley. Much of western Iowa is within the trade area of metropolitan Omaha. A significant amount of the traffic from western Nebraska moves to Denver, Colorado.

Nebraska has modern highways that cross the state in all directions. The most important route is Interstate Highway 80, which carries heavy traffic between Iowa and the western borders of Nebraska, Colorado, and Wyoming. The state is served by five railroads. Omaha is an important port for commercial barge traffic on the Missouri. Air carriers serving Nebraska include both nationwide lines and those providing local and regional "feeder" service.

ADMINISTRATION AND SOCIAL CONDITIONS

Structures of government. *State level.* Nebraska functions under a frequently amended constitution dating from 1875. Since 1937 it has had a one-house legislature whose members are elected without political-party affiliation—the only such legislative body in the nation. The 49 members of the legislature, or "Unicameral," as it is commonly called, are popularly elected for four-year terms following primary and runoff elections in their districts, which are equally proportioned by population. Unicameral meets for sessions of 90 legislative days in odd-numbered years and 60 in even-numbered years.

The unicameral, nonpartisan legislature

The nonpartisan feature of the legislature has many critics, who charge that the lack of political parties in the Unicameral results in a lack of leadership in that body. The results of an independent study of American state legislatures released in 1971, however, ranked Nebraska's Unicameral ninth among the 50 legislatures in terms of procedural qualities. In addition, although the majority of legislators generally are Republican in political inclination, the system allows the electoral campaigns to be fought on the basis of local issues rather than party loyalties. Omaha's black community long has had representation in Unicameral.

The governor, elected for a four-year term on a partisan ballot, is the chief executive officer; his tenure is limited to two consecutive terms. The governor is responsible for the operation of 15 administrative departments and is an ex officio member of various boards and commissions. As the state's chief fiscal officer, he must present a detailed budget to the legislature, which needs an affirmative three-fifths vote to appropriate more funds than recommended by the governor or for overriding a veto. Other elected officers also run on partisan ballots.

Nebraska's court system comprises a Supreme Court with seven justices and 21 district courts with a total of 38 judges. In addition, there are two conciliation courts, the Nebraska Industrial Relations Court, a Workmen's Compensation Court, 93 county courts, and over 1,500 local justice of the peace courts. Some of the larger communities have police courts, while Omaha and Lincoln each have a municipal court and a juvenile court. Nebraska has adopted the Missouri Plan, originated in that state in an attempt to create a nonpolitical judiciary. Judicial nominating commissions, chosen by the governor and the Nebraska Bar Association, compose lists of nominees to fill vacancies on the bench. The governor then appoints one of the nominees to fill a particular position. Nebraska also has a merit system under which the electorate determines whether judges will be retained in office.

The nonpartisan judicial system

Local government and political life. County government is vested in boards of supervisors or commissioners of from three to seven members, who like other county officials are elected on partisan ballots. The city manager and mayor–council forms of government are used in Nebraska's cities, and governmental authority in villages is vested in elective boards of trustees.

The Territory of Nebraska was the creation of a Democratic administration in Washington, D.C., and Democrats dominated Nebraska politics until the Civil War. The 30 years after 1860 was a period of Republican pre-eminence in Nebraska, but the political ferment during and after the 1890s brought an end to one-party rule. Although a slight majority of Nebraska's voters are registered Republicans, Democrats often are elected to office.

The social milieu. The social and economic problems of Nebraska, while manageable, are pressing. Better housing, improved educational and employment opportunities, and an upgrading of police–community relations are fundamental necessities in Omaha's black community. An active civil rights movement has pressed for these improvements in the conditions of life. In certain rural counties as well, the level of income is far below state and national averages. Overall wage levels in Nebraska are slightly below the national average, yet inflation and unemployment have been less pronounced than in many other parts of the country.

Health and welfare. Nebraska maintains programs of public assistance: medical aid and aid for dependent children, the aged, the blind, and other disabled persons. Federal funding in 1970 provided over 60 percent of Nebraska's public-assistance expenditures, compared to the national average of about 50 percent. Although federal, state, and county welfare funding has steadily increased, the average monthly public-assistance payments in Nebraska are below the national averages. Nebraska also has participated in various antipoverty operations of the federal Office of Economic Opportunity.

The state maintains a system of mental hospitals and other specialized health, correctional, and care facilities. Health conditions in Nebraska are generally sound. Although expanded medical facilities are needed in some rural areas, Omaha ranks as a medical centre of national significance.

Education. Nebraska's public elementary and secondary schools and ten vocational–technical schools had a per pupil expenditure of $682 in 1971, compared with the national average of $867. Since 1967 an increasing

amount of direct state aid has supplemented local appropriations. Beginning in 1949, the number of school districts was cut drastically to offer more efficient educational facilities and programs.

Higher education

Of the institutions of higher education in Nebraska, 10 are private, five are state operated, and six are public junior colleges. The University of Nebraska (established 1869) is the largest in the state and is composed of three semi-autonomous campuses, the main one in Lincoln and two, including the medical school, in Omaha. Graduate-degree programs are offered at the University of Nebraska and Creighton University (Omaha), both of which offer programs in medicine, law, and dentistry, and at state colleges in Chadron, Kearney, and Wayne.

CULTURAL LIFE AND INSTITUTIONS

The arts and historical preservation. In less than two generations, Nebraska was converted from a wilderness inhabited by a small number of native Indians to a settled commonwealth of over 1,000,000 persons. This conquest of space was one of the most important achievements of the 19th century, and it is natural that the cultural contributions of Nebraska, like those of the other western states, would focus upon this frontier experience.

Such Nebraska novelists as Willa Cather, Mari Sandoz, and Bess Streeter Aldrich were among those who wrote perceptively of life on the plains and won national audiences. The relationship of the pioneers to a capricious natural environment, the life styles and interaction of settlers of diverse social and ethnic backgrounds, and the plight of the Indians were important themes of these writers. The epic poetry of John G. Neihardt spoke feelingly of plains Indian life and re-created the adventures of the explorers of the 19th-century West. In the early 1970s his poetic narrative *Black Elk Speaks* achieved national recognition some 40 years after its original publication.

The Nebraska State Historical Society, organized in 1878, continues to make important contributions to an understanding of life in Nebraska and the West. In 1960 the University of Nebraska Press launched the paperback Bison Series, reprints of early and modern works on the American West, including histories, collections of lore from Indian and white settlers, and other important documents, many of which had been long out of print.

Increased urbanization and affluence have promoted continued artistic expression in Nebraska. The Joslyn Art Museum in Omaha and the Sheldon Memorial Art Gallery in Lincoln are the state's major collections in the visual arts. Omaha and Lincoln have theatrical groups and symphony orchestras that enrich the life of the state.

Folk and historical festivals

Various folk observances, such as the Czech Festival at Wilber, are periodic reminders of the diverse origins of the people of Nebraska. Ogallala, a roaring cow town in the 1870s and 1880s, relives its colourful past with its "Front Street" festivities held each summer. Each October, the Knights of Ak-Sar-Ben (Nebraska spelled in reverse), an Omaha civic organization founded in 1895, crown a king and queen of Quivira. This elaborate social event commemorates the search through the plains in 1541 of the Spanish explorer Francisco Vásquez de Coronado for the legendary Seven Cities of Cibola in the Kingdom of Quivira. Over the years the University of Nebraska football team has attained national prominence and no single subject holds the attention of Nebraskans as does the plight of the "Cornhuskers." The benefits derived from football enthusiasm are significant in several areas of the state's economy.

Communications. Journalism has always been important to the cultural and political development of Nebraska. The state has about 50 newspapers, with a circulation of over 3,000 copies each; of these, 18 are dailies. The two middle-of-the-road Lincoln dailies have a wide circulation in southeastern and south central Nebraska, while the generally conservative *Omaha World Herald* is read throughout Nebraska and southwestern Iowa. The state is served by 14 commercial television stations, while the Nebraska Educational Television Network, a noncommercial system operated by the University of Nebraska, serves the schools and general public across the state. There are more than 60 standard broadcasting stations in Nebraska.

Prospects. Agriculture remains vital, either directly or indirectly, to the livelihood of Nebraska. The agrarian stereotype of the state is increasingly invalid, yet Nebraska's large towns and cities maintain a certain closeness to the land, in spirit as well as in economic and geographic fact. Although Nebraska inevitably will become more urbanized, this physical and spiritual proximity to the land may enable the state to minimize the problems of city life. It is likely, also, that the urban experience will propel the people of Nebraska toward newer themes of cultural expression.

BIBLIOGRAPHY. JAMES C. OLSON, *History of Nebraska*, 2nd ed. (1966), is an excellent general historical work, including a useful bibliography. The NEBRASKA STATE HISTORICAL SOCIETY, *Proceedings and Collections*, 25 vol., and *Nebraska History*, a quarterly magazine, are essential to an understanding of the history of the state. The NEBRASKA LEGISLATIVE COUNCIL, *Nebraska Blue Book and Historical Register* (published biennially since 1915), is a useful compendium of information on government. The NEBRASKA DEPARTMENT OF ECONOMIC DEVELOPMENT, *Nebraska Statistical Handbook* (1970), has information on economic and other aspects of Nebraska life.

(H.A.D.)

Nebuchadrezzar II

Nebuchadrezzar, or Nebuchadnezzar, II was the oldest son and successor of Nabopolassar, founder of the Chaldean (or Neo-Babylonian) Empire, and the greatest king of his dynasty. He was celebrated for his military might; the splendour of his capital, Babylon; and for his important role in Jewish history. He is known from cuneiform inscriptions, the Bible and later Jewish sources, and from classical authors. His name, from the Akkadian Nabu-kudurri-usur, means "O Nabu, protect my boundary stone," "boundary stone" connoting line of succession.

While his father disclaimed royal descent, Nebuchadrezzar claimed the third-millennium ruler Naram-Sin as ancestor. The year of his birth is uncertain, but it is not likely to have been before 630 BC, for according to tradition Nebuchadrezzar began his military career as a young man, appearing as a military administrator by 610. He is first mentioned by his father as working as a labourer in the restoration of the temple of Marduk, the chief god of the city of Babylon and the national god of Babylonia.

Service as a military commander

In 607/606, as crown prince, he commanded an army with his father in the mountains north of Assyria, subsequently leading independent operations after Nabopolassar's return to Babylon. After a Babylonian reverse at the hands of Egypt in 606/605, he served as commander in chief in his father's place, and by brilliant generalship he shattered the Egyptian army at Carchemish and Hamath, thereby securing control of all Syria. After his father's death on August 16, 605, Nebuchadrezzar returned to Babylon and ascended the throne within three weeks. This rapid consolidation of his accession and the fact that he could return to Syria shortly afterward reflected his strong grip on the empire.

On expeditions in Syria and Palestine from June to December of 604, he received the submission of local states, including Judah, and captured the city of Ashkelon. With Greek mercenaries in his armies, further campaigns to extend Babylonian control in Palestine followed in the three succeeding years. On the last occasion (601/600), Nebuchadrezzar clashed with an Egyptian army, with heavy losses; this reverse was followed by the defection of certain vassal states, Judah among them. This brought an intermission in the series of annual campaigns in 600/599, while Nebuchadrezzar remained in Babylonia repairing his losses of chariots. Measures to regain control were resumed at the end of 599/598 (December to March). Nebuchadrezzar's strategic planning appeared in his attack on the Arab tribes of northwest Arabia, in preparation for the occupation of Judah. He attacked Judah a year later and captured Jerusalem on March 16, 597 BC, deporting King Jehoiachin to Babylon. After a further brief Syrian campaign in 596/595, Nebuchadrezzar had to act in east Babylonia to repel a threatened invasion, probably from Elam (modern southwest Iran).

Tensions in Babylonia were revealed by a rebellion late in 595/594 involving elements of the army, but Nebuchadrezzar was able to put this down decisively enough to undertake two further campaigns in Syria during 594.

His further military activities are known not from extant chronicles but from other sources, particularly the Bible, which records another attack on Jerusalem and a siege of Tyre (lasting 13 years, according to the Jewish historian Flavius Josephus) and hints at an invasion of Egypt. The siege of Jerusalem ended in its capture in August 586 and in the deportation of prominent citizens, with a further deportation in 582. In this respect Nebuchadrezzar followed the methods of his Assyrian predecessors.

Much influenced by the Assyrian imperial tradition, Nebuchadrezzar consciously pursued a policy of expansion, claiming the grant of universal kingship by Marduk and praying to have "no opponent from horizon to sky." From cuneiform fragments he is known to have attempted the invasion of Egypt, the culmination of his expansionist policy, in 568/567.

In addition to being a brilliant tactician and strategist, Nebuchadrezzar was also prominent in international diplomacy, as shown in his sending an ambassador (probably Nabonidus, a successor) to mediate between the Medes and Lydians in Asia Minor. He died in August or September 562 and was succeeded by his son Amel-Marduk (Evil-Merodach of II Kings).

The rebuilding of Babylon

Nebuchadrezzar's main activity, other than as military commander, was the rebuilding of Babylon. He completed and extended fortifications begun by his father, built a great moat and a new outer defense wall, paved the ceremonial Processional Way with limestone, rebuilt and embellished the principal temples, and cut canals. This he did not only for his own glorification but also in honour of the gods. He claimed to be "the one who set in the mouth of the people reverence for the great gods" and disparaged predecessors who had built palaces elsewhere than at Babylon and had only journeyed there for the New Year Feast.

Little is known of his family life beyond the tradition that he married a Median princess, whose yearning for her native terrain he sought to ease by creating gardens simulating hills. A structure representing these hanging gardens cannot be positively identified in either the cuneiform texts or the archaeological remains.

Assessment

Despite the fateful part he played in Judah's history, Nebuchadrezzar is seen in Jewish tradition in a predominantly favourable light. It was claimed that he gave orders for the protection of Jeremiah, who regarded him as God's appointed instrument whom it was impiety to disobey, and the prophet Ezekiel expressed a similar view at the attack on Tyre. A corresponding attitude to Nebuchadrezzar, as God's instrument against wrongdoers, occurs in I Esdras and, as protector to be prayed for, in Baruch. In Daniel and in Bel and the Dragon, Nebuchadrezzar appears as a man, initially deceived by bad advisers, who welcomes the situation in which truth is triumphant and God is vindicated.

There is no independent support for the tradition in Daniel of Nebuchadrezzar's seven years' madness, and the story probably arose from a fanciful later interpretation of texts concerned with events under Nabonidus, who showed apparent eccentricity in deserting Babylon for a decade to live in Arabia.

In modern times Nebuchadrezzar has been treated as the type of godless conqueror; Napoleon was compared to him. The story of Nebuchadrezzar is the basis of Giuseppe Verdi's opera *Nabucco*, while his supposed madness is the theme of William Blake's picture "Nebuchadnezzar."

BIBLIOGRAPHY. There is no standard biography. Most of Nebuchadrezzar's building inscriptions, containing some historical material, are edited in STEPHEN LANGDON, *Die neubabylonischen Königsinschriften*, pp. 70–209 (1912). Relevant chronicles are edited in D.J. WISEMAN, *Chronicles of Chaldaean Kings (626–556 B.C.) in the British Museum* (1956). The principal classical references to Nebuchadrezzar are contained in I.P. CORY, *Ancient Fragments*, 2nd ed. (1832).

(H.W.F.S.)

Nebula

"Nebula" (Latin: "mist") is a term formerly used in astronomy to denote any object situated beyond the solar system that is seen as a bright or dark area, in contrast to the pointlike images of stars. This definition, adopted at a time when very distant objects could not be resolved into great detail, unfortunately includes two unrelated classes of objects: the extragalactic nebulae, now called galaxies, which are huge collections of stars and gas; and the galactic nebulae, which are much smaller clouds of gas (with traces of solid particles) within a single galaxy. The term nebula is now restricted to the latter class of object. The galactic nebulae collectively amount to only a few percent of the mass of their galaxy, the rest being made up by the stars within the galaxy. Galactic nebulae are also called gaseous nebulae. The present article will be concerned only with the galactic nebulae.

CLASSES OF NEBULAE

The galactic nebulae are divided into two subclasses on the basis of their appearance: (1) dark nebulae appear as black patches in the sky that are usually very irregular in shape, blotting out the light of the stars beyond them; (2) bright nebulae appear as faintly luminous, glowing surfaces, either emitting their own light or reflecting the light of nearby stars.

The bright nebulae can be subdivided into four classes on the basis of their origin and the details of their appearance: (*a*) diffuse nebulae, which are usually faint and irregular in shape and produce the radiation they emit; (*b*) reflection nebulae, which reflect light from a nearby bright star; (*c*) planetary nebulae, which typically appear as rather symmetrical round disks or rings of fairly high surface brightness (the name arises from the fact that many of them appear rather like out-of-focus images of planets seen through a telescope of poor resolution); and (*d*) supernova remnants, which are clouds of gas expanding away from the site of a stellar explosion called a supernova, which disrupts the star and ejects most of its mass at speeds of thousands of miles per second.

Densities in nebulae

The diffuse nebulae, reflection nebulae, and dark nebulae are basically the same type of physical entity, seen under different conditions. All are irregular clouds of gas, typically several light-years across but with a density of ten to 1,000 atoms per cubic inch, which is 1,000 times less dense than an excellent laboratory vacuum. About 1 percent of their mass is in the form of "dust," or small, solid particles that are efficient in absorbing and scattering starlight. When the gas cloud happens to be near one of the massive, hot O-type stars or very close to a slightly cooler B star, it glows because these stars, with temperatures five to ten times that of the Sun, emit enough ultraviolet radiation to ionize the gas in the cloud; the ionized gas then emits the radiation seen in the visible part of the spectrum. (An ion is an atom that has lost one or more of its electrons.) If the nebula is close to a cooler star that does not emit enough ultraviolet radiation to ionize the gas, the starlight may still be partially reflected from the dust grains, and the cloud is then seen as a reflection nebula. Finally, if the cloud lies between the observer and more distant stars, the dust either scatters the starlight into other directions or absorbs it, in either case reducing the amount of starlight that comes to the observer directly. Therefore, the starlight is dimmed, and, when the cloud is extensive enough, it is seen as a gap in the star field; that is, as a dark nebula. Thus, the same cloud could appear to hypothetical observers at different points in space as a reflection nebula (if viewed with a star near to it but in front of or beside it) or a dark nebula (if viewed with the same star behind it, so that it blocks the starlight).

Sizes of nebulae

Nebulae differ greatly in size. The only one easily visible to the naked eye is the beautiful Orion Nebula (see Plate I), which is seen as the central star in Orion's sword. It is the brightest and best studied diffuse nebula. Some nebulae are extremely large; the largest, the Gum Nebula (named after its discoverer, the Australian astronomer Colin S. Gum), is 40° in angular diameter. The whole constellation of Orion is embedded in extremely faint diffuse nebulosity that cannot be detected by eye.

Such large nebulae are so faint that they require special techniques (see below *Advances brought by photography*) to detect them. Most supernova remnants are similar to these large, faint diffuse nebulae.

The planetary nebulae are all telescopic objects. The largest known, the Helix Nebula, in the constellation Aquarius, subtends an angle of about 20′ of arc—two-thirds of the angular size of the Full Moon; such large nebulae, however, have very low surface brightness and are not easily seen. Most bright planetary nebulae are less than 0.5′ of arc in size—that is, only 0.02 of the angular diameter of the Moon. The Crab Nebula, the best known supernova remnant, is slightly larger—approximately 3′ of arc.

PRE-20TH-CENTURY OBSERVATIONS OF NEBULAE

Naked-eye observations of "cloudy stars" were mentioned by the Greek astronomers Hipparchus (*c.* 190–125 BC) and Ptolemy (*c.* AD 138), but these objects are known to have been compact clusters of stars. In 1610, two years after the invention of the telescope, the Orion Nebula, which looks like a star to the naked eye, was discovered by the French scholar and naturalist Nicolas-Claude Fabri de Peiresc. In 1656 Christian Huyghens, the Dutch scholar and scientist, using his own greatly superior instruments, was the first to describe the bright inner region of the nebula and to determine that its inner star is not single but a compact quadruple system —the well-known "Trapezium." The inner regions of the nebula are still referred to by his name.

Early-18th-century observational astronomers gave high priority to comet seeking. A by-product of their search was the discovery of many bright nebulae. The first catalog of nebulae was compiled by Edmund Halley, second astronomer royal, whose name was given to the great comet whose appearances he correctly predicted. He included in his catalog six nebulae, which he correctly surmised were composed of a "lucid medium shining with its own proper lustre" and "filling space immensely great." Several catalogs of special objects were subsequently compiled by other comet researchers; by far the best known is that of the Frenchman Charles Messier, who, in 1781, compiled a catalog of 103 nebulous, or extended, objects to avoid their confusion with comets. Most are clusters of stars, 35 are galaxies, and 11 are nebulae. Even today many of these objects are commonly referred to by their Messier catalog number: M 20, for instance, is the great Trifid Nebula, in the constellation Sagittarius (see Plate I).

The work of the Herschels. By far the greatest observers of the early and middle 19th century were the English astronomers William Herschel and his son John. Between 1786 and 1802 William Herschel, aided by his sister Caroline, compiled three catalogs totalling about 2,500 clusters, nebulae, and galaxies. He also observed patchy areas in which there appeared extremely few stars even in the midst of the brilliant star clouds of the Milky Way. He tentatively but incorrectly interpreted these dark nebulae, which were obscuring the stars behind them, as holes in the distribution of stars through which it is possible to see the empty distances of space. John Herschel later added to the catalogs 1,700 other nebulous objects in the southern sky visible from the Cape Observatory in South Africa but not from London and 500 more objects in the northern sky visible from England.

The origin of NGC and IC numbers

The catalogs of the Herschels formed the basis for the great *New General Catalogue* (NGC) of J.L. Dreyer, published in 1888. It contains the location and a brief description of 7,840 nebulae, galaxies, and clusters. In 1895 and 1908 it was supplemented by two *Index Catalogues* (IC) of 5,386 additional objects. The list still included galaxies as well as true nebulae, for they were often at this time still indistinguishable. Most of the brighter galaxies are still identified by their NGC or IC numbers according to their listing in the *New General Catalogue* or *Index Catalogues*.

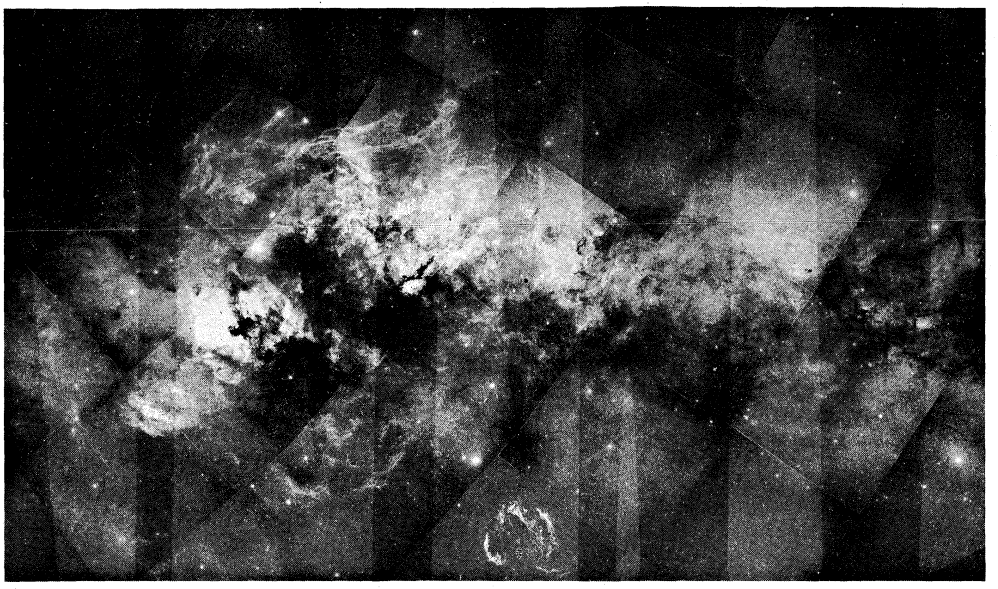
Advances brought by photography. The advent of photography, which allows the recording of faint details invisible to the naked eye and provides a permanent record of the observation for study of fine details at leisure, caused a revolution in the understanding of nebulae.

In 1880 the first photograph of the Orion Nebula was made, but really good ones were not obtained until 1883; these early photographs showed a wealth of detail extending out to distances unsuspected by visual observers. Around 1885 the use of modestly sized but short-focal-length ("fast") lenses, instead of the long-focus instruments used previously, greatly increased astronomers' ability to photograph nebulae. The long-focus instrument is useful if a large-scale photograph is wanted, because it spreads the light from a given area in the sky into a large area of photographic plate, widely separating stellar images. For nebular photography, however, the nebular light is also spread over a wide area and is consequently fainter if photographed with a long-focus lens rather than a short-focus lens of the same aperture. Furthermore, the short-focal-length lenses can be made to cover a much wider area of the sky in one photograph than the long-focus instrument can, especially if the lens consists of several elements to reduce aberrations. In 1885 the faint reflection nebulosity in the Pleiades star cluster was photographed for the first time, and in 1889 the full extent of the Pleiades nebulosity was photographed with a wide-angle camera at Lick Observatory. Later the sky was systematically photographed and many nebulae recorded, especially in the Milky Way (see illustration).

The study of nebular spectra. The physical nature of an astronomical object can be discovered in detail by studying its spectrum; that is, the resolution of its light into different wavelengths (or colours). Study of the spectrum of a cloudy-looking object provides a decisive test as to whether it is composed of unresolved stars (as are galaxies) or glowing gas, because the natures of the spectra of stars and of gas are quite different. According to the laws of radiation as stated by the chemist Robert Bunsen and the physicist Gustav Kirchoff, both German, in 1859, opaque, hot bodies (such as stars) should radiate continuously at all wavelengths, with perhaps dark absorption lines superimposed, while hot, transparent gas clouds should radiate only emission lines at certain wavelengths, or colours, characteristic of their constituent gases. In 1864 observation of the spectrum of the Orion Nebula showed bright emission lines of glowing gases, with conspicuous hydrogen lines and some green lines even brighter. The spectrum of galaxies was found to be stellar, so that a distinction between galaxies and nebulae—that nebulae are gaseous and galaxies are stellar—was appreciated at this time, although the true sizes and distances of galaxies were not demonstrated until the 20th century.

TWENTIETH-CENTURY DISCOVERIES

Observationally, there have been at least four major developments in the 20th century. (1) Very fast, wide-angle telescopes of excellent resolution, patterned after a type invented by the German optical worker Bernhard Schmidt in 1930, have made it possible to detect very faint nebulae; increased speed of available photographic plates has also been important in this regard. (2) Other powerful observational techniques using photoelectric devices (in which the light liberates electrons from specially sensitized materials) have been devised; the principles are discussed below. These techniques can be more than 100 times more sensitive than photography. They also allow quantitative measurements of the faint nebular spectrum. (3) Better methods for obtaining spectra have been devised. These include the use of interference filters that reject all colours but those in a very narrow spectral range, or of Fabry-Perot or Fourier spectrometers. (4) Spectral observation has been extended to wavelength ranges inaccessible to the eye or photographic plates. The most important new range is the radio region, where waves are relatively unaffected by the absorption of cosmic dust. The infrared region (wavelengths longer than the visible) has also become accessible; many of the important emission lines are in this region, and their study will become increasingly important. So far, the ultraviolet (short wavelength) region of the spectrum has not been of great value in the study of nebulae.



Red-light photomosaic shows many star clouds, glowing gas, and dark dust in the Milky Way, near the galactic plane. The brightest object (centre left) is the North American Nebula; the loop-shaped object (bottom) is the Cygnus Loop. The rich pattern of bright and dark nebulae, stars, and star clouds is due to the direction of the line of sight, in this region nearly parallel to the outer spiral arms of the Galaxy. (Photo taken by 48 inch Schmidt telescope; photograph, The Hale Observatories.)
By courtesy of © National Geographic Society—Palomar Observatory Sky Survey

During the 20th century scientists have been able to acquire a reasonably good theoretical understanding of the nebulae, although challenging and exciting problems still remain. During the 1920s astrophysics received added impetus from the detailed quantum-mechanical study of atoms (their emission and absorption processes and spectra), and this was applied to the study of nebulae. During the 1930s and 1940s the ionization and heating of nebulae were studied, on the assumption that motions are negligible. During the 1950s problems of gas dynamics were considered, and since about 1960 the use of high-speed electronic computers has made it possible to take rather accurate account of ionization processes, heating and cooling, and even (to a limited extent) the dynamics of the resultant nebular motion. The agreement of these models with observations is encouraging.

Modern observations. Two types of telescopes, which allow complementary types of observations, are used for nebular photography. To photograph the finest details possible, a conventional parabolic reflector is used, with a special additional lens to correct the worst image defect, coma. Even with the correcting lens, the usable field for the 200-inch telescope is very small, about 10′ of arc—which is about twice the angular diameter of a pinhead held at arm's length. The size of the smallest visible details is set by the degree of steadiness of the air above the telescope. Under the best conditions, one can observe details as small as 0.5″ of arc—the width of the pinhead as seen from 500 metres away but still larger than our solar system at the distance of even the nearest nebula.

Devices for the photography of very fine details

The structure of large nebulae. The structure of large nebulae, covering many angular minutes or degrees in size, can be best revealed in photographs made with a Schmidt-type telescope. This has extremely good definition over an area of several (usually about 5 × 5) square degrees, which is about the angular size of a playing card held at arm's length, or ten times the angular dimensions of the Moon.

Spectrographic studies

A great deal of information may sometimes be obtained about an object by photographing it in a limited wavelength region; for example, by cutting out the light beyond the desired region by filters. Some filters can be used to photograph nebulae in the light of only one emission line while rejecting all others. The appearance of the nebula in the light of each emission line indicates how the ions giving rise to that line are distributed throughout the nebula. Another device, the Fabry-Perot etalon, makes it possible to calculate the speeds of the gas inside the nebula in the line of sight of the observer by measuring the Doppler shift of the nebular emission lines. The wavelength shift of the centre of the emission lines determines the average speed of the gas in the nebula, while the width of the line indicates how much difference in velocity there is in material along the line of sight.

The spectra of nebulae are usually obtained with spectrographs of design similar to those used for stars. For large, very faint nebulae, however, it is more useful to be able to receive light from a large portion of the sky, whereas for stars the light comes from a point of the sky. The so-called nebular spectrographs are very fast cameras (*i.e.*, with focal length as little as 0.65 times the aperture, or f/0.65) that photograph the spectrum produced by a large diffraction grating.

Small planetary nebulae are most easily discovered from photographs made with an objective prism placed in front of the optics of a Schmidt telescope. This arrangement yields low-dispersion spectra of all objects appearing on the plate. Stars appear as streaks because their spectra are continuous, but nebulae show up as dots because much of their energy is concentrated in a few colours, especially the red emission lines of hydrogen and nitrogen near 6563 angstroms and of hydrogen and oxygen near 5000 angstroms.

Physical conditions in nebulae. There are several different types of nonreflecting bright nebulae (diffuse, planetary, supernovae remnants), but all of their basic physical processes are similar. These processes must be understood in order to understand the physical nature of the nebulae (for individual characteristics, as well as the origin and evolution of each nebular type, see below).

The composition of cosmic matter

Analysis of cosmic matter shows almost a uniform chemical composition, regardless of the object. This material is predominately hydrogen (about 65 percent by weight, or 90 percent by number of atoms). There is about 3 percent by weight of heavy elements—*i.e.*, of all elements heavier than hydrogen and helium. By number of atoms these heavy elements make up only 0.1 percent of the atoms. The rest—about 30 percent by weight, 10 percent by number of atoms—is helium, which on Earth is one of the rare, or noble, gases.

The diffuse nebulae, planetary nebulae, or supernovae remnants consist of such material in which every atom has been ionized; that is, had one or more electrons stripped from it. The other types of nebulae—dark and

reflection—are composed of almost completely neutral, or nonionized gases. The reason for this separation into two classes, one almost completely ionized and the other almost completely neutral, can be seen from consideration of a pure hydrogen cloud, because the basic idea applies equally well to a real nebula. The hydrogen in the cloud can be ionized by ultraviolet light at a wavelength of less than 912 angstroms or, equivalently, a photon energy of more than 13.6 electron volts. Such radiation, if produced by a star inside a cloud of gas, will ionize a region around the star. At a typical point within the ionized region, only one hydrogen atom in 10,000 is neutral. Since only neutral hydrogen can absorb the radiation, the absorption, once the gas has been ionized, is rather low, just enough to offset the "recombinations" (or captures of a free electron by an ion, resulting again in a neutral atom that must eventually be re-ionized). The intensity of ionizing radiation decreases with distance from the source, and the fraction of neutral material therefore increases. But the increasing fraction of neutral material absorbs the remaining radiation: when the fraction of neutral atoms has increased to 1 percent, the absorption is 100 times as large as it was when the matter was only 0.01 percent neutral. The remaining radiation is very quickly absorbed. Thus, the edge to the ionized region is very sharp, and one can divide the gas into almost completely ionized or almost completely neutral regions.

When several species of atoms are present, it is still found that the ionized region has a sharp boundary, because hydrogen and helium behave similarly and together make up 99.9 percent of the atoms of cosmic bodies. There is, in a sense, a competition for the ultraviolet radiation among the various species of ions and atoms, and a balance is quickly reached in which each stage of ionization is destroyed at the same rate as it is produced. A given stage of ionization is produced by (1) recombination (electron capture) from the next higher stage of ionization and (2) ionization (by absorption of radiation) of the next lower stage of ionization. It is destroyed by either the capture of an electron or the absorption of a photon so that the ionization goes a step further. The exact composition of the gas, for a given radiation field, is completely determined by the balance of these processes.

Physics of the nebular spectrum. The simplest atom, hydrogen, consists of a positively charged nucleus encircled by a negatively charged electron. The electron is attracted to the nucleus because of its opposite charge and therefore has a binding energy. One of the rather astonishing results of quantum mechanics, or the laws of atomic behaviour, is that only certain energies characteristic of the type of atom are available to the bound electron. The levels of hydrogen are designated by a quantum number, n. The higher levels are much more closely spaced than the lower ones. The lowest energy level is called the ground level; higher levels are said to be excited. An atom can change its level; if it radiates energy, it emits a quantum or photon of light and drops to a lower energy level, such that the photon energy exactly matches the difference in energy of the initial and final atomic levels. According to Planck's law of radiation, the energy of the photon is related to its wavelength, so that when an atom drops from one level to another a photon of a characteristic wavelength is emitted. Hα, the red hydrogen line at 6563 angstroms, corresponds to a transition from level 3 to level 2; the blue-green Hβ, 4 to 2. All of the hydrogen transitions ending on the level 2 are called the Balmer series and are the only hydrogen lines in the visible region of the spectrum. The Lyman lines ending on the ground level, n = 1, are stronger but are all at short wavelengths (1215 angstroms or less) in the ultraviolet absorbed by the Earth's atmosphere.

Energy levels of atoms

Not all changes in energy levels are downward to lower energies: an atom can absorb a photon, using the photon's energy to move the electron into a larger orbit. If the absorbed photon is energetic enough, the electron may even be removed completely, and the atom is ionized. The energy required to ionize the ground level is called the ionization potential of the atom.

Besides emitting and absorbing photons, the atom can change its energy because of its close approach to another atom, ion, or electron when an exchange of energy may take place. Most such "collisions" do not change the energy level of either particle and are "elastic"; that is, the particles have the same kinetic energy after the collision that they did before. Occasionally, however, the collision can either raise the energy of the atom to a higher energy level, at the expense of the kinetic energy (an inelastic collision), or cause the atom to drop to a lower energy level, increasing the kinetic energy (a superelastic collision).

Collisions between atoms and other particles

The spectrum of the nebulae can be understood in terms of such transitions between the energy levels of the atoms in the nebula.

Recombination. The hydrogen emission lines in nebulae are produced by the hydrogen nucleus (proton) passing near enough to an electron to recapture it into some bound energy level, emitting a photon in the process. Since the free electron can have any energy above the ionization energy, the photon produced can have any energy, from the ionization energy of the level upward. This gives rise to a continuum of radiation.

The atom is left neutral but, in general, is in an excited state after the recapture. It then drops down to lower energy levels, remaining for only a few hundred-millionths of a second in each level. Most of the nebulae are transparent to the emitted photons except those corresponding to transitions to the lowest, or ground, level; hence, the emitted photon usually escapes. The nebula is opaque to photons matching transitions of the Lyman series (n = 1) of hydrogen because almost all of the neutral atoms are in the ground level (they cascade down into it from excited states almost instantaneously), and the photon has exactly the right energy to be absorbed by that level. The result is that the photon can rarely escape, and the nebula is opaque to the Lyman-series lines. The probability of the hydrogen atom emitting the various lines can be calculated from atomic theory; the results agree very well with the observations of actual nebulae.

Since the higher levels are closely spaced, the transitions between two high levels have low energies and correspondingly long wavelengths that are usually in the radio region. Here, there is some action of the nebula to amplify the emission by "stimulated emission," which is the action of the laser or maser. One of the commonly observed radio lines is between the 110th and 109th levels. It has a wavelength of 2.4 inches (six centimetres).

After an atom has regained its electrons through recombination and cascaded down to the ground state, it remains in the ground state until it absorbs another ionizing photon and is re-ionized. In a typical diffuse nebula this takes about one year. It then remains ionized for about 10,000 years, on the average, before recombining again. During the time it is ionized, the ion undergoes elastic encounters with electrons every few hours as it wanders through the vast spaces of the nebula at a speed of seven miles (ten kilometres) per second.

All elements undergo recombination and ionization, but, of course, the strength of the lines emitted varies with the atomic abundance. Since hydrogen dominates the atomic abundances, it has by far the strongest recombination lines. The helium lines, however, are easily observed. There are weak recombination lines of carbon and oxygen detected.

Collisionally excited lines. The strongest lines in most nebulae are green lines; they are not emitted from hydrogen and are never observed in the laboratory. When first discovered, they were attributed to a hypothetical element, "nebulium," but it was known that there was no place for another element in the periodic table and that nebulium must be a known element in an unusual state of ionization or excitation. In 1927 the U.S. astronomer Ira Sprague Bowen identified nebulium as doubly ionized oxygen, which can be produced in the laboratory. The green nebular lines are not observed terrestrially because they are "forbidden"; that is, the transition from the upper level to the lower level that produces the emission requires a long time to take place. In the laboratory the

"Forbidden" lines

excited atom hits another particle or the container walls before it emits the photon. In the nebulae the atom remains undisturbed long enough to emit the photon.

The doubly ionized oxygen ion enters the upper level of the transition through collisions with electrons, typically undergone about once every year for each ion. This is a comparatively rapid rate. In comparison, ionized hydrogen takes about 10,000 years to recombine and emit a photon at a hydrogen-line wavelength. The "rapid" rate of collisional excitation of the oxygen more than makes up for the fact that its abundance is lower than hydrogen by a factor of about 1,000.

The hydrogen lines are not produced by collisions, because the energy-level spacing is different in hydrogen and doubly ionized oxygen. Practically no electrons have enough energy to excite the hydrogen atom. In contrast, doubly ionized oxygen and many other heavy ions have low-lying levels that can be excited by the electrons.

In nebulae the forbidden lines provide an important means of measuring the density and temperature. The rate at which collisions excite the various upper levels depends on the temperature, which can be determined by comparing the strengths of lines arising from different upper levels of the same atom. The derived values are 6,000°–9,000° K (0° K = −273° C or −460° F) for diffuse nebulae and 7,000°–17,000° K for planetary nebulae. Similarly, for rather high-density objects the densities can be found by comparing two forbidden lines that are collisionally de-excited at different rates. The strongest lines of singly ionized oxygen, at 3726 and 3729 angstroms, are usually used for this purpose. The densities derived for the centre of the Orion Nebula are about 10,000 atoms per cubic centimetre (160,000 atoms per cubic inch).

The nebular continuum. In addition to the emission lines, there is a rather faint continuous background spectrum of continuum. It has already been mentioned that recombination to a bound level can give rise to some of it. Another cause is the bremsstrahlung, radiation emitted by free electrons as they are accelerated by passing near ions. This is the strongest source of nebular continuous radiation in the radio-wavelength region. For short radio wavelengths the radiation escapes from the nebula, but for longer wavelengths the nebula absorbs the radiation, and only the radiation emitted close enough to the edge to escape can be observed.

In the diffuse nebulae there is also a continuum contribution from starlight reflected from dust grains within the nebula. Thus, a diffuse nebula may have a weak reflection component, but its own line emission greatly dominates the light produced.

The temperature of nebulae. Nebulae differ from one another by factors of hundreds in mass, size, and density; their sources of ionization vary from stars whose temperatures are 30,000° K (B-type stars, for some diffuse nebulae) to the central stars of small, peculiar planetary nebula that may be as high as perhaps 150,000° K. Yet practically all nebular temperatures fall in the range from about 6,000° to about 12,000° K. To see why the temperature is kept within so small a range, it may be recalled that the temperature is a measure of the kinetic energy of the free electrons; hence, it is fixed by the balance of the energy losses and gains of the electrons. An energy gain occurs when an atom absorbs an energetic ionizing photon. The electron is ejected with a kinetic energy equal to the excess of the photon energy over the ionization energy of the atom—often several electron volts. The electrons constantly interact with one another (and with ions) through their electric forces. The fast electrons from ionizations quickly share their energy with the many lower speed neighbouring electrons, which have a definite distribution of speeds set up by the collisions. This distribution is characterized by the temperature. On the other hand, there are also processes removing kinetic energy from the collection of electrons; the main one is inelastic collisions of electrons with the heavy ions (usually O^{2+}) that excite the ion while the colliding electron is left with less energy after the collision than it had before. Recall that this process results in the emission of a forbidden-line photon that escapes the nebula; the energy carried away by this photon was provided by the electrons.

The reason that nebulae have narrow ranges of temperatures is that the cooling process is very inefficient below about 5,000° K; not enough energetic electrons exist to produce forbidden lines. Hence, only a small amount of heating suffices to bring the temperature up to this level. Above this value the cooling increases very rapidly, and a great deal of heating is needed to bring the temperature much above 10,000° K. The heavy ions of the nebula act as a thermostat set on the rather formidable temperature of 8,000° K.

DIFFUSE, DARK, AND REFLECTION NEBULAE

Forms taken by diffuse nebulae. The diffuse nebulae are clouds of gas and dust, ionized by one or more young, massive stars (the "exciting" stars) within or near them. Their sizes and masses vary enormously; there is really no lower limit to their size, since a very faint nebula must exist around almost every star, while the largest nebulae are about 200 light-years across and contain about 100,000 solar masses of ionized gas. A typical diffuse nebula is about 30 light-years across and has a density of roughly ten atoms per cubic centimetre (160 atoms per cubic inch), which represents an ounce of matter for every 10^9 cubic miles. In spite of this low density, the total mass of gas in such a nebula would be several times that of the Sun. The density distribution is chaotic and irregular, and the obscuration caused by the dust within and in front of the nebula makes it difficult to get an accurate idea of the gas distribution by optical means. Radio waves penetrate the dust unhindered, and radio observations reveal that on a large scale diffuse nebulae are more regular than their photographs would suggest. For instance, the Trapezium, a group of bright stars exciting the Orion Nebula, appears in visible light near the edge of the nebula (see Plate I). The radio contours reveal that the nebula is really rather symmetrical about the Trapezium; there is a dark cloud that acts as a curtain in front of about half of it. One problem with the radio observations is that the resolving power (ability to distinguish detail) of a single radio telescope is much poorer than that of an optical telescope: details covering about 10′ of arc rather than 1″ can be separated, so that the rather smooth radio contours may to some extent be illusory. It is possible but expensive in observing time to overcome this limitation by an interference technique called aperture synthesis, in which the outputs of two widely separated radio telescopes are used. The resolving power increases as the distance between the radio receivers is increased, and the radio-resolving power achieved in this way is comparable to or better than optical resolution.

Luminous filaments

High-resolution studies of diffuse nebulae reveal one of the surprises that make the study of astrophysics delightful. Instead of the smooth structure that might be expected of a gas, a delicate tracery of luminous filaments can be detected, down to the smallest scale that can be resolved. In the Orion Nebula this is about 400 astronomical units, or ten times the diameter of the orbit of Pluto, the outermost known planet, about the Sun. Even finer detail may exist, and there is evidence from spectra that much of the matter may be gathered into dense condensations, or knots, the rest of the space being comparatively empty. The time it would take unrestrained gas to fill a vacuum between the visible filaments is only about 200 years, much less time than the age of the nebula. How the gas is held together in the small knots is one of the intriguing questions still to be answered.

The largest diffuse nebulae in the Galaxy are revealed only by the radio telescope—they are so heavily obscured by dust that they are inconspicuous optically. These giant nebulae are hundreds of times as bright at radio wavelengths as the Orion Nebula. Similar objects, about 200 light-years in diameter, can be detected in other galaxies; if they were located at the Orion Nebula, they would cover the entire constellation of Orion with brightly glowing gas. The sizes of these giants seem to be about the same in all of the several nearby galaxies in

which they are observed and whose distances are known. If it is assumed that such nebulae at greater distances are also of about the same dimensions, a very useful yardstick becomes available for distance measurement. The angular size of such nebulae in other galaxies can be determined at great distances—up to 30,000,000 light-years. If both the true and angular size are known, the distance of the nebula and thus of the galaxy can be determined.

Forms taken by dark nebulae. Intimately associated with the diffuse nebulae are cold, neutral clouds of material that show up as dark nebulae because of the obscuring dust they contain. (The word cloud is used for convenience; actually there is probably some material everywhere, and cloud is used to mean a region with a local increase in density.) Sometimes the dark nebulae project as long, serpentine shapes seen against a background of stars beyond them; more typically, their borders are not sharp or obvious. A few dark nebulae are very opaque, allowing only $\frac{1}{1,000}$ of the incident light or less to pass through them, but more usually they absorb half of it or less. Most (perhaps 90 percent) of the mass of gas in the Galaxy is in this neutral material and can be seen as dark nebulae only when concentrated in dense clouds. On the average, the galactic material absorbs half of the light in a distance of about 3,000 light-years.

The dark nebulae are neutral and cold because not enough ultraviolet radiation is available in the regions they occupy to cause any appreciable ionization of hydrogen or helium (although the X-rays and cosmic rays in the Galaxy will always provide at least some ionization, about 1 percent or less). Such a lack of radiation energy is very important to the physical state of the gas; theoretical studies suggest that at typical pressures the gas separates into two different phases according to its state of ionization, one at about 50° K and the other nearer 1,000° K but with the same pressure (see also INTERSTELLAR MEDIUM).

Absorption is a major difficulty for optical-observational astronomers studying stars, since it dims the starlight being studied, but it does provide the means for studying the nature of the tiny particles that cause the absorption. Their exact nature is still not known, but it appears that they are in general a few millionths of an inch in size, spaced about 100 yards apart. An ounce would be distributed over a volume larger than the Earth. It is very likely that a significant fraction of these particles are graphite, similar to soot; there may well be particles of rocklike materials (silicates), silicon carbide (Carborundum), or other materials as well. An individual particle may possibly consist of a matrix of several different constituents. It has recently been shown that ordinary water ice does not exist in large quantities on the particles.

The origin of the particles is an interesting theoretical problem. While the grains are probably composed of common substances, solid grains cannot form from a completely gaseous, dilute medium, such as interstellar clouds. Nuclei of a few hundred atoms must somehow first stick together into a tiny "seed" particle; until the grain is large enough to radiate away the energy gained by a collision with an approaching atom, the atom will simply rebound instead of sticking. Such grains may form in the outer atmospheres of cool supergiant stars where the gas density is comparatively high (perhaps 10^9 times what it is in the nebula). If so, the gas in the star can carry away the heat liberated by the grains in capturing atoms. The grains are then blown out of the atmosphere into space by the radiation pressure (the mechanical force of the light they absorb and scatter). Calculations indicate that refracting materials, such as the constituents of the grains proposed above, should condense in this way.

The formation of stars

In the inner regions of dark nebulae important events take place: the formation of stars. The discussion that follows will be slightly oversimplified by the assumption that the nebula has very little net rotation. Since the density in a star is immensely greater than that in a nebula, star formation must occur through condensation. The force of gravity in the nebula constantly pulls it together, working against the disruptive collisions of any one nebula or cloud with its neighbours. But, even if gravity can hold the cloud together against other forces, the cloud cannot collapse unless it can cool; gravitational energy is released in a contraction and would heat the gas cloud, thus increasing the outward pressure and preventing further collapse. The dust grains are efficient emitters of infrared radiation, which escapes, removing the energy from the nebula. A cloud, therefore, contracts under its own gravity, radiating away half of the gravitational energy of the contraction, the other half going into heating the gas. As it contracts, the density and gravitational binding increase, and finally the cloud's gravity so dominates the internal pressure that the material rushes inward in almost a free fall.

While the entire cloud has been collapsing, it does not have a smooth density distribution but rather consists of a chaotic jumble of smaller clouds. These cloudlets pull themselves together by their own gravity into "protostars," each of which is destined to be an individual stellar system. Most of these protostars are smaller than one solar mass, but a very few may be several (up to about 100) times as massive as the Sun. These few massive stars have a profound influence in the evolution of the nebula.

Each protostar collapses very quickly; its gas falls inward in free fall. A protostar can collapse from a size equal to the outer diameter of the solar system to about 30 times the Sun's size (the size of Mercury's orbit) in about six months. After it reaches that size, the material becomes so hot it has become a star; that is, an opaque body radiating energy only from its surface. Around it is a whirling ring of cold, dusty material, which will also break up into even smaller fragments, the "protoplanets." It is now believed that planetary systems are made out of this dusty chaos.

Many such newly formed stars continue to contract, becoming hotter but less luminous in the process, until they start to produce heat by converting hydrogen into helium rather than by contracting. Such a star would then pour forth this energy from thermonuclear reactions into the dark clouds around it and begin to ionize the nebular material, producing a bright nebula around it.

These ideas are given encouraging confirmation by observations of dark nebulae in very long wavelength infrared radiation. Some of the brightest infrared sources are associated with such dark dust clouds; a good example is the class of T Tauri variables, named for their prototype star in the constellation Taurus. The T Tauri stars are known for a variety of reasons to be extremely young. The variables are always found in or near dark nebulae; they are often also powerful sources of infrared radiation, corresponding to warm clouds of dust heated by the T Tauri star to a few hundred degrees K. There are some strong infrared sources (especially in the constellation of Orion) that have no visible stars with them; these are presumably "cocoon stars" completely hidden by their veils of dust.

Formation of reflection nebulae. In 1912 it was discovered that the bright nebula in the Pleiades star cluster has a spectrum like that of the nearby stars rather than a typical nebular spectrum of bright emission lines. The interpretation is clear; the nebula is shining by starlight reflected from the dust particles rather than by any emission of the gas itself. The star illuminating such a reflection nebula is almost always a bright star of type B2 (temperature about 20,000° K) or cooler; hotter stars would ionize the gas and cause nebular emission lines. Studies of the angular sizes of the reflection nebulae led to the conclusion that about 60 percent of the light incident upon the dust is reflected (scattered) rather than absorbed, showing that the material is almost as reflective as snow. Calculations show that even graphite, which, of course, is black in bulk, can produce this effect when dispersed into tiny particles.

According to present theories the fate of the gas remaining in a cloud after star formation depends critically on the kind of star formed; most stars can be created in the interior of the cloud without much influence on the gas (other than to use it up in their formation). The situation

changes abruptly when one or more of the rare, massive hot stars, called O stars, form. These stars produce ultraviolet radiation that ionizes the gas, heating it from perhaps 50° to 8,000° K in the process. Each atom is separated into two parts, the nucleus and the electron, by the ionization, and the pressure in the gas is consequently increased by about 300 times. The ionized gas expands into the surrounding neutral, cold gas like a giant piston, compressing it tremendously. Just as in an explosion, a shock front is formed between the compressed gas and the quiet, undisturbed material ahead of it. Rapid star formation may occur in the compressed region, producing an expanding group of young stars. Such expanding groups, the so-called O Associations (with O stars) or T Associations (with T Tauri stars), have been observed, but it is not known from observations that these associations are formed in the manner just described. The great concentration of gas very close to the Trapezium stars ionizing the Orion Nebula shows clearly that the stars are very young indeed; otherwise, the gas would have expanded and smoothed itself out. It has been suggested that their age is only 20,000 or 30,000 years—hardly more than four-millionths of the age of the Sun, or comparable to two hours in the average life expectancy of man. Early men would not then have been able to see the fascinating "star" in the middle of Orion's sword.

Stellar associations

The gas expansion ends when the ionized gas has reached pressure equilibrium with the neutral material around it; it can hardly act as a piston after this but more resembles a hot bubble. In the time available the gas may not reach this state, which needs perhaps 10,000,000 years to achieve, a time that is comparable to the lifetime of the star that provides the outgoing energy. When the star has "died" after consumption of its nuclear fuel, the diffuse nebula is without its hot central star, and its material presumably recombines—unless the star's life ends with a spectacular supernova explosion, described below.

This picture of the evolution of diffuse and dark nebulae is one of constant turmoil, a few transient O stars serving to keep the material stirred, in constant motion, constantly producing new stars and churning clouds of gas and dust. In this way the thermonuclear energy of stars is converted into kinetic energy of interstellar gas.

Location of diffuse and dark nebulae in the Galaxy. One of the remarkable features of the diffuse and dark nebulae is their extreme concentration in the spiral arms in the plane of the Galaxy. While there is no rigid boundary to the arms, which have irregularities and bifurcations (or "spurs"), the nebulae in other spiral galaxies are strung out along these narrow lanes and form a beautifully symmetric system when viewed from another galaxy. The nebulae are remarkably close to the galactic plane; most are within 300 light-years, only 1 percent of the distance of the Sun from the centre.

The details of the explanation of why the gas is largely confined to the spiral arms is beyond the scope of this article (see GALAXY, THE). Briefly, the higher density of the stars in the arms produces sufficient gravity to hold the gas to them.

The question may be asked: why doesn't the gas simply condense into stars and disappear? The present rate of star formation is about one solar mass per year in the entire Galaxy, which contains something like two times 10^9 solar masses of gas. Clearly, if the gas received no return of material from stars, it would be depleted in roughly two times 10^9 years, about one-fifth the present age of the Galaxy. Several processes by which gas is returned to the interstellar medium from stars have been observed. Possibly the most important is the ejection of planetary nebula shells, discussed below; other processes are ejection of material from massive O- and B-type normal stars or from cool M giants and supergiants. This rate of gas ejection is roughly equal to the rate of star formation, so that the mass of free gas is declining very slowly.

This cycling of the gas through stars presumably has had one major effect: the chemical composition of the gas has been changed by the nuclear reactions inside the stars (see CHEMICAL ELEMENTS, ORIGIN OF). There is excellent evidence that the Galaxy originally consisted of almost pure hydrogen with (probably) some helium and that all of the heavier elements have been produced inside stars by being subjected to the exceedingly high temperatures in the central regions. Thus, most of the atoms and molecules on the Earth, as well as in human bodies, owe their very existence to processes that take place within stars.

Original composition of the Galaxy

Chemical composition of diffuse nebulae. The composition of the diffuse nebulae can be estimated by relating the strengths of the emission lines found in their spectra to the numbers of atoms producing them, and great strides have now been made in calculating the necessary atomic properties. The principal difficulties in gauging the accuracy of the determination are, however: (1) the use of average temperatures and densities when the material is actually clumped into regions with very different physical conditions and (2) making the right adjustment to the line strengths to allow for the absorption by the dust both within the nebula and between the nebula and the observer on the Earth. Nevertheless, nebulae appear to be remarkably consistent in their composition. In particular, the ratio of numbers of helium atoms to hydrogen atoms has been shown to be about 1 to 10 in many different objects, observational and theoretical uncertainties being about 15 percent; even nebulae in several other galaxies show the same value. Oxygen is the most abundant heavy element (the number of oxygen atoms is one-thousandth of the number of hydrogen atoms) and nitrogen is about one-sixth as abundant as oxygen. A few apparent exceptions occur in nebulae in which the helium abundance is measured by using radio (rather than the optical) recombination lines, always in the sense that the helium abundance seems low. One possible explanation is that hydrogen is ionized in a larger region than helium, which would make the helium lines appear weak or absent in the measurements.

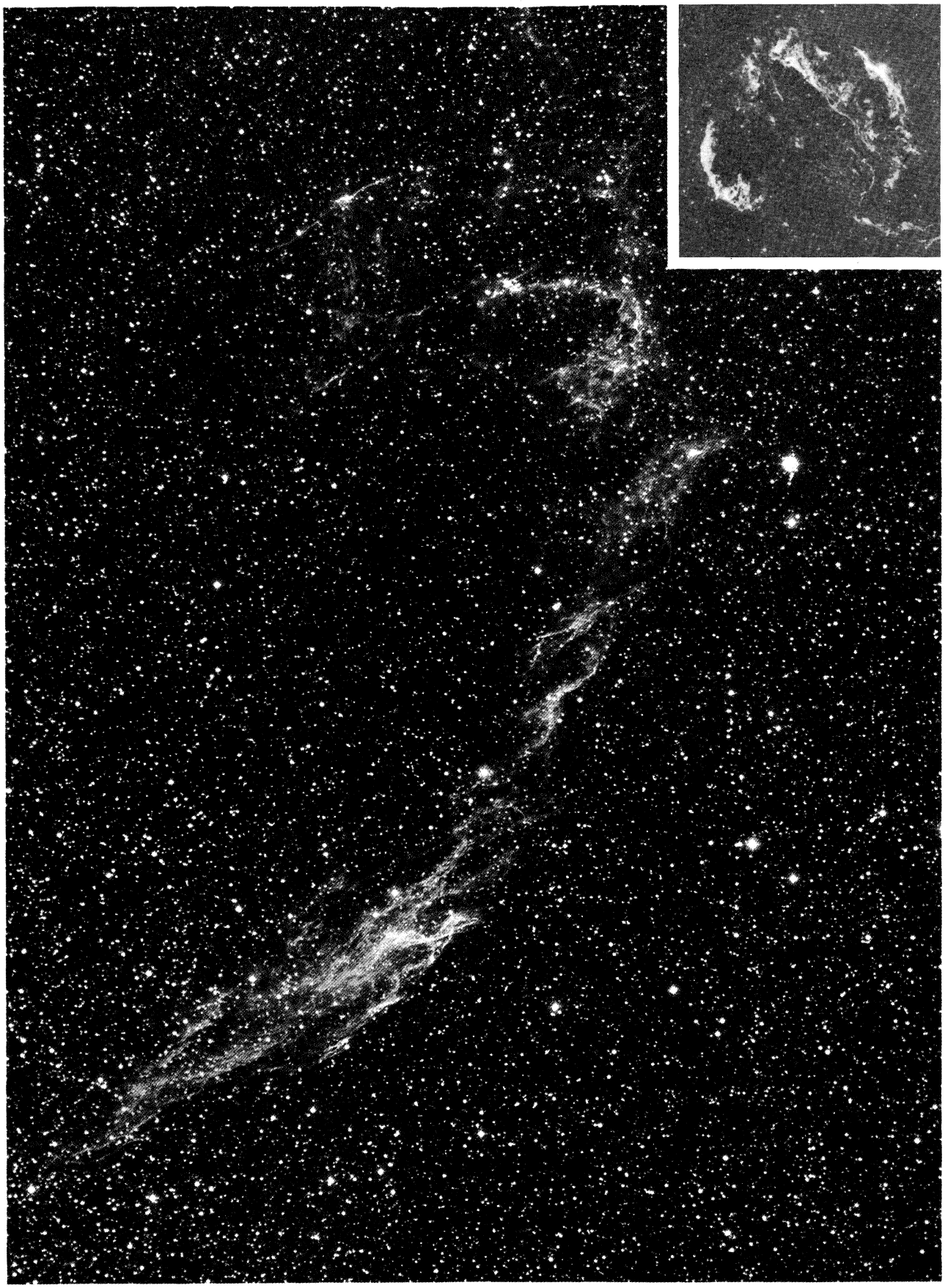
THE PLANETARY NEBULAE

Forms and structure. Compared to the diffuse nebulae, planetary nebulae are small objects, having a radius, typically, of one light-year and containing a mass of gas equivalent to about 0.3 solar mass. They are considerably denser than most diffuse nebulae, typically containing 1,000–10,000 atoms per cubic centimetre (16,000–160,000 atoms per cubic inch) within their dense regions. Many are so far away that they appear stellar when photographed directly, but the conspicuous examples measure up to 20′ of arc across, 10″–30″ being usual. Those that show a bright disk have much more regular forms than the chaotic diffuse nebulae, but there are still usually some brightness fluctuations over the disk. The planetaries generally have regular, sharp outer boundaries; often they have a relatively regular inner boundary as well, giving them an appearance like a ring. Many have two lobes of bright material, resembling arcs of a circle, connected by a bridge—somewhat resembling the letter Z.

Most planetaries show a central star, called the nucleus, which provides the ultraviolet radiation required for ionizing the gas in the ring or shell surrounding it. Those stars are among the hottest known and are in a state of comparatively rapid evolution.

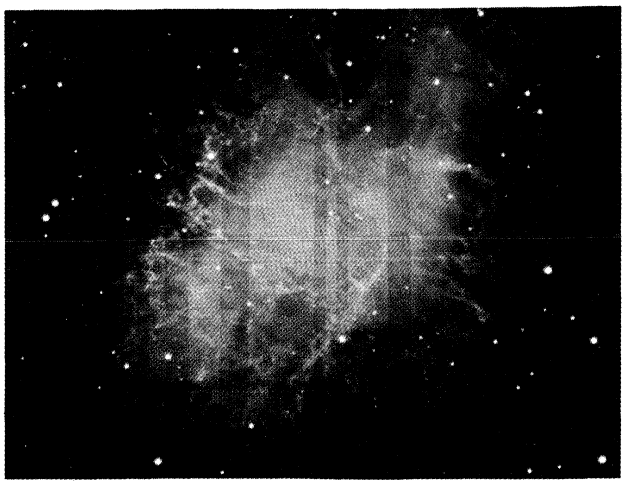
The central star

As with diffuse nebulae, the overall structural regularity conceals large-scale fluctuations in density. High-resolution photographs of a planetary nebula usually reveal tiny knots and filaments down to the resolution limit of the photograph. The spectrum of the planetary nebula is basically the same as that of the diffuse nebula; it contains bright lines from hydrogen and helium recombinations and the forbidden lines (defined above) of other ions. In general, the spectra of planetaries differ from those of diffuse nebulae in that they show much higher degrees of ionization. In some planetaries most of the helium is doubly ionized, and appreciable amounts of five-times-ionized oxygen and argon and four-times-ionized neon exist. In diffuse nebulae helium is mainly once ionized and neon and argon only once or twice. This difference in the states of the atoms results from the temperature of the planetary nucleus (up to about 150,000° K), which is

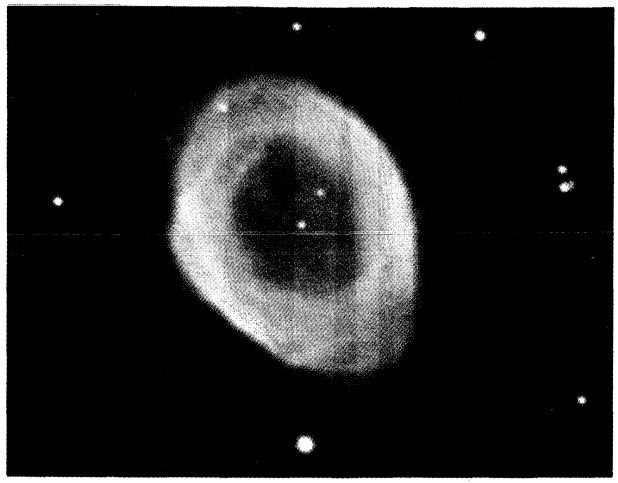


Veil Nebula (NGC 6992 and 6995) in the northern constellation Cygnus, which is part of a much larger spherical supernova remnant (inset). The nebula glows as it collides with dust and gas in interstellar space. Blue light is emitted from the hot leading edge of the nebula, where the most energetic collisions occur; the red glow is from hydrogen in the cooler gas.

Plate 1: By courtesy of Hale Observatories; photograph, © California Institute of Technology and Carnegie Institution of Washington, (inset) © National Geographic Society—Palomar Sky Survey

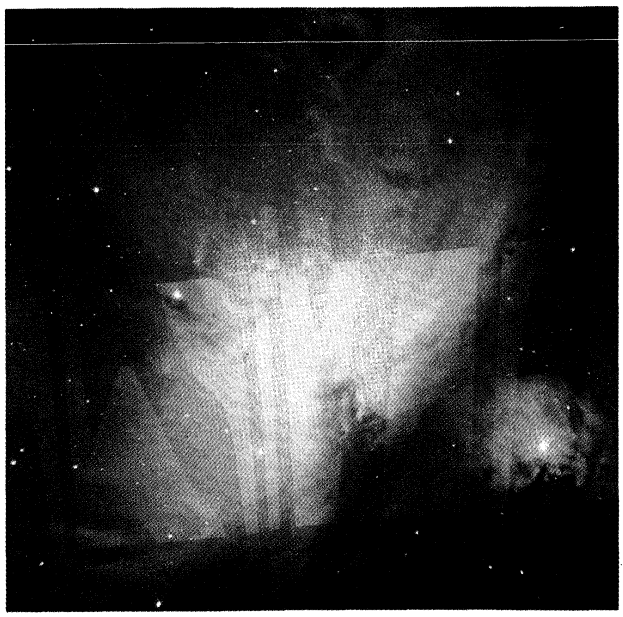


Crab Nebula (M1, NGC 1958) in the constellation Taurus, a gaseous remnant of the galactic supernova of AD 1054, 6,000 light-years away, expanding at 700 miles/second.

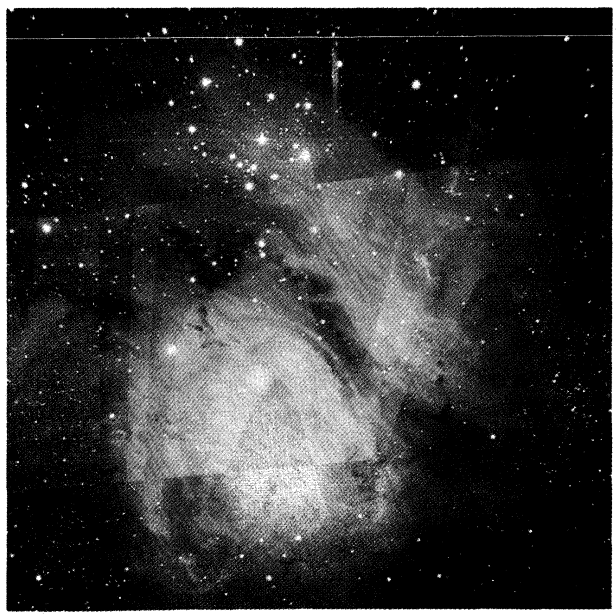


Ring Nebula (M57, NGC 6720) in the constellation Lyra, a planetary nebula consisting mainly of gases thrown off by the star in the centre.

(Above) Trifid Nebula (M20, NGC 6514) in the constellation Sagittarius, a bright diffuse nebula with lanes of dark material in front of it.
(Left) Bright nebulosity in the Pleiades (M45), distance 420 light-years. Cluster stars provide the light and surrounding clouds of dust reflect and scatter the rays from the stars.



Great Nebula in the constellation Orion (M42, NGC 1976), surrounded by glowing gas cloud, one of the largest gas-dust complexes in the Galaxy. Radiation from stars excites atoms of gas, causing them to emit the fluorescent light visible here.



Lagoon Nebula (M8, NGC 6526), in the constellation Sagittarius. This bright diffuse nebula is so large that light from the involved stars does not penetrate its boundaries and the bright nebula appears to be seen against a larger darker one.

Plate 2: By courtesy of Hale Observatories; photographs, © California Institute of Technology and Carnegie Institution of Washington

much higher than that of the exciting star of the diffuse nebula (less than 60,000° K for an O star, the hottest). One of the conspicuous features of planetaries is that the high stages of ionization are found close to the central star. The rare heavy ions, rather than hydrogen, absorb the photons of several hundred volts' energy. Beyond a certain distance from the central star, all the photons of energy sufficient to ionize a given species of ion have been absorbed, and that species therefore cannot exist farther out. Detailed theoretical calculations have rather successfully predicted the spectra of the best observed nebulae.

The planetary nebulae contain no optically detectable dust; their spectra do not contain an observable reflection component. There is, however, evidence gathered from the anomalously strong infrared emission from several well-observed planetaries that some dust exists—too little to detect optically but enough to absorb the strong ultraviolet emission line of hydrogen called Lyman alpha. Presumably, absorption of the Lyman-alpha radiation heats the dust to about 200° K, and the dust radiates in the infrared.

Expansion of planetaries

The spectra of the planetary nebulae reveals another interesting general fact: they are expanding from the central star at speeds of 15 to 35 miles per second. The gravitational pull of the star is quite small at the distance of the shell from the star, so the shell will continue its expansion until it finally merges with the interstellar gas around it. The expansion is proportional to the distance from the central star, consistent with the entire mass of gas having been ejected at one brief period from the star in some sort of instability.

The distances of planetary nebulae. Not a single planetary nebula is close enough to the Sun to allow a direct determination of its distance; the nearest one (NGC 7293) is about 300 light-years away. It is necessary, therefore, to use indirect methods to find the distances. One method is to estimate the true rate of emission of one of the hydrogen lines from the nebula (from the gas density and temperature) and compare this with the observed brightness. From the intrinsic and apparent brightness the distance can be found. One planetary, NGC 246, has a normal G star as a companion and its distance can be estimated if it is assumed that the companion is similar in intrinsic brightness to nearby G stars of similar type. Once the distance is known, other quantities (such as size and mass) can easily be found. Unfortunately, different methods of estimating distances give results differing by factors of more than two. A more reliable way of estimating distance is to consider the planetaries found in the Magellanic Clouds, the nearest external galaxies, whose distance (about 150,000 light-years) is known from reliable distance estimates of their other stars. At this distance, however, planetaries are very faint and correspondingly difficult to observe.

From the best available average distance determination, the true size of any nebula can be found, within the limits of error, from its angular size; typically, planetary nebulae are found to be a few tenths of a light-year in radius. If this distance is divided by the expansion speed, the age of the nebula since ejection is obtained, with values up to roughly 30,000 years.

Chemical composition. The chemical composition of planetaries can be found from their spectra in the way discussed above for diffuse nebulae. Almost all planetaries are found to have compositions remarkably similar to that of the diffuse nebulae. In particular, the helium abundance is about the same or only slightly higher, indicating that the ejection mechanism does not greatly disrupt the helium-rich interior of the parent star. The heavy-element content is similar, with two very interesting known exceptions: the planetary in the globular cluster M15 and a planetary in the "halo" (tenuous outer regions) of the Galaxy. Both have very low heavy-element content (down from normal by factors of about 50) but normal helium. Both objects are very old, and this result suggests that, while the primeval gas in the Galaxy had a low heavy-element content, it did have helium. The origin of helium in the Galaxy may very possibly have been the initial explosion of the universe itself.

Positions in the Galaxy. One of the best indicators of the average age of astronomical objects is their position and motion in the Galaxy. The youngest are in the spiral arms, near the gas out of which they have formed; the oldest are not concentrated to the plane of the Galaxy, nor are they found within the spiral arms. By these criteria, the planetaries reveal themselves to be rather middle-aged; they are moderately but not strongly concentrated to the plane; rather, they are concentrated toward the galactic centre, as the older objects are. Their motions in the Galaxy follow elliptical paths, whereas circular orbits are characteristic of younger stars. They belong to the type of distribution often called a "disk population," to distinguish it from Population II (very old) and Population I (young) objects of U.S. astronomer Walter Baade. It is possible that there is a wide variation in the ages of planetaries and that some are very young objects.

Stellar populations

Only about 1,000 planetaries are known, but their discovery is much hampered by obscuring dust. Their local (nearby) density suggests that there may be 50,000 in the Galaxy; if so, about 98 percent of them are hidden.

Evolution of planetary nebulae. A description of the evolution of a planetary nebula after it has been ejected from the central star begins with the cloud of gas perhaps a few light months from the star, at which distance gravity of the star is negligible. The expansion speed of the gas is typically 30 miles (50 kilometres) per second. Nebulae in this stage are bright but have starlike images as seen from the Earth, because they are too small to show a disk. The gas is at a comparatively high density—about 1,000,000 atoms per cubic centimetre—but becomes more dilute as the gas expands.

Now the chance that an ion will collide with an electron and recombine in a given length of time increases with the density. When the density is high, the matter nearest to the star can absorb the entire amount of ionizing radiation, and most of the gas is then in a neutral region outside the ionized (bright) part. That there is neutral hydrogen present can be deduced from the presence of forbidden lines (see above) of neutral oxygen in the spectrum; oxygen and hydrogen happen to have the same ionization potential, so, if neutral oxygen is present, neutral hydrogen must be also. During this stage the nebula appears to expand faster than the individual atoms of gas in it are moving; the ionized shell is "eating into" the neutral material as the density falls.

The middle stage of evolution occurs when the density has dropped to the point where the entire mass of gas is ionized. After this stage is reached, some of the ultraviolet radiation escapes into space, and the expansion of the nebula is caused entirely by the motion of the gas. Most planetaries are now in this middle stage. Finally the central star becomes less luminous and can no longer provide enough ultraviolet radiation to keep even the dilute nebula ionized. Once again the outer regions of the nebula become neutral and therefore invisible. Eventually the gas is mingled with the general interstellar gas.

A curious feature of several planetaries is that faint rings surrounding the bright, inner nebula can be observed. Possibly the outer rings are the remnants of a previous shell ejected earlier by the star. Alternately, they might be at the outer edge of the expanding neutral (and invisible) gas, which is ramming into quiescent interstellar gas originally present when the planetary ejection occurred. There would be a strong shock wave at this interface, and the shocked matter would be heated to several thousand degrees, causing the emission. The matter would quickly radiate its heat and become cold and invisible again. The key to the choice of those alternatives would be the spectrum of the faint ring, if it could be determined.

The central stars. The central stars are known from their spectra to be very hot. A common type of spectrum has very broad emission lines of carbon or nitrogen, as well as of ionized helium, superimposed upon a bluish continuum. These spectra are indistinguishable from those from the very bright rare stars known as Wolf-Rayet stars, but the planetary nuclei are about 100 times fainter than true Wolf-Rayet objects. The line widths are

interpreted as Doppler shifts caused by velocities in the line of sight of over 1,000 miles (1,600 kilometres) per second, enough for the mass to escape from the star. Thus, the stars appear to be losing some mass at the present time although evidently not enough to contribute appreciably to the shell.

The evolution of the central star. The presence of the nebula allows a comparatively precise determination of the evolution of the central star. The temperature of the star can be estimated from the nebula from the amounts of emission of ionized helium and hydrogen by a method devised by the Dutch astronomer H. Zanstra. The amount of ionized-helium radiation is determined by the number of photons of more than 54 volts' energy, while hydrogen is ionized by photons of more than 13.6 volts. The relative numbers of photons in the two groups depend strongly on the temperature, since the spectrum shifts dramatically to higher energies as the temperature of the star increases. Hence, the temperature can be found from the observed strengths of the hydrogen and helium lines. The rate of evolution of the stars can be found from the sizes of their nebulae, as the time since ejection of the shell is simply the radius of the nebula divided by the expansion rate. The energy output, or luminosity, of the central star can be estimated from the brightness of the nebula, because the nebula is converting the star's invisible ultraviolet radiation (which contains the greater part of the star's luminous energy) into visible radiation. The results pertain to only one nebula at one particular epoch, but consideration of the statistics of many nebulae provides a consistent picture of the evolution, just as one could learn about the human aging process by considering the vital statistics of an entire human population at any instant.

The resulting theoretical description of the star's evolution is quite interesting. While there seem to be real differences in stars at a given stage, the trends are quite clear. The central stars in the youngest planetary nebulae are about as hot as the massive O and B stars—35,000°–40,000° K—but roughly ten times fainter. They have half the diameter of the Sun but are 1,000 times as luminous. As the nebula expands, the star increases its brightness and temperature, but its radius decreases steadily. It reaches a maximum energy output, when it is roughly 10,000 times as luminous as the Sun, about 5,000 years after the initial expansion. This is an amazingly small fraction of the star's age of several times 10^9 years; it represents a period equivalent to about half an hour in a human life. From this point on the star becomes fainter, but for some time the temperature continues to increase while the shrinkage continues. At its hottest the star is perhaps 200,000° K, almost five times hotter than the hottest of most of the stars. It then cools and after about 10,000 years becomes a very dense white dwarf star, hardly larger than the Earth but with a density of several tons per cubic inch. From this point it cools very slowly, becoming redder and fainter indefinitely.

Final stages of central star's evolution

While there is not yet a very detailed theoretical picture of this contraction, a few results have emerged rather clearly: (1) white dwarf stars must obtain almost all of their energy from the contraction noted above, not from nuclear sources; therefore, (2) they must contain practically no hydrogen or helium, except perhaps in a very thin shell on their surfaces. These conditions would have to be met for the evolution to take place so quickly.

The absence of hydrogen in the star's interior is quite surprising; the planetary nebulae are all found to have a normal hydrogen abundance of about 1,000 times as many hydrogen atoms as heavy elements, such as oxygen. Thus, the mechanism of expulsion of the envelope must be very efficient at ejecting the hydrogen-rich outer layers of a star while leaving heavy-element-rich material behind.

The nature of the progenitor stars. A complete agreement on the kind of star that produced the nebula or the means by which it produced the nebula had not been reached in the early 1970s, but rather specific limits could be set on the probable nature of the progenitors.

First, the progenitor must have mass not much in excess of a solar mass because of the distribution of the planetaries in the Galaxy. Very massive stars are young and more closely confined than are nebulae to the galactic plane. Also, the mass of the nebula is roughly 0.3 solar mass, and the mass of a typical white dwarf (the final state of the central star) is roughly 0.7 solar mass. Next, the expansion velocity of the nebula is probably comparable with the velocity of escape from its progenitor, which implies that the progenitor was a red giant star, large and cool, completely unlike the small, hot, blue, nuclear star remaining after the ejection. Likely candidates are members of the class of Long-Period Variable stars, which have about the right size and mass and are known to be unstable. Symbiotic stars—*i.e.*, stars with characteristics of both cool giants and very hot stars—are also candidates. The novae (stars that brighten temporarily while ejecting a shell explosively) are definitely not candidates; the nova shell is expanding at hundreds of miles per second.

The cause of the ejection is not yet known; at least three different plausible explanations have been advanced. Perhaps the true situation is produced by a combination of several effects.

FILAMENTARY BRIGHT NEBULAE: SUPERNOVA REMNANTS

Some of the most fascinating objects in the sky are the remnants of supernovae. The supernova phenomenon is a spectacular explosion in which a star is almost completely disrupted and most of its mass ejected in a violently expanding cloud of debris that soon becomes an observable and remarkable nebula. At the brightest phase of the explosion, the expanding cloud radiates as much energy in a day as the Sun has done in the last 3,000,000 years. Such explosions are infrequent—one may occur roughly every 50 years within an entire galaxy. They have been observed less frequently in the Galaxy because most of them have been hidden by the obscuring clouds of dust. The known galactic supernovae were observed in 1006 in Lupus, in 1054 in Taurus, in 1572 in Cassiopeia (Tycho's nova, named after Tycho Brahe, its observer), and finally in 1604 in Serpens, called Kepler's nova. The stars became bright enough to be visible in the daytime.

Known galactic supernovae

In spite of the rarity of supernovae in the Galaxy, they have been found in large numbers in other galaxies. For a few days the brightness of the supernova is comparable to that of the entire galaxy, and it is easy to detect a supernova within a not-too-distant galaxy. Hundreds have been photographed and studied spectroscopically. At the site of Kepler's and Tycho's novae, there are heavy obscuring clouds, and the optical objects remaining are now inconspicuous knots of glowing gas. Near Tycho's nova, in Cassiopeia, there are similar optically insignificant wisps that appear to be remnants of yet another supernova explosion. To a radio telescope, however, the situation is spectacularly different: the Cassiopeia remnant is the strongest radio source in the entire sky (easily outshining the Sun, although the Sun is almost 10^9 times closer). Study of this remnant, called Cassiopeia A, reveals that a supernova explosion occurred there in approximately 1700, missed by observers because of the obscuring dust.

The Crab Nebula. At the site of the 1054 supernova is one of the most remarkable objects in the sky, the Crab Nebula. Photographed in colour, it is revealed as a beautiful red, lacy network of long and sinuous, glowing hydrogen filaments surrounding a bluish, structureless region whose light is strongly polarized. The filaments emit the spectrum characteristic of a diffuse nebula—*i.e.*, the emission lines of hydrogen, oxygen, and other elements. The gas is expanding at 700 miles (1,100 kilometres) per second—but still slower than the 4,000 miles (6,400 kilometres) per second in the shells of new supernovae in other galaxies.

The bluish, amorphous inner region of the Crab Nebula is even more remarkable. The glow is produced by the so-called synchrotron radiation—*i.e.*, from electrons spiralling about a magnetic field at almost the speed of light. This radiation is dramatically different from what it would be if it were emitted from electrons moving at low speeds: it becomes (1) strongly concentrated in the

Synchrotron radiation

forward direction, like the beam of a searchlight, (2) spread out over a broad range of frequencies but with the average frequency increasing with the electron's energy, and (3) highly polarized. Electrons of many different energies produce radiation in a large range of wavelengths: radio, infrared, optical, and ultraviolet. Even X-rays are emitted copiously by the Crab, which is the second-brightest X-ray source in the sky, indicating (for reasonable values of the magnetic field) that the electrons must have energies in the cosmic-ray range. After almost 1,000 years, the nebula is still losing 100,000 times as much energy per second as the Sun.

On the basis of this huge outpouring of energy, it is easy to calculate how long the nebula can shine without a new supply of energy. The electrons emitting the X-rays should decay, or drop to lower energies, in about 30 years —far less than the age of the nebula. The source of energy within the Crab Nebula supplying energy to the electrons that emit the X-rays was discovered in 1969 to be the pulsar NP 0532, the remains of the star that exploded to form the nebula (see PULSARS).

In the case of the Crab Nebula, the pulsar has been found to flash optically as well as at radio wavelengths, blinking on and off with the same period, believed to be a result of rotation: 0.033099324 second (on June 28, 1969). This period is slowly increasing (at the rate of 0.0012 second per century), which implies the pulsar is slowing down and thereby losing its energy to the nebula. The corresponding rate of energy loss is about equal to the nebula's rate of energy loss, convincing evidence that a tiny, extremely dense pulsar can supply the energy to the nebula, which is about four light-years across.

The Crab Nebula is unique in being a young supernova remnant and relatively free from obscuration, while Tycho's and Kepler's supernovae are conspicuous radio sources, apparently radiating by synchrotron emission; in neither case has a detectable pulsar been found. The Crab's pulsar, NP 0532, has the shortest known period, which suggests that it is the youngest known pulsar. The failure to detect pulsars at the sites of Kepler's and Tycho's novae, or Cassiopeia A, does not mean that they are not there; the radiation from the pulsar is probably strongly beamed, and the Earth may not be located in the direction to which the pulsars are sending their energy.

The Crab Nebula is still in a stage of violent expansion, and it is interesting to consider how it will appear a few thousand years hence. The gas will continue to expand and the pulsar lose energy until the nebula enters a second stage: one in which the pulsar can no longer put much energy into the nebula, and the nebula expands like a hot bubble of gas into the cold, surrounding interstellar gas. As the hot gas rams into the cold much faster than its speed of sound, a strong shock wave results. The cold gas is heated to about 100,000° K as it is hit by the shock wave and then cools by radiating forbidden and recombination lines. Almost all of the mass inside the shocked "bubble" will consist of this interstellar material swept up by the expanding shock wave.

The Cygnus Loop. As a general mechanism this picture seems well confirmed. The best observed old supernova remnant is the Cygnus Loop (or the Veil Nebula; see Plate II), a beautiful, filamentary object roughly in the form of a circular arc in Cygnus. Its patchiness is striking: the loop consists of a series of wisps rather than a continuous cloud of gas (see illustration). The most likely interpretation of this patchiness is that the interstellar medium into which the shock wave is propagating contains small clouds of denser material; many lines of reasoning from other evidence lead to the same result (see INTERSTELLAR MEDIUM). The present speed of the filaments is about 60 miles (100 kilometres) per second; the approximate age of the Cygnus Loop is about 50,000 years.

Other supernova remnants. Many other such old supernova remnants in the Galaxy are recognized, because they are also sources of radio emission from synchrotron radiation. Optically, in general they show a wispy appearance similar to that of the Cygnus Loop.

The eventual fate of these supernova remnants is to cool down, stop expanding, and finally merge with the interstellar medium. In the process they have modified the interstellar medium in two important ways: (1) They leave behind the elements that have been made in the presupernova star by a long process of heating and thermonuclear processing. Some very heavy elements are made in the very high temperatures right after the explosion. These elements are ejected into the supernova nebula and are finally merged with the interstellar medium. In this way the heavy-element content of the interstellar medium is increased. There is observational confirmation of this; both the Crab Nebula and the remnant of Tycho's supernova seem heavy-element rich. (2) They have provided kinetic energy to the surrounding gas by violent expansion, heating and accelerating the interstellar gas clouds. In addition, the original explosion emitted large quantities of very high energy particles that probably are the cosmic rays. These cosmic rays, as well as the X-rays emitted by the remnant, serve to heat the interstellar gas. It seems that supernovae play a very important role in the history of a galaxy, just as remarkable individuals shape the course of human history.

BIBLIOGRAPHY. B.M. MIDDLEHURST and L.H. ALLER (eds.), *Nebulae and Interstellar Matter* (1968), an excellent collection of chapters on various aspects of nebulae, each written by an expert in the field, especially good on the observational characteristics of nebulae; L.H. SPITZER, *Diffuse Matter in Space* (1968), a brief but excellent work on the physics of nebulae and interstellar matter; L.H. ALLER, *Gaseous Nebulae* (1956), a general monograph, slightly out of date, especially good on observational results; D.E. OSTERBROCK and C.R. O'DELL (eds.), *Planetary Nebulae* (1968), the proceedings of an International Astronomical Union conference on planetary nebulae in which the status of almost every aspect of this subject is summarized; H.J. HABING (ed.), *Interstellar Gas Dynamics* (1970), the proceedings of another IAU conference containing valuable discussions of supernova remnants and dynamics of diffuse and planetary nebulae; JEAN DUFAY, *Galactic Nebulae and Interstellar Matter* (1957), a somewhat more general summary than the above.

(J.S.M.)

Necker, Jacques

A Swiss banker and minister of finance under the French king Louis XVI, Jacques Necker was overpraised in his lifetime for his somewhat dubious skill with public finances and unduly depreciated by historians for his alleged vacillation and lack of statesmanship in the opening phases of the French Revolution.

Necker was born in Geneva on September 30, 1732, the younger son of Charles-Frédéric Necker, a lawyer from Küstrin in the Electorate of Brandenburg who had become a citizen of the Genevan republic in January 1726. At the age of 16, Jacques Necker became a clerk in the bank of a friend of his father; and in 1750 he was transferred to the bank's headquarters in Paris. In 1762 Necker was promoted to the position of junior partner. As a result of adroit speculation in the public funds and in the grain trade during the Seven Years' War, Necker became a prominent and wealthy banker. In 1764 he married Suzanne Curchod, the cultivated and talented daughter of a former Vaudois pastor (among her earlier suitors had been the historian Edward Gibbon). She encouraged him to embark on a public career. Geneva appointed him its resident minister in Paris in 1768; and he also became a director of the French East India Company. In 1772, at his wife's suggestion, Necker transferred his banking responsibilities to his brother Louis. He then acquired a reputation as a writer on financial topics with a eulogy written in memory of Jean-Baptiste Colbert, Louis XIV's great finance minister, and with an attack (1775) on the current free-trade policy in corn. Though he was a foreigner and a Protestant, Necker was placed in virtual control of French finances, as director of the royal treasury, on October 22, 1776, and by his appointment as director general of the finances on June 29, 1777.

First ministry

In his first ministry, Necker made several cautious experiments in social and administrative reform. He abolished mortmain (possession of lands by a corporation) on the royal domains in August 1779; reduced the numbers of the general tax farmers from 60 to 40; and

Necker, portrait by Augustin de Saint-Aubin (1736–1807), after a painting by Joseph-Sifford Duplessis.
H. Roger-Viollet

established "provincial assemblies" for Berry and for Haute-Guyenne with administrative powers in which the Third Estate (the commons) had as many representatives as the clergy and the nobility combined and voting was by head (1778–79). The first and the last of these experiments met with opposition from the privileged orders and were not extended, as had been hoped, to the country as a whole. Necker's principal mistake, however, was his misguided attempt to finance French participation in the American War of Independence without recourse to additional taxation. In trying to raise the necessary loans, Necker published in 1781 his celebrated *Compte rendu au Roi,* claiming a surplus of 10,000,000 livres in the hope of concealing an actual deficit of 46,000,000. The opposition of the leading minister, Jean-Frédéric Phelypeaux, comte de Maurepas, and the hostility of the queen, Marie-Antoinette, forced Necker to resign on May 19, 1781. He retired to Saint-Ouen, where he wrote his *De l'administration des finances de la France* to justify his policy (1784).

Outbreak of revolution

After both Charles-Alexandre de Calonne and Étienne-Charles de Loménie de Brienne had failed to solve the financial problems for which Necker was at least partially responsible, Necker himself was recalled as finance minister on August 26, 1788. France was now on the verge of bankruptcy, in spite of the aristocracy's agreement to surrender its immunity from taxation. In the face of the financial crisis it was decided to summon a meeting of the States General (the representatives of the clergy, nobility, and commons), which was to set in motion the French Revolution (see FRANCE, HISTORY OF). As the decision to summon the States General for 1789 had already been taken, Necker's main preoccupation lay in the field of politics rather than in finance, though he was too complacent in assuming that the surrender of the fiscal immunities of the nobility would remove his financial anxieties. In preparing for the meeting of the States General, Necker had to steer a difficult course between the claims of the Third Estate for double representation, which were conceded by the royal council on his recommendation in December 1788, and the insistence of the privileged classes on the traditional method of debate, order by order. He has often been blamed for not having clearly laid down the government's proposals for political as well as financial reform in his opening speech to the States General on May 5, 1789. He did, however, propose a reasoned program of social and constitutional reforms that was overlooked in the struggle over procedure. This struggle might have been avoided if Necker's suggestion for conciliation had been adopted. His program of liberal concessions to the National Assembly (formed between June 10 and June 17) was drastically modified by the court reactionaries just before the "royal session" of June 23, 1789. His objective was a limited constitutional monarchy with a bicameral legislature on the English model. His dismissal, on July 11, 1789, an overt sign of court reaction, did much to provoke the disturbances in Paris that culminated in the storming of the Bastille. Having retired to Switzerland, he received the summons to return to office on July 20, 1789.

In his third ministry Necker struggled ineffectively with the rapidly mounting deficit and was overshadowed by Talleyrand and the Comte de Mirabeau, who dictated revolutionary finance at that stage. Necker's chief weakness as a politician was his vanity and his anxiety to preserve his popularity at all costs. After his final resignation (September 18, 1790), he lived in retirement until his death at Coppet on April 9, 1804.

BIBLIOGRAPHY. No definitive biography in English has, as yet, been written. The most scholarly account (in French) of Necker and his Franco-Swiss background is EDOUARD CHAPUISAT, *Necker: 1732–1804* (1938). Other more popular biographies are EUGENE LAVAQUERY, *Necker, fourrier de la Révolution, 1732–1804* (1933); and PIERRE JOLLY, *Necker* (1947). For technical studies of Necker's financial policies, see MARCEL MARION, *Histoire financière de la France depuis 1715 . . .* , vol. 2 (1919, reprinted 1965); and two articles, the first by HENRI CARRE, "Le Premier Ministère de Necker, 1776–1781" in the review *La Révolution française*, 44:97–136 (1903); and the second by JULES FLAMMERMONT, "Le Second Ministère de Necker," *Revue historique*, 46:1–67 (1891).

(Al.G.)

Negro Churches (in the United States)

Negro churches are the churches of those Americans who have been known variously as Negroes, Africans, Afro-Americans, and blacks. These churches have also increasingly been termed black churches.

NATURE AND SIGNIFICANCE

Influence of the slave system

The historical and cultural importance of the black church in American Christianity is tied up with the *origin* of the black people in America. Unlike European immigrants, who went to America seeking religious and political freedom or better living conditions, Africans did not go by choice but as bondmen, chained on ships. The slave system shaped black life, especially its religious expressions. An investigation of the nature and significance of the black church in the United States, therefore, involves an assessment of what in pre-Civil War times was called the "peculiar institution" of slavery, and of its brutalizing effect on black life in America—its reduction of black people to a brute, subhuman existence.

The black churches as a distinguishable group during the 18th and 19th centuries represented the enslaved black people who were trying to come to terms with their existence as slaves in the acclaimed land of "freedom" and "equality." Black slavery in America meant not only unrewarded manual labour; it meant also a disruption of African family life and culture. Black people were forced to live in environmental conditions without accustomed cultural and religious values necessary for the regulation of their lives, and it was thus incumbent upon them to create order amid an imposed societal disorder. The creation of the black churches provided the much needed spiritual and social order that enabled oppressed black people to withstand the pain and suffering of slavery without losing their emotional and mental balance.

Though black people were enslaved physically and culturally, they usually refused to accept those white interpretations of Christianity that attempted to reconcile slavery with the gospel of Jesus Christ. The significance of this refusal is not found in formalized church creeds; the black churches did not differ significantly from their white counterparts. Almost without exception black people merely adopted the already existing creeds of the white Protestant churches. Their quarrel with white Christians originated not from abstract theological doctrines about God, Jesus Christ, or the Holy Spirit but from the overt and covert hypocrisy of the doctrinal formulations of white churches and their ethical ambiguity when applied to slavery and racial discrimination.

The liturgical style of the black churches has provided a religious framework for black affirmations and black negations in a society in which black people have lacked

social, political, and economic justice. The distinctive black "soul" is found in the spirit of black people who realize that their humanity does not depend upon the legal and moral codes of their oppressors. This spirit has shaped the distinctive liturgical style of the black churches in America.

Significance of black worship forms and styles

During the pre-Civil War period of enslavement "invisible" black churches (comprising blacks who worshipped with whites or held secret services) in the South expressed their reaction to involuntary servitude by projecting their view of life's meaning into an eschatological future (the consummation of history), a supernatural world in which God would make right what men had made wrong. In the "free" North, black people became involved in political, economic, and social affairs, attempting to make earthly justice a reality and not just a theoretical concept. Radical black preachers such as Henry Garnet (1815–82), Nathaniel Paul, and Daniel Payne (1811–93) were representatives of this tradition. Black spirituals (songs, or hymns) were sometimes the expressions of black people looking for justice in a heavenly world.

In worship, black people tend to transcend the formalized moral and legal codes of the white culture; their worship is related to their spiritual and physical survival. God is seen from a viewpoint arising out of the black condition, and worship is a spiritual response to the political and social limitations that were placed on black people. Thus, black worship emphasizes the *immediacy* of God's presence with black people, especially during adverse times. Black people tend to believe that "God will make a way out of no-way"—which means that they need not despair about seemingly impossible circumstances—and that as God delivered the Hebrews from Egyptian bondage, so also will he deliver black people from their oppressors. Black worship is an expression of this certainty, which explains the freedom of spirit in black worship that is characterized by spontaneous shouting, praying, singing, and preaching. Black people believe that God is at work in the world, and that despite the overwhelming power of various oppressors, his righteousness will become a reality in his ordained future.

Black theology

In the second half of the 20th century, some black American theologians have attempted to create a "black theology," based on the expression of certainty and peculiar style of black worship. These theologians believe that the essence of the Christian gospel is found in human liberation. God is the God of the oppressed, and his revelation is understood to be found in his historical identification with oppressed men, liberating them from human bondage. In America, this means that God is identified with the goal of strivings for liberation from white injustice.

The black theological emphasis on liberation and the freedom expressed in styles of worship are the chief contributions of the black churches to ecumenical discussions. Moreover, from the viewpoint of black theology, the very presence of the black church is ecumenically significant: it reminds Christians of European descent of past and present injustices committed against black people in the name of God and country and challenges them to recognize that the gospel of Jesus Christ cannot be reconciled with slavery and racism.

Academic study of black churches. Written sources about black churches are limited. Since black people were, by and large, excluded from the mainstream of American life, most scholars in religion simply ignored the black churches. Except for a few brief references, the black churches have seldom been mentioned in textbooks on American Christianity.

The black churches have been a subject of academic interest since the beginning of the 20th century, when black scholars began to reflect on the meaning of black life in America. Among the black scholars who have contributed studies of the black churches have been W.E.B. Du Bois (1868–1963), Charles Wesley (1891–), Carter G. Woodson (1875–1950), Benjamin Mays (1895–) and E. Franklin Frazier (1894–1962). These men, however, were not trained *church* historians, and a thorough study of the black churches from a historical-theological perspective still remains to be done.

HISTORY

The period of slavery (from the 17th century to the Civil War). In 1619, one year before the arrival of the ship "Mayflower" in Massachusetts, 20 Africans were sold as indentured servants at Jamestown, Virginia. At that time, no one could predict the profound implications of their presence, for not until 1661 did the colony of Virginia reveal the future significance of the presence of blacks in America, when it legally decreed that black people and their children would be slaves—chattels—for the duration of their earthly existence. Maryland legalized black slavery two years later. Shortly thereafter, other English colonies followed suit.

Early slave religion

For black people, slavery not only meant involuntary servitude; it also meant that slaveholders used various methods to destroy their African religious practices and beliefs. Initially, slaveholders not only forbade Africans to communicate in their native languages or to practice the worship of their gods but they also opposed teaching Christianity to slaves because of the unwritten law that a Christian could not be held as a slave. The established Anglican Church, however, assured slaveholders that the Christian gospel did not affect the civil status of men, and that the biblical emphasis on equality had nothing to do with the political equality of Africans with Englishmen. The Anglican Church contended that Christianity made Africans "better" slaves. With that assurance, slaveholders reluctantly permitted slaves to be baptized. But the Anglican Church had little success in gaining black converts. This is usually attributed to its alleged indifference to the civil freedom of black people, along with its rigid liturgical style and the apparent philosophical abstractness of Anglican Church doctrine.

Not until the 18th century, with the appearance of the Baptists and Methodists in the English colonies, did black people respond enthusiastically to Christianity. The appeal of the Baptists and Methodists may be found in their emphasis on civil freedom and equality, and in their simple styles of church worship. They provided an opportunity for black people to experience the emotional aspects of their religion and at the same time to affirm the right of civil freedom.

As a rule, slaveholders did not permit black slaves to organize independent churches because of the fear of insurrections and repressed early attempts to establish black churches. Slaves were at first required to worship with their masters. Most scholars believe that the first black Baptist church was founded sometime between 1773 and 1775 by a Mr. Palmer (about whom little is known) at Silver Bluff, South Carolina. This church was driven into exile when Savannah was evacuated by the Americans during the Revolution in 1778. Some of the members of the Silver Bluff church formed the nucleus for the first black Baptist church in Savannah, Georgia. George Liele (*c.* 1750–?) was a key figure in the organization of the church in Savannah; and before leaving for Jamaica, where he became an influential preacher among the blacks, he baptized Andrew Bryan (1737–1812), who later became an important leader in the Savannah church. Although Bryan did not speak out against the cruelty of the slave system, the slaveholders were fearful that his great influence among the slaves would result in insurrection. Many slaves were beaten and arrested for attending his church. Eventually Bryan and his brother Sampson were imprisoned and beaten severely.

Effects of the Nat Turner Rebellion

The Nat Turner Rebellion in 1831 confirmed the slaveholders' fear that black preachers were a serious threat to the stability of the slave system. Turner (1800–31) was a Baptist preacher who believed that God had ordained him to be the Moses of his people and led black people in armed revolt against their white masters. Fifty-five white people died as a result of his vision, and black preachers were consequently outlawed in most Southern states. The Mississippi law of 1831 is representative of such legislation: "It is unlawful for any slave, free Negro, or mulatto to preach the gospel under the pains of receiving thirty-nine lashes upon the naked back of the preacher."

Black churches in the North. Because slavery was not an important economic necessity in the North, most

Northern white churches did not seriously object to the organization of independent black churches, and indeed, they often encouraged it, particularly when the numbers of blacks increased, necessitating "Negro Pews" or galleries in white churches. The oldest independent black church was organized through the efforts of Richard Allen (1760–1831), who became the first bishop of the African Methodist Episcopal Church. In 1787 Allen and his black brothers were forced out of the church service during prayer at St. George Methodist Episcopal Church of Philadelphia. The blacks left and formed the Free African Society. The incident at St. George's Church convinced Allen that black people needed a church of their own where they would be free to worship God without having to encounter the racist attitudes and actions of white Christians. He was instrumental in organizing the Bethel African Methodist Church. In 1816 the Methodist Societies of Free Africans met at Philadelphia and founded the African Methodist Episcopal Church, of which Allen was elected the first bishop.

Events similar to the one that occurred at St. George's Church happened throughout the Northern states, resulting in the formation of the African Methodist Episcopal Zion Church in 1821, and of many independent African Baptist churches. White churches did not welcome black people, and it is noteworthy that segregation in public places had its origin in the churches. The post-Reconstruction South merely practiced in a wider sense what white churches had already initiated in the North.

The independent black churches and their preachers became involved in the movement to abolish slavery. Most white historians, however, have overlooked the significant role of black abolitionists, concentrating on white abolitionists, such as William Garrison (1805–79) and Arthur (1786–1865) and Lewis (1788–1873) Tappan. Mid-20th-century studies in black history by black scholars have changed this one-sided view. Many black Northern churches served as Underground Railroad "stations," and their ministers spoke out vehemently against slavery. Henry Garnett (1815–82), a Presbyterian preacher and an opponent of the Garrisonian abolitionists, even urged rebellion and cited the names of Nat Turner and Denmark Vesey (*c.* 1767–1822) as prominent examples of those who had led slave insurrections. Because of the involvement of the independent black churches in protest movements against slavery, slaveholders attempted to exclude the churches from operating in the slave states.

The period of entrenchment (from the Emancipation Proclamation of 1863 to World War I). During the post-Civil War period the involvement of the black churches in the struggle for the civil freedom of black people was discontinued. By and large, they replaced concern about earthly justice with concern for heavenly justice and became primarily concerned with the next life rather than with the problems of this life. The black churches withdrew from the politics of society at large and became a small isolated society, serving as a haven for oppressed black people. They became centres for the circumscribed political, economic, and social aspirations of black people who had been denied participation in the overall social and political affairs affecting their lives. Ministers of these churches, to a great extent, were more interested in who would be elected bishop or president of a convention than they were in the social and political liberation of the black people.

The retreat from civil concerns

In part, this change of emphasis is understandable as a response to new forms of oppression that emerged in post-Civil War America, "invisible" forms of slavery that were difficult and dangerous to cope with. Black ministers no doubt recognized this reality and attempted, consciously or not, to create a style of Christian life and worship that would not subject black people to mass murder. It could be argued that the black church (perhaps unknowingly) preserved the lives of black people for a later time when black unity against oppression could be more fully developed.

Not all black ministers, however, retreated from political life. During the post-Civil War Reconstruction period, black people entered politics, and a significant number of black politicians were preachers. Hiram Revels (1822–1901), an African Methodist minister from Mississippi, became the first black United States senator. Henry M. Turner (1834–1915) was a representative (1868) in the Georgia State Legislature until he and other blacks were expelled because of their race. He also served as postmaster in Georgia until he was forced to resign (1869) because of racial biases. Later, Turner was elected a bishop (1880) in the African Methodist Episcopal Church and became a strong advocate of the "Back to Africa" nationalist tradition.

Henry Turner, however, was an exception to the general fortunes of black people after the Civil War. Though the Thirteenth Amendment had ended legalized slavery, the masses of black people were subject to the vicissitudes of racial policies being developed by whites. The black church not only provided a structure for the regulation of the "morality" of black people; it also protected them from the fears of white people, assuring the societal leaders that black religion did not conflict with the values of American society.

Segregation in the churches

The "invisible" black churches in the South merged with existing independent black churches. Baptist and Methodist churches were the most successful. In most cases, however, white Southern churches excluded blacks. The Colored Methodist Episcopal Church (since 1954 called the Christian Methodist Episcopal Church) was founded in 1870. This church was an outgrowth of the Methodist Episcopal Church, South. Segregation thus became a way of life for all churches, North and South. Even churches that retained blacks on their membership lists (Methodists and Presbyterians among them) practiced segregation and refused to allow blacks to participate in policy-making decisions.

Nor was the Roman Catholic Church's relation to black people significantly different from that of the Protestant's. A vulnerable minority itself, it failed to extend itself on behalf of black freedom, and not until the Second Plenary Council of Baltimore, in 1866, did black people become a matter of interest in the Roman Catholic Church. At the Council the Roman Catholic Church, following the racial ethos of America, decided to organize separate schools and churches for black people. The first black Catholic church, however, had been founded in Baltimore in 1863. The most significant black priest of the period was Father James A. Healy (1830–1900), who was appointed a bishop of the diocese of Maine and later became an assistant to the papal throne.

By the beginning of the 20th century, the political disfranchisement of black people was virtually complete, and the black churches had withdrawn from the larger society. Most blacks apparently believed that earthly justice was hopeless, and they, therefore, placed their hopes in an eschatological future located in a transcendent, heavenly reality.

The period of urbanization (from World War I to World War II). During and after World War I (1914–18) large numbers of black people migrated from southern rural areas to northern industrial cities. Before this period, only one-tenth of the blacks in the United States lived in northern areas. World War I created an unprecedented demand for unskilled labour in northern industrial centres. Because the war cut off the immigration of European workers, there was a great demand for black labour. Other factors contributing to the influx of blacks to northern cities included the long-standing oppression of blacks in the South, the utopian image of the North held by blacks, and the effect of the boll weevil in the cotton-producing areas in creating a high percentage of unemployment among blacks throughout the South.

New styles of black churchmanship

In this period the black ghetto emerged in northern cities and with it new styles of black ministry and black churches. Unlike the rural South, where the church was the centre of the social and moral stability of the black community, the urban northern environment produced a breakdown of traditional black moral and religious values. Though many urban blacks continued to hold their traditional religious affiliations in existent Protestant churches, the religious behaviour of blacks was no longer

the same. This change has often been described as the "secularization" of the black churches—which means that the otherworldly outlook became this-worldly, focussing upon the existing social conditions of black people. Strict rules regarding card playing, dancing, and theatregoing (regulations derived from Baptist, Methodist, and other Protestant churches) were no longer enforced, and this may be cited as an example of the new secular emphasis. More important, however, was the black churches' growing interest in community affairs that involved relationships with the National Urban League (a voluntary interracial community service agency founded in 1910) and the National Association for the Advancement of Colored People (a voluntary interracial organization founded in 1909 to combat racial discrimination and segregation). Black ministers began to preach sermons related to the political affairs of black people. A most notable example of this practice was Adam Clayton Powell, Jr. (1908–72), pastor of Abyssinian Baptist Church in New York City, who led boycotts and marches against stores and other businesses that discriminated on the basis of colour. He was elected to the United States Congress (1945–70) and rose to the post of chairman of the prestigious House Committee on Labor and Education.

This was also the period of the rise of the "storefront" churches (congregations worshipping in empty stores) and black sects. The emergence of the storefront churches was primarily caused by the inability of traditional Protestant churches to satisfy the particular religious spirit of the southern blacks, who were accustomed to intimate and warm associations in the church. Most of the storefront churches were related to Baptist, Holiness, and some Methodist churches. The members of the storefront churches were, to a great extent, concerned about the freedom of worship; *i.e.*, the freedom to "shout" and praise the name of God without being sensitive about a lack of intellectual sophistication. They were concerned about the presence of the Holy Spirit and his effect on the lives of men. Thus they emphasized that people can know the *real* Christians by what they say and do. They believed that the sins of the flesh (drinking, smoking, adultery, etc.) are never found among Christians, and that God will never allow unclean persons to enter heaven.

Also present during this period were the black cults (religious groups), which should not be confused with the storefront churches. Both the cults and the churches often practiced in abandoned buildings and stores but differed markedly on religious beliefs. While the storefront churches tended to maintain traditional Protestant beliefs, the cults were characterized by new notions about God and, what is very important, about black people and being black. Some of the cults focussed on sanctification and holiness, attempting to restore the purest form of religion. Others focussed on black people as a race, attempting to relate religion to the economic, political, and social conditions of blacks. The former usually affirmed beliefs similar to those of traditional Protestant churches, and the latter tended to reject Christianity as a white man's religion, often affirming that Islām was the black people's religion.

The period of reformation (from the end of World War II to the present). Even before the civil rights movement in the 1950s and 1960s, Adam Clayton Powell, among others, had begun to relate the church to the political, social, and economic conditions of black people. He used his church as a power base for the black people's struggle against racial discrimination in New York City.

The Martin Luther King era

It was not until the appearance in 1955 of Martin Luther King, Jr. (1929–68), black clergyman and later Nobel Peace Prize recipient, that the black churches became focal points of protests against racial injustices and discrimination. Because of various movements and groups led by King—the Montgomery, Alabama, bus boycott (1955), initiated after Mrs. Rosa Parks refused to give up her seat to a white person, the organization of the Southern Christian Leadership Conference, a civil rights group, and the March on Washington (1963) by 200,000 blacks and white sympathizers calling for an end of segregation practices—it was no longer possible for the black churches to remain on the sidelines, without being condemned for their silence. The black churches and their leaders responded with enthusiasm to the call for civil justice, and the churches became meeting places for planning strategies and for rallies. Black ministers were often chosen as the leaders of civil rights marches and sit-ins (nonviolent noncooperation in public buildings).

During the period of Martin Luther King's leadership, the main emphasis was on the Indian leader Gandhi's (1869–1948) concept of nonviolence and on Christian love. Appealing to the biblical tradition of both the Old and New Testaments and relevant philosophical and theological ideas, King invited black people to follow him in nonviolently confronting the racism of white Americans with the aim of transforming it. Many blacks followed him, and the black churches once again became involved in the political life of the United States.

The movement led by King was neither regional nor racial; it involved the South and North, blacks and whites. White students and churchmen became involved in the civil rights movement and some died in the attempt to change the social, political, and economic conditions of blacks. The largest number of those who died as a result of their involvement in the civil rights movement, however, were members of the black community.

Before the assassination of King in 1968, a new emphasis in black concepts and practices emerged with profound implications for the civil rights movement and the black churches. In 1966 Stokely Carmichael, leader of the Student Nonviolent Coordinating Committee, verbalized the deep feelings and aspirations of many black Americans when he called for "Black Power." Advocates of Black Power held the view that blacks should not expect to achieve self-determination as a particular community as long as they depended on white people for aid and assistance; that as the oppressed, they should never expect their oppressors to relinquish power by merely appealing to Christian love but must be willing to fight for what is rightfully theirs—using violence if necessary.

This was the message of Black Power, and it was difficult for the black churches to decide how to respond to it. Christianity is generally known for its emphases on human equality and racial justice, as well as for its emphases on love and turning the other cheek. Thus, black churches found it difficult to respond to the new mood of many blacks who believed that the emphasis on nonviolence was politically irrelevant for the black people.

Faced with this dilemma, many black ministers and members followed the conservative tradition of emphasizing love and obedience to law as the only answer. This emphasis is found particularly in many black Baptist churches that are affiliated with the National Baptist Convention, USA, Inc., whose main spokesman is Joseph H. Jackson (1900–), president of the convention since 1953.

On the other hand, other black ministers and theologians argued that Black Power is a significant political reality not unrelated to the religious aspirations of black people. In 1966, the National Committee of Negro Churchmen (later, the National Committee of Black Churchmen) responded to criticisms of whites about Black Power by insisting that white racism was the basic reason for black unrest in America.

In 1968, black ministers and laymen in the United Methodist Church met in Cincinnati, Ohio, and proclaimed that their church was racist. They also endorsed Black Power as a relevant concept in the religious and political lives of black people. Soon after these two important events, caucuses of black ministers and laymen arose in various Protestant and Catholic churches.

The rise of Black Theology

Also, during this period, the term Black Theology emerged as a significant theological concept. The term was first used by an ad hoc group of black theologians within the National Committee of Black Churchmen, and it received academic respectability with the publication of *Black Theology and Black Power* and *A Black Theology of Liberation* by James H. Cone. In 1969, the National Committee of Black Churchmen issued a formalized statement on Black Theology, and in April 1970, the first

Consultation on Black Theology was held at Atlanta, Georgia.

The affirmation of a black theology is an attempt of the black churches to relate themselves to the self-determination of the black community as defined by Black Power. The advocates of Black Theology hold that the Christian gospel and black liberation are identical realities, and they refuse to accept a view of Christianity that denies the significance of blackness in the American society. They assert that oppressed people are God's chosen people, and that in America that means black people. They insist that the oppressed are chosen, not for suffering, but for liberation, and that they are free to achieve it by whatever means they regard as suitable.

Not all black theologians, however, have accepted the view as proposed by the adherents of Black Theology. The opponents contend that the emphasis on "black" is too exclusive and thus reductionist; they emphasize that the gospel is for all men—blacks and whites.

BLACK CHURCHES AND SECTS

Major Protestant churches. *Baptist churches.* The first black Baptist church, as noted earlier, was organized between 1773 and 1775 at Silver Bluff, South Carolina, across the river from Augusta, Georgia. During the slavery era black people were attracted to the Baptists because of their lack of rigid ritualistic requirements, and this denomination has continued to be the most influential church among the masses of black people. It was not until 1886 that the black National Baptist Convention was organized, and that was the result of an attempt by white Baptists to control the associations and conventions of black Baptists. When, in order to please their Southern white constituency, the American Baptist Publication Society (white) refused to accept the contributions of black ministers, Richard Henry Boyd (1843–1922) established the National Baptist Publishing House for black Baptist churches. Schisms among the black Baptists occurred in 1917 and 1961. The schism of 1917 resulted in the foundation of the National Baptist Convention, USA, Inc. (about 6,000,000 members in 1970) and the National Baptist Convention of America (nearly 3,000,000 members). The 1961 schism occurred in the larger group (of the USA, Inc.), producing The Progressive National Baptist Convention, Inc. (521,692 members).

Methodist churches. The oldest black Methodist denomination in America is the African Methodist Episcopal Church, organized in 1816 at Philadelphia, Pennsylvania. Richard Allen is usually referred to as the founder, and he was elected the first bishop. Like the doctrine and polity of all black Methodist churches, those of the African Methodist Episcopal Church do not differ significantly from those of white Methodist bodies, for its origin was due to racial discrimination rather than theological disagreement. This church is the largest of all among black Methodists, with over 1,000,000 members and over 6,000 churches. It also has churches in Africa and the West Indies.

The African Methodist Episcopal Zion Church, the second oldest and largest of the black Methodist Churches, held its first annual conference in 1821. James Varick (1750–1828) was elected the first bishop. Its doctrine and polity are similar to that of the African Methodist Episcopal Church, and it has nearly 1,000,000 members and over 4,000 churches in the United States, Africa, the West Indies, and South America.

The African Union First Colored Methodist Protestant Church, Inc., is the result of a union that occurred in 1866 between the African Union Church and the First Colored Methodist Protestant Church. Its doctrine is similar to other Methodist bodies but the polity is significantly different. This church has no bishops, and laymen have equal power with ministers in annual and general conferences. There are approximately 10,000 members.

Another Methodist denomination is the Christian Methodist Episcopal Church (originally called Colored Methodist Episcopal Church, its name changed in 1954). This church is an outgrowth of the Methodist Episcopal Church, South, in 1870. It has nearly 500,000 members.

Other churches. Another significant denomination is the African Orthodox Church, which is an outgrowth of the Protestant Episcopal Church. Like other black churches, it was founded because blacks believed that they should have a church of their own. The African Orthodox Church was founded by a Protestant Episcopal rector, George Alexander McGuire, in 1919, but it was not until 1921 that the first general synod was held and the present name adopted. McGuire was elected the first bishop. Of particular importance was McGuire's association with the black nationalist leader Marcus Garvey (1887–1940); he became a chaplain of Garvey's Universal Negro Improvement Association. This church strongly emphasizes apostolic succession, sacraments, and rituals. In 1957 there were 24 churches and 6,000 members in the United States.

Blacks in white Protestant churches

No exact figures are available for the number of blacks found in the major denominations with predominantly white memberships. The largest number of blacks own and operate their own churches, and only a small percentage is found in predominantly white denominations. The United Methodist Church has the largest number of black members (approximately 400,000) among such denominations. Blacks are found, however, in all major denominations, and recently almost all have organized black caucuses within these church bodies in order to confront the white majorities with their stated need for black self-determination as defined by various expressions of the Black Power movement.

Protestant sects. Most of the black religious sects owe their origins to Protestant Christianity and the urban migration of the second decade of the 20th century. Black sects often use biblical images and are usually found in large northern cities, such as New York City, Chicago, Detroit, and Philadelphia, and also, more recently, in large southern cities such as Atlanta. These sects seek to speak to particular needs of black people by attempting to offer "salvation" from political, social, and economic conditions. Release from humiliation and oppression stands at the centre of their messages. Often a state of "holiness," the freedom from sins of the flesh, is offered as the only cure to man's wretched condition. To be saved is to accept the Holy Spirit into one's heart, and this is believed to enable men to live the "good" and "holy" life and thus to prepare them to be received into heaven.

Of the many black sects, only a representative number that best illustrate the influence of the Protestant tradition can be mentioned here. See also NEGRO CULTS (IN THE UNITED STATES).

Pentecostal and Holiness sects

The Fire-Baptized Holiness Church separated from the white Fire-Baptized Holiness Association of America in 1908 and adopted its present name in 1922. Like most Pentecostal sects, its doctrines include repentance, regeneration, divine healing, premillennial (before the 1,000-year reign of Christ) Second Coming, justification, sanctification, pentecostal baptism, and speaking in tongues. It is ruled by one bishop, and it has over 1,000 members and 53 churches, with headquarters in Atlanta, Georgia.

The Church of Christ (Holiness) USA was founded in 1894 by C.P. Jones, a Baptist preacher, who "was dissatisfied with his religious experience and longed for a new faith which would make him one of wisdom's true sons and, like Abraham, 'a Friend of God.' " This church is similar to other holiness sects, emphasizing original sin, the gift of the Holy Spirit, and Christ's Second Coming. It is reported to have nearly 10,000 members and 146 churches, and it performs missionary work in Liberia and Nigeria.

The Mt. Sinai Holy Church of America, Inc., was founded in 1924 by Ida Robinson at Philadelphia. Bishop Robinson, as she is called, is the supreme head of the church because her authority is believed to come directly from God. The beliefs of this church include the gift of the Holy Spirit, divine healing, sanctification, and baptism.

The United House of Prayer for all People had its origin in the South. The founder was Bishop Charles Emmanuel Grace (1882–1960), who began preaching in 1925. "Daddy Grace," as he was often called, was the undisputed head of the church. Though the basic beliefs of this church

are similar to other holiness churches, "Daddy Grace," before his death, was reputed by his followers to have attained an authority equal with God. He often admonished his worshippers: "Never mind about God. Salvation is by Grace only. . . . Grace has given God a vacation, and since God is on vacation, don't worry him. . . ."

The Roman Catholic Church. The Roman Catholic Church's behaviour in regard to segregation and integration has not significantly differed from the Protestant churches'. Its schools and churches were segregated before the United States Supreme Court decision in 1954, and the progress of integration in them since that decision has been relatively slow. There is perhaps one significant difference, however, between the Roman Catholic Church and Protestant churches. The Roman Catholic Church has been able to enforce its decisions on integration in a more authoritative fashion, undoubtedly because of the locus of power in its hierarchy. Some black people joined the Roman Catholic Church during the 1950s and '60s precisely because of its power to enforce integration.

BLACK CHURCH PRACTICES

Worship. The distinguishing characteristic of black church worship is its freedom of expression. The desire for free expression probably accounts for the large number of blacks in Methodist, Baptist, Pentecostal, and Holiness churches. Some scholars have attributed this desire for an unstructured expression in worship to the black people's African heritage, and others have claimed that it is due to their peculiar economic, political, and social condition and to the unique ways in which black people have appropriated Christian beliefs and practices to meet their emotional and spiritual needs. Whatever the reason or reasons, black people have not responded enthusiastically to denominations with rigid ritual and liturgical requirement. Both their African heritage and their remaking of white Christianity under the most adverse conditions have probably shared in the shaping of black worship styles. A complete removal of African memories was not possible, and it may be assumed that such memories are reflected in black church worship. Nevertheless, the fact remains that black worship styles in the United States originated in the context of slavery and have been nurtured in oppression and racial injustice.

African and American influences on worship

Black worship styles were created and have endured because they are attempts by blacks to deal with their own peculiar position in America. Hence black ministers preached about an otherworldly reality, and black spirituals (religious songs) were poetic expressions of that reality. To preach and sing about heaven was to assert that, regardless of the cruelty inflicted on the blacks, the whites are not God and cannot have ultimate authority over the blacks. The freedom of expression in worship is regarded as an expression of this conviction.

Further, and more doctrinally, ritual may be viewed by most blacks as impeding the work of the Holy Spirit, the free-flowing power of God that can change the character of men, making them new persons. Spontaneous prayer, singing, preaching, and shouting are key ingredients in black worship.

Prayer is usually extemporaneous, and it is accepted as a sign of the work of the Spirit. Carefully prepared, written prayers are rare in black worship. Prayer is generally understood as man's communication with God through his Spirit; and this means that there is no reason to write it down as if one is addressing someone he does not know. In prayer, most blacks insist, one should come just as he is, not worrying about his speech or education, for Jesus is to be regarded as friend, and God his Father. The black man believes that God knows his condition, his sorrow, and wretched pain. He does not come to God to inform him of something that God does not know; he comes because he wants to renew his faith so that he can continue the troublesome journey. It is this emphasis that creates the freedom of expression in black worship.

Singing and shouting are also important aspects in black worship. Black spirituals are a tribute to the creativity of black people in poetry and song; they are a way of saying that, though enslaved, oppressed, and segregated, black people have refused to let their humanity be determined by white governmental officials and have affirmed their dignity through song and poetry. The spirituals were the creations of the "black soul" reaching for life and beauty in the midst of death and ugliness. Black people profess that they are a beautiful people because God has created them, and they lift their praise to him in song. There can be no authentic black worship without the singing of God's praises, for singing is viewed as a sign of the presence of the Holy Spirit; to sing is to express thanksgiving to God for the revelation of the certainty that oppression in this world is not the extent of life's existence. Shouting and the clapping of hands thus often accompany songs.

Preaching stands at the centre of black church worship. Those who are able to expound the Word of God are believed by black worshippers to have been "called" or "set aside" by the power of the Holy Spirit. Whatever may be said of the importance of prayer or song, they are merely prologues to the heart of black worship—the preaching of the Word. According to black worshippers, to be able to preach means that the individual has been grasped by the Holy Spirit, and it is not uncommon for the preacher to tell the hour and place that the Spirit first moved him. The content of the sermon is not confined to personal reflections on the preacher's conversion experience or call to the ministry; he also uses biblical images and stories as well as contemporary experiences of the community in order to demonstrate his understanding of the gospel message. The worship service usually ends after the sermon with a collection, song, and benediction.

Preaching: the centre of black worship

The role of the minister. The black minister is not only the chief spiritual adviser in the black community, he is often the leader of all its major community activities. During the 19th century, the minister was the major spokesman for the black community. Not until World War I and the black migration from the South to urban centres did black ministers begin to experience significant competition from other black professionals for leadership in the black community. Even today, however, despite the prestige of other professional groups, the black minister is still regarded as the most influential person in the black community.

He visits the sick, buries the dead, administers the sacraments, and plans the overall spiritual life of the community. But he also seeks and is often elected to political office. He is viewed as the key to the spirit and ethos of the black community. This is not too difficult to understand in view of the authority of the minister among blacks even in the more traditional Baptist and Methodist churches, and particularly in the Pentecostal churches. The churches are accepted by blacks as the most important institutions that provide structure to the life-style of the black communities, and the minister, the central figure in the church, occupies a significant role in that community. The minister is often accepted as the one who represents the goals and aspirations of the community as a whole, conveying to the larger society what the black community believes and feels.

Other practices. Since the black church has been the most important institution in the black community, it is the centre for many social, cultural, and political activities. Of particular importance are the revivals, the occasions when a special evangelist is invited to preach for a week or two weeks in order to add more members to the church. The entire community is always invited, and it is a time for renewal of faith and the offering of the opportunity to believe to those who do not believe.

The black churches' concern for its youth is reflected in the establishment of many educational institutions—high schools and two- and four-year colleges. For many black people, the black church schools offered the only possibility for higher education. Before blacks were more freely accepted in white colleges and universities, the church colleges laid the basic foundation for black college education in America.

CONCLUSION

Born in the context of slavery and oppression, the black churches have created a style of existence related to the

political, economic, and social condition of their members. Theologically, the churches may constitute a religious community of victims. Its activities naturally differ significantly from the churches of its asserted victimizers.

BIBLIOGRAPHY. C.G. WOODSON, *The History of the Negro Church*, 2nd ed. (1945), the best comprehensive history of the black church available; E. FRANKLIN FRAZIER, *The Negro Church in America* (1964), an excellent work on the black church by an expert sociologist; J.R. WASHINGTON, *Black Religion: The Negro and Christianity in the United States* (1964), a controversial work that describes black religion as folk religion and the white church as authentic Christianity—a good introduction to the black church despite its disputable interpretations; B.E. MAYS, *Negro's God* (1938, reprinted 1968), a good account of the black church's view of God as reflected in its literature; C.H. WESLEY, *Richard Allen: Apostle of Freedom* (1935), an excellent biography of the founder of the African Methodist Episcopal Church; H. THURMAN, *Negro Spiritual Speaks of Life and Death* (1947, reprinted 1969), and *Deep River* (1945, reprinted 1969), good theological treatises on the spirituals; A.H. FAUSET, *Black Gods of the Metropolis* (1944, reprinted 1970), the best book available on the black cults; J.H. CONE, *Black Theology and Black Power* (1969), the first book on Black Theology, seeks to reconcile Black Power and Christianity, and *Black Theology of Liberation* (1970), a systematic treatment of the important Christian doctrines in the light of Black Theology's emphasis on liberation.

(J.H.C.)

Negro Cults (in the United States)

The Negro cult—an extremely varied urban phenomenon—is a syncretistic religious group in revolt against the mainstream of American white and Negro religion and society, which it regards as unresponsive to its needs. The main roots of Negro cults—Judaism, Islām, and Christianity—each have passed through a cult phase (an early stage of a religious movement), as did countless other groups in Western society that have failed to grow into a sect, denomination, universal church, or religion. Negro cults are, with few important exceptions, short-lived, unstable, local, small in size, structureless, dominated by a charismatic leader, schism-prone, and mystically oriented.

"Race consciousness"—which may be described as the will to realize the natural dignity and integrity of all Negro people by reconstructing their moral, spiritual, economic, and political life—is instinctive with the Negro masses and is constantly renewed by the lives and teachings of the earliest esteemed "race men," or leaders. This earlier indigenous leadership was replaced by World War I with the "talented tenth," or black bourgeoisie. In the breach, the Negro masses—cut off from the regenerating source of a "race ethos" that was capable of inspiring cooperation for liberation and overmatched by objective limitations—were pressed by various forms of frustration, confusion, disillusionment, and apathy. They clutched their only possessions: a primordial race consciousness, religion, superstition, magic, rhythm, music, antiphonal singing, and improvisation. These floundering people, deprived both socially and economically, were forced to confront a leadership vacuum, which was filled by purveyors of voodoo (a religion of African origin), prophets, fortune-tellers, miracle workers, and other such personages who dealt with the aspirations of the Negro people. In this historical and cultural crucible the Negro cults evolved, originating among a plethora of refugees from a hundred different countries, replete with dozens of foreign dialects, and worshipping many strange gods. Before the advent of these later religious beliefs and languages, Islām and Judaism were unfamiliar to Afro-Americans.

Leadership of "race men"

HISTORY

The formative period. The ancestors of the Afro-American Negroes arrived in chains from a variety of complex African cultures, dominated by pervasive religions. In North America their religious needs were largely ignored from the beginning. Religion reached the slaves as tidal waves of enthusiasm, initially emanating from the revivalistic upheavals of the 18th-century Great Awakening and the later revivals of the 19th-century Second Awakening. These demonstrative ecstatic ventures of white American religion were primed by fervent preaching and pumped by wild frenzies, shouting, sobbing, weeping, fainting, swooning, and incoherent ejaculations. Isolated and segregated, Negroes first embraced and then re-created white religion in their own experience; they fashioned the "invisible institution," or black religion, a combination of other worldly aspirations and this-worldly protest, which developed in "field hollers" and "praise houses," shouts and plantation churches.

The expansionist period. While Southern Negro peasants were demoralized by lingering slave mores and "Jim Crow" (segregation) conditions after the Civil War, the new city-dwelling whites were unnerved by urban "worldliness." Their reaction to urban licentiousness created the "Holiness Movement," which built on the demand for freedom from sin, conversion, or justification (first blessing) that was generated in the revivals, and added the quest for perfect sanctification by grace or holiness (second blessing). This movement became national in scope and appeal around 1867 and encouraged the formation of the "Pentecostal Movement" that was distinguished by faith or spiritual healing and by baptism of the Holy Spirit (third blessing), which was manifested by "speaking in tongues." These movements welcomed the new urban Negroes who had moved to the cities and caught their imagination in storefront fellowships (churches occupying unused stores) where they were integrated into older Negro religious views and practices and restyled.

Types of cults that developed in urban centres. In the main, Negro cults follow two different, but not inherently conflicting, patterns. One, the Pentecostal or Holiness type, emphasizes separation from the world with separate religious and moral standards and has the tendency to substitute status gained through trances, visions, and gifts of tongues for social status. Strict spiritually oriented living, severe puritan morality, and concentration on personal problems serve as surrogates for the power to change conditions through unified Negro efforts. Explicitly, Pentecostal and Holiness cults equally abhor race consciousness and racism. Consequently, cults that are solely Pentecostal or Holiness oriented, without ties to the black protest heritage, do not ultimately satisfy and, indeed, increase the opportunity for more exotic cults to develop, along with some attendant charlatans who exploit the Negroes' urban miseries for personal gain and power, and offer simple solutions to complex problems of poverty, illness, despair, and discrimination.

Holiness, Pentecostal, black protest/black consciousness cults

Cults of the second pattern may have Pentecostal, Holiness elements, but they emphasize the connection with black religion in the form of black protest and black consciousness, or incipient black nationalism. In these cults, spiritual concerns serve various moral and social issues. Their major concern is the union of black people for self-direction and self-acceleration through a cherishing of the African heritage and Afro-American experience. These cults are militant and aggressive, and they exert an influence far beyond that of their numerical strength or that of their pessimistic, world-rejecting religious rivals.

SPIRITUALIST CULTS

Spiritualist cults, with their alleged beginnings in the voodoo- and Roman Catholic-influenced environs of New Orleans, combine typical Holiness and Pentecostal jubilant worship and spiritual healing, emotional Methodist and Baptist gospel hymns to the beat of secular rhythm and blues music, and the accoutrements from the Roman Catholic traditions such as altars, candles, and statues. The spiritualistic cultic leaders of the lower class segments that seek pleasure with religious blessings create a cult to fit the desires of Negroes for success in love, marriage, gambling, employment, and healing instead of remolding them to fit a religious ideal. Leaders of spiritualist cults often claim to be mediums in contact with the spirit of wisdom, rather than of the dead. Negroes who work hard but also yearn for "extras," or just sheer good times in social drinking, dancing, card playing, or gambling, find spiritual advisers and readers to be comforting

and inexpensive counsellors. Spiritualist advisers meet a need born of individual inclinations toward superstition, magic, religion, and spiritual power. They do not advertise themselves as advocates of the displacement of ability or initiative but simply as persons who can augment the natural through contact with the supernatural wisdom. The healing of illness and advice for any problem are available from spiritualist leaders who sell good-luck charms, holy flowers, sacred statues, or candles as means by which the objects of desires may be gained.

CULTS CENTRED AROUND INDIVIDUALS

Father Divine

Peace Mission movement. Father Divine (1877–1965) is regarded by his followers as God, Dean of the Universe, and Harnesser of Atomic Energy, according to the unique beliefs of the Peace Mission movement, which he founded. Worship of this five-foot-two rotund Negro by whites and blacks comprised its salient practice. According to some accounts, he was given the name George Baker at birth and reared in Savannah, Georgia, during the post-Reconstruction period when, out of unimaginable misery, black messiahs flourished, offering answers to the Negro's problems. He gained indelible impressions through immersion in primitive mysticism and the holiness and pentecostal movements. The transition from George Baker to Major J. Devine to Father Divine was completed when he set up his first heaven in Sayville, Long Island, in 1919. Legal entanglements forced him to relocate in Manhattan and Philadelphia and also fostered phenomenal growth through extensions in cities of the North and West. Heaven, according to Father Divine, is symbolized by separation of sexes and union of all races in a communion composed of a multicourse feast. He also taught a belief that there should be total racial integration, that the Bible is the *New Day* newspaper, that all things and persons are to be forsaken for the Father, and that heaven is on earth. He owned nothing, but a cult of 500,000 provided him free access to a $1,500,000 fortune. Father Divine supported blacks in businesses that became successful and fulfilled basic needs. Following the death of his wife, Penninah, called Mother Divine, he married in 1946, at the age of 70, a 21-year-old white follower, Edna Rose Richings. The new Mother Divine remained a virgin. The key to Father Divine's "success" was the devotion of competent disciples. At present, the movement lives off its past; it is contracting rather than expanding.

The United House of Prayer for All People. The United House of Prayer for All People differs from traditional Holiness cults in its belief that its founder, Bishop Charles Emmanuel Grace (1882–1960), is God; he is worshipped as "Daddy Grace." His followers genuflect and pray before his picture. A cook on a Southern railway, he was "called" to preach in 1926, assumed the name "Grace," and proclaimed himself "Bishop." The key to the success of this cult was that the many offerings went directly to "Daddy Grace" to advance the sale of his healing-power products: soap, stationery, tea, coffee, cookies, toothpaste, facial creams, talcum powder, hair dressing, and the *Grace Magazine*. Today this cult continues in skeleton form, a mere shadow of its earlier substance.

NATIONALISTIC CULTS

Marcus Garvey

Universal Negro Improvement Association. Marcus Garvey, born in 1887, established in his native Jamaica the Universal Negro Improvement Association (1914). It failed for lack of support. He re-established it in New York (1916) with extraordinarily successful results. In ten years he built the only Negro mass-based organization in American history. His effectiveness was due to putting into words and deeds what blacks were thinking, by means of great oratory and production of the *Negro World*, regarded by many as the best edited black weekly in New York. Thousands, and some reports indicate millions, followed him. He frequently spoke of the greatness of coloured civilizations and taught that God is black and Jesus was black, that subservience would end when blacks stopped worshipping a white God, and that black is beautiful. By his teachings he encouraged the Negro masses to organize, cooperate, and awaken to black consciousness. The principal plank of his movement was "back to Africa," to set up there an independent nation so that blacks everywhere could gain self-respect from pride, strength, and protection. The strength of this appeal, however, was his undoing. A master propagandist, his program was fiercely protestant, strongly emotional, and bitterly chauvinistic, thereby creating enemies who allegedly entangled him in a mail fraud charge resulting in his deportation from the United States (1927). Garvey remains important as the principal architect of black nationalism, which advocates possession of a distinct territory, a common heritage and culture, and self-help for self-determination.

Black Jewish cults. The Church of God (Black Jews of Philadelphia) is headed by Prophet F.S. Cherry. It is open to any black person. Members identify as black Jews those who believe that black people originated with Jacob and that white Jews should worship Jesus. The worship service begins and ends with the Prophet beating a drum. Castanets and tambourines, without piano, play a large part in the service. There is no Hebrew service or "Temple worship"; Baptism is observed, but Passover replaces the Lord's Supper.

The Black Jews of Philadelphia, Harlem, and Washington

The Commandment Keepers of the Living God (Black Jews of Harlem) is a tightly knit West Indian community of about 3,000 members, led by rabbinically trained (in Jewish traditions) Wentworth D. Matthew. The group holds Judaism to be their own and the one true religion, that Negroes are really Ethiopian Hebrews or Falashas, and that Jesus was black. These are the only black Jews with extensive contact among white Jews; they strictly observe the dietary laws and celebrate all Jewish holidays. Christianity is rejected by this group, and non-believing Negroes are held in low esteem.

The Church of God and Saints of Christ (Black Jews of Washington, D.C.) began under the leadership of Prophet William S. Crowdy. He passed his mantle of leadership to Bishop William Plummer who announced himself as "Grand Father Abraham." This group believes that all Jews were originally black and that Negroes are descendants of the "lost tribes of Israel." Their beliefs centre around the "Seven Keys," the "Stone of Truth," and the Ten Commandments. They cultivate 1,000 acres of land in Virginia, operate several industries, a school, and a home for orphans and the aged.

Moorish Science Temple of America. The Moorish Science Temple of America was founded by Timothy Drew in Newark, New Jersey, in 1913. He is believed to have been ordained Prophet Noble Drew Ali by Allāh. A Holy Koran, the sacred text, was created out of his study of Oriental philosophy. The prophet's central teaching is that salvation will come when Negroes discover and identify with their national origin as Asiatics, Moors, or Moorish Americans. Men are to wear fezzes at all times. The term Negro, or black, signifies death and the loss of power, authority, and their God. All laws are to be obeyed; prayers are offered three times daily; alcohol, sports, eggs, smoking, and meat are forbidden. This group was a forerunner of the Black Muslims (see also BLACK MUSLIMS). Drew died mysteriously during a period of internal strife, and the cult split and withered.

NATIVISTIC MOVEMENTS

Voodooism is the belief that the spirit-power resides in the python (snake) and is communicated through a priest. It entered the United States in 1809 by way of Ewe-speaking African slaves brought from Haiti by planters escaping the Cuban revolution at that time. It became virtually extinct as a cult in the United States with the death of its last queen, Marie Laveau (died 1881). Voodoo imposters, however, with Voodoo-like conjurations gain occasional hearings today among black Americans.

CONCLUSION

Negro cults do not differ in religious intention, intensity, form, or substance from other cults. Even the stated worship of a man (Father Divine) is not historically unique, as can be noted in the case of Jesus. Moreover, in such

Negro cults the spirit believed to be in the man, and not the man himself, is what is revered. Father Divine, for example, did not claim to be God. The significant distinction to be noted about Negro cults is their collective aspect as impacted depositories of residual Afro-American cultural elements during their phenomenal rise and subsequent decline into harbingers of change for Negroes in the United States.

BIBLIOGRAPHY. A.H. FAUSET, *Black Gods of the Metropolis: Negro Religious Cults of the Urban North* (1944); and R.J. JONES, *A Comparative Study of Religious Cult Behavior Among Negroes; with Special Reference to Emotional Group Conditioning Factors* (1940), are two classic cultural studies. E.T. CLARK, *The Small Sects in America*, rev. ed. (1949), sets Negro cults in the context of white Holiness and Pentecostal primitive religion. ST. CLAIR DRAKE and H.R. CAYTON, *Black Metropolis*, 2 vol., rev. ed. (1962), is most important as a source of religious creations in urban ghettoes. R.A. PARKER, *The Incredible Messiah* (1937), traces the influences bearing upon Father Divine. Two key critical essays are C.E. LINCOLN, *The Black Muslims in America* (1961); and E.U. ESSIEN-UDOM, *Black Nationalism* (1962). H. BROTZ, *The Black Jews of Harlem: Negro Nationalism and the Dilemmas of Negro Leadership* (1964), is an important but brief exploration of the phenomenon of Black Jews. E. FRANKLIN FRAZIER, *The Negro Church in America* (1964); and J.R. WASHINGTON, JR., *Black Religion: The Negro and Christianity in the United States* (1964), are more general works that supply a framework within which the cults' deviations find perspective.

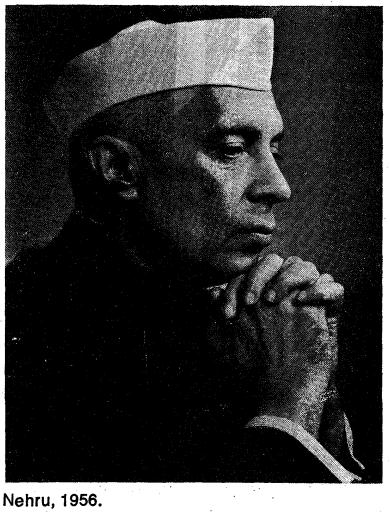
(J.R.W.)

Nehru, Jawaharlal

Jawaharlal Nehru, the first prime minister of independent India, was in many ways the antithesis of his leader, Mahatma Gandhi, the architect of India's freedom. Reared in a patrician household, educated at an exclusive English school and at Cambridge University, Nehru's outlook was more modern and forward-looking than that of the Mahatma. Nehru did not conceal his differences with Gandhi on many basic economic, social, and political issues. He did not share Gandhi's aversion to industrialization, and he saw to it that the early five-year plans after independence were geared to heavy industrialization. If Nehru accepted Gandhi's nonviolence, he did so not as a matter of principle but because he regarded nonviolence as a useful political weapon and the right policy for India under prevailing conditions. He deprecated the inordinate value the older man attached to austerity and asceticism and questioned his idealization of poverty. While Gandhi was more interested in individuals than in institutions and preferred to build upward from the people rather than to impose his ideas from above through a parliament, Nehru's economic thinking was basically ideological, tied to Socialist doctrine.

Early years. Nehru came of a family of Kashmiri Brahmins, noted for their administrative aptitude and scholarship, that had migrated to India early in the 18th century. He was born at Allāhābād in what was then known as the United Provinces, India, on November 14, 1889, the son of Motilal Nehru, a renowned lawyer and one of Gandhi's prominent lieutenants. Jawaharlal was the eldest of four children, two of whom were daughters. A sister, Mrs. Vijaya Lakshmi Pandit, was to earn celebrity in her own right as the first woman president of the U.N. General Assembly.

© Karsh



Nehru, 1956.

Education

Until the age of 16, Nehru was educated at home by a series of English governesses and tutors. Only one of these, a part-Irish, part-Belgian theosophist, Ferdinand Brooks, appears to have made any impression on him. Jawaharlal also had a venerable Indian tutor who taught him Hindi and Sanskrit. In 1905 he went to Harrow, a leading English school, where he stayed for two years. His housemaster describes him as "a very nice boy, quiet and very refined. He was not demonstrative but one felt there was great strength of character. I should doubt if he told many boys what his opinions were, or the masters with whom he had a good name, as he worked well and seldom (almost never) gave trouble."

Nehru's academic career at school and college was in no way outstanding. From Harrow he went to Trinity College, Cambridge, where he spent three years earning an honours degree in natural science. On leaving Cambridge he qualified as a barrister after two years at the Inner Temple, London, where in his own words he passed his examinations "with neither glory nor ignominy."

The seven years Nehru spent in England left him in a hazy half-world, at home neither in England nor in India. "I have become," he wrote some years later, "a queer mixture of the East and West, out of place everywhere, at home nowhere." He came back to India to discover India. The contending pulls and pressures that his experience in England was to exert on his personality were never completely resolved.

Four years after his return to India, in March 1916, Nehru married Kamala Kaul, who came from a Kashmiri family settled in Delhi. Their only child, Indira Priyadarshini, now Mrs. Indira Gandhi, prime minister of India, was born in 1917. Kamala Nehru, who suffered from poor health for the greater part of her brief life, died in 1936.

Political apprenticeship. On his return to India in 1912, Nehru at first tried to settle down as a lawyer. But, unlike his father, he had only a desultory interest in his profession and did not relish either the practice of law or the company of lawyers. At this time he might be described, like many of his generation, as an instinctive nationalist who yearned for his country's freedom, but, like most of his contemporaries, he had not formulated any precise ideas on how it could be achieved.

Nehru's autobiography discloses his lively interest in Indian politics not only as an undergraduate at Cambridge but as a boy at Harrow. His letters to his father over the same period reveal their common interest in India's freedom. But not until father and son met Gandhi and were persuaded to follow in his political footsteps did either of them develop any definite ideas on how freedom was to be attained. The quality in Gandhi that impressed the two Nehrus was his insistence on action. A wrong, the Mahatma argued, should not only be condemned, it should be resisted. Earlier, Nehru and his father had been contemptuous of the run of contemporary Indian politicians, whose nationalism, with a few notable exceptions, consisted of interminable speeches and long-winded resolutions. Jawaharlal was also attracted by Gandhi's insistence on fighting Britain without fear or hate.

First meeting with Gandhi

Nehru met Gandhi for the first time in 1916 at the annual meeting of the Indian National Congress Party in Lucknow. Gandhi was 20 years his senior. Neither seems to have made any initial impact on the other. Gandhi's autobiography, *The Story of My Experiments with Truth*, which he dictated while in prison between 1922 and 1924, contains not a single reference to Jawaharlal. The omission is understandable, since Nehru's role in Indian politics was secondary until his election as Congress president in 1929, when he presided over the historic Lahore session that proclaimed complete independence

as India's political goal. Until then the objective had been dominion status.

Nehru's close association with the Congress Party dates from 1919 in the immediate aftermath of World War I. This period saw a wave of nationalist activity and governmental repression culminating in the massacre of Amritsar in April 1919; 379 persons were killed and at least 1,200 wounded when the local British military commander ordered his troops to fire on a crowd of unarmed Indians assembled for a meeting.

When, late in 1921, the prominent leaders and workers of the Congress Party were outlawed in some provinces, Nehru went to prison for the first time. Over the next 24 years he was to serve another eight periods of detention, the last and longest ending in June 1945, after an imprisonment of almost three years. In all, Nehru spent over nine years in jail. Characteristically, he described his terms of incarceration as normal interludes in a life of abnormal political activity.

General secretary of Congress Party

His political apprenticeship with the Congress lasted from 1919 to 1929. In 1923 he became general secretary of the party for two years and again, in 1927, for another two years. His interests and duties took him on journeys over wide areas of India, particularly in his native United Provinces, where his first exposure to the overwhelming poverty and degradation of the peasantry (which he describes so movingly in his autobiography) had a profound influence on his basic ideas for solving these vital problems.

Though vaguely inclined toward Socialism and interested in the Soviet experiment, until then Nehru's radicalism had set in no definite mold. The watershed in his political and economic thinking was his tour of Europe and Soviet Russia during most of 1926–27. Nehru's real interest in Marxism and his Socialist pattern of thought stem from that tour, even though it did not appreciably increase his knowledge of Communist theory and practice. His subsequent sojourns in prison enabled him to study Marxism in more depth. Interested in its ideas, but repelled by some of its methods, such as the regimentation and heresy hunts of the Communists, he could never bring himself to accept Karl Marx's writings as revealed scripture. Yet from then on, the yardstick of his economic thinking remained Marxist, adjusted, where necessary, to Indian conditions.

Struggle for Indian independence. After the Lahore session of 1929, Nehru emerged as the leader of the country's intellectuals and youth. Hoping that Nehru would draw India's youth, at that time gravitating toward extreme leftist causes, into the mainstream of the Congress movement, Gandhi had shrewdly elevated him to the presidency of the Congress Party over the heads of some of his seniors. The Mahatma also correctly calculated that, with added responsibility, Nehru himself would be inclined to keep to the middle way.

After his father's death in 1931, Jawaharlal moved into the inner councils of the Congress Party and became closer to the Mahatma. Although Gandhi did not officially designate him his political heir until 1942, the country as early as the mid-1930s saw in Nehru the natural successor to Gandhi. The Gandhi–Irwin pact of March 1931, signed between the Mahatma and the British viceroy, Lord Irwin (later Lord Halifax), signalized a truce between the two principal protagonists in India. It climaxed one of Gandhi's more effective civil-disobedience movements, launched the year before, in the course of which Nehru had been arrested.

Hopes that the Gandhi–Irwin pact would be the prelude to a more relaxed period of Indo-British relations were not borne out; Lord Willingdon (who replaced Irwin as viceroy in 1931), jailed Gandhi in January 1932, shortly after the Mahatma's return from the Second Round Table Conference in London. He was charged with attempting to mount another civil-disobedience movement; Nehru was also arrested and sentenced to two years' imprisonment.

The Round Table Conferences in London, held to advance India's progress to self-government, eventually resulted in the Government of India Act of 1935, giving the Indian provinces a system of popular autonomous government. Ultimately, it provided for a federal system comprising the autonomous provinces and princely states. Although for various reasons federation never came into being, provincial autonomy was implemented.

During the middle '30s Nehru was much concerned with developments in Europe, which seemed to be drifting toward another world war. He was in Europe early in 1936, visiting his ailing wife, shortly before she died in a sanitarium in Switzerland. Even at this time he emphasized that in the event of war India's place was alongside the democracies, though he insisted that India could only fight in support of Britain and France as a free country.

When the elections following the introduction of provincial autonomy brought the Congress Party to power in a majority of the provinces, Nehru was faced with a dilemma. The Muslim League under Mohammed Ali Jinnah (who was to become the creator of Pakistan) had fared badly at the polls. Congress, therefore, unwisely rejected Jinnah's plea for the formation of coalition Congress–Muslim League governments in some of the provinces, a decision on which Nehru had not a little influence. The subsequent clash between the Congress and the Muslim League hardened into a conflict between Hindus and Muslims that was ultimately to lead to the partition of India and the creation of Pakistan.

Imprisonment during World War II. When, at the outbreak of World War II in September 1939, the viceroy, Lord Linlithgow, committed India to war without consulting the autonomous provincial ministries, the Congress Party's high command withdrew its provincial ministries as a protest against this action. Congress's action left the political field virtually open to Jinnah and the Muslim League.

Nehru's views on India's war role

Nehru's views on the war differed from those of Gandhi. Initially, the Mahatma believed that whatever support was given to the British should be given unconditionally and that it should be of a nonviolent character. Nehru, on the other hand, held that nonviolence had no place in defense against armed aggression and that India should support Britain in a war against Nazism, but only as a free nation. If it could not help, it should not hinder. He held broadly to this view, regardless of Gandhi's shifting postures.

In October 1940, Gandhi, abandoning his original stand, decided to launch a limited civil-disobedience campaign in which leading advocates of Indian independence were selected to participate one by one. Nehru, the second of these leaders, was promptly arrested and sentenced to four years' imprisonment—his eighth term of detention. After spending a little over a year in jail, he was released along with other Congress prisoners, three days before Pearl Harbor.

When, after Pearl Harbor, the Japanese carried their attack through Burma to the borders of India in the spring of 1942, the British government, faced by this new military threat, decided to make some overtures to India. Prime Minister Winston Churchill dispatched Sir Stafford Cripps, a member of the war Cabinet who was politically close to Nehru and also knew Jinnah, with proposals for a settlement of the constitutional problem. Cripps's mission failed, however, for Gandhi would accept nothing less than independence, even though Nehru would have accepted an arrangement enabling India to participate honorably in the war effort.

The initiative in the Congress Party now passed to Gandhi, who called on the British to leave India; Nehru, though reluctant to embarrass the war effort, had no alternative but to join Gandhi. Following the Quit India resolution passed by the Congress Party in Bombay on August 8, 1942, the entire Congress working committee, including Gandhi and Nehru, were arrested and imprisoned. Nehru emerged from this—his last detention—only on June 15, 1945.

Within two years India was to be partitioned and free. A final attempt by the viceroy, Lord Wavell, to bring the Congress Party and the Muslim League together failed. The Labour government that had meanwhile displaced Churchill's wartime administration dispatched, as one of

its first acts, a Cabinet mission to India and later also replaced Lord Wavell with Lord Mountbatten. The question was no longer whether India was to be independent but whether it was to consist of one or more independent states. Hindu–Muslim antagonism, culminating late in 1946 in clashes costing 7,000 lives, made the partition of India inevitable. While Gandhi refused to accept it, Nehru reluctantly but realistically acquiesced. On August 15, 1947, India and Pakistan emerged as two separate independent nations. Jawaharlal Nehru became independent India's first prime minister.

Achievements as prime minister. In the 35 years from 1929, when Gandhi chose Nehru over his seniors as president of the Congress session at Lahore, until his death as prime minister in 1964, Nehru remained—despite the Chinese debacle of 1962—the idol of his people. Superficially, his secular approach to politics contrasted with Gandhi's religious and traditionalist attitude, which during the Mahatma's lifetime had given Indian politics a religious cast—misleadingly so, for, although Gandhi might have appeared to be a religious conservative, he was actually a social nonconformist trying to secularize Hinduism.

Nehru as modernizer

The real difference between Nehru and Gandhi was not in their attitude to religion but in their attitude to modern civilization. While Nehru talked in an increasingly modern idiom, Gandhi was harking back to the glories of ancient Indian civilization. Yet within the Hindu framework, Gandhi was socially as much a protestant as Nehru outside it. Both men were, in the eyes of orthodox Hindus, secular in respecting their own religion while simultaneously respecting the religions of other Indians and other peoples.

The importance of Nehru in the perspective of Indian history is that he imported and imparted values and ways of thinking, which he adapted to Indian conditions. Apart from his stress on secularism and on the basic unity of India, despite its racial and religious diversities, Nehru was deeply concerned in carrying India forward into the modern age of scientific discovery and technological development. In addition, he aroused in his people an awareness of the necessity of social concern with the poor and the outcast and of respect for democratic values. One of the achievements of which he was particularly proud was the reform of the ancient Hindu civil code that finally enabled Hindu widows to enjoy equality with men in matters of inheritance and property.

Internationally, Nehru's star was in the ascendant until October 1956, when India's attitude on the Hungarian revolt against the Soviets brought his policy of nonalignment with either the Communist or non-Communist powers under sharp scrutiny by the non-Communist countries. In the United Nations, India was the only nonaligned country to vote with the Soviet Union on the invasion of Hungary, and thereafter it was difficult for Nehru to command credence in his calls for nonalignment. In the early years after independence, anticolonialism had been the cornerstone of his foreign policy, but after Chou En-lai, the Chinese prime minister, stole his thunder at the Bandung Conference of Afro-Asian states in 1955, Nehru's interest in this issue waned. By the time of the Belgrade conference of nonaligned countries in 1961, Nehru had substituted nonalignment for anticolonialism as the most pressing concern of the developing countries.

The Sino-Indian conflict of 1962 exposed the limitations of Nehru's wishful thinking on nonalignment. When the Chinese threatened to overrun the Brahmaputra Valley in November 1962, they exposed the hollowness of Nehru's euphoric commitment to the policy of "Hindu-Chini bhai bhai" ("Indians and Chinese are brothers"); his call for all-out Western aid made virtual nonsense of nonalignment.

The claims on Kashmir and Goa

Kashmir—claimed by both India and Pakistan—remained a perennial problem throughout Nehru's term as prime minister. His tentative efforts to settle the dispute by adjustments along the cease-fire lines having failed, Pakistan, in 1948, made an unsuccessful attempt to seize Kashmir by force. In solving the problem of the Portuguese colony of Goa—the last remaining colony in India—Nehru was more fortunate. Although its military occupation by Indian troops in December 1961 raised a furor in many Western countries, in the hindsight of history, Nehru's action is justifiable. With the withdrawal of the British and the French, the Portuguese colonial presence in India became an anachronism. Both the British and the French had withdrawn peacefully. If the Portuguese were not prepared to follow suit, Nehru had to find ways to dislodge them. After first trying persuasion, in August 1955 he had permitted a group of unarmed Indians to march into Portuguese territory in a nonviolent demonstration. Even though the Portuguese opened fire on the demonstrators, killing nearly 30, Nehru stayed his hand for six years, appealing meanwhile to Portugal's Western friends to persuade its government to cede the colony. When India finally struck, Nehru could claim that neither he nor the government of India had ever been committed to nonviolence as a policy.

Nehru's health showed signs of deteriorating not long after the clash with China. He had a slight stroke in 1963, followed by a more debilitating attack in January 1964. He died in New Delhi on May 27, 1964, from a third and fatal stroke.

Assessment. While consciously assertive in his Indianness, Nehru never exuded the Hindu aura and atmosphere clinging to Gandhi's personality. Because of his modern political and economic outlook, he was able to attract the younger intelligentsia of India to Gandhi's movement of nonviolent resistance against the British and later to rally them around him after independence had been gained. Nehru's Western upbringing and his visits to Europe before independence had acclimatized him to Western ways of thinking. Of all the leaders of the Congress Party, including Gandhi, he alone had given serious thought to India's place in the world community. This enabled him not only to educate his people on foreign affairs before independence but to project his own views on India's foreign policy when freedom came. If Gandhi made Indians aware of India, Nehru made them also aware of others. When India achieved independence the image it presented to the world was really Nehru's image: in the early years of India's independence, the world identified India with Nehru.

Throughout his 17 years in office, Nehru held up democratic Socialism as the guiding star for his country. Emphasizing the necessity of both democracy and Socialism for India, he set out to achieve Socialism through democracy. With the help of the overwhelming majority that the Congress Party maintained in Parliament during his term of office, he was able to advance toward that goal. The four pillars of his domestic policies were democracy, Socialism, unity, and secularism. He succeeded to a large extent in maintaining the edifice supported by these four pillars during his lifetime.

BIBLIOGRAPHY. For a biography of Nehru, see MICHAEL BRECHER, *Nehru* (1959); and *The New States of Asia: A Political Analysis* (1963), which contains, as an appendix, the author's talks with Nehru on many subjects. See also WALTER R. CROCKER, *Nehru: A Contemporary's Estimate* (1966); and GEOFFREY W. TYSON, *Nehru: The Years of Power* (1966). In A.B. SHAH (ed.), *Jawaharlal Nehru: A Critical Tribute* (1965), ARNOLD TOYNBEE's article published in *Encounter* is reprinted. BIMLA PRASAD, *The Origins of Indian Foreign Policy*, 2nd ed. (1962), contains a wealth of material on Nehru's views on foreign affairs before 1947. The appendixes include his letter to Roosevelt dated April 12, 1942, and his inaugural address to the Asian Relations Conference, March 23, 1947. TAYA ZINKIN, *Reporting India* (1962), contains material unpublished until then. It is excellent on the relationship between father and son, covering the years 1861 to 1931. SUDHIR GHOSH, *Gandhi's Emissary* (1967), discusses in part 2, "The Nehru Era," and throws much light on Nehru after the Chinese invasion and on his relations with Kennedy. Other recommended works include: V.B. KULKARNI, *The Indian Triumvirate; A Political Biography of Mahatma Gandhi, Sardar Patel, and Pandit Nehru* (1969); the memoirs of ABUL KALAM AZAD, *India Wins Freedom* (1959); and H.V. HODSON, *The Great Divide: Britain, India, Pakistan* (1969), on Nehru's relations with Mountbatten.

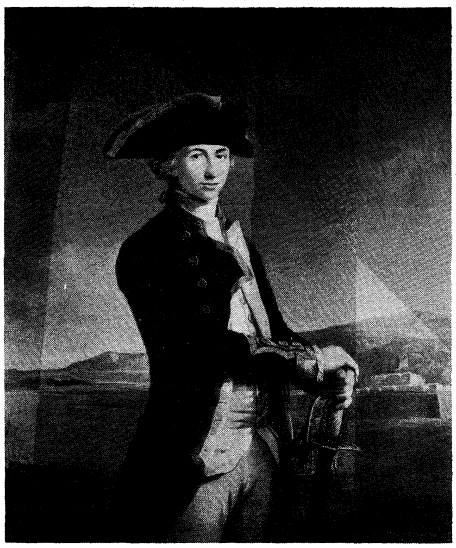
(F.R.M.)

Nelson, Lord

Horatio Nelson, the most successful and popular British naval commander in the wars with Revolutionary and Napoleonic France, set abiding standards of originality and initiative in strategical and tactical thinking and in leadership for every major navy and is still generally accepted as the most appealing of British national heroes.

Nelson's flair and forcefulness as a commander in battle were decisive factors in his two major victories—the battles of the Nile and Trafalgar. In the former, he destroyed the French fleet upon which Napoleon Bonaparte had based his hopes of Eastern conquest, and in the latter he destroyed the combined French and Spanish fleets, thus ensuring the safety of the British Isles from invasion and the supremacy of British sea power for more than a century. Spectacular success in battle, combined with humanity as a commander and a scandalous private life, raised Nelson to godlike status in his lifetime; and after his death at Trafalgar in 1805, he was enshrined in popular myth and iconography.

By courtesy of the National Maritime Museum, Greenwich, England



Nelson, oil painting by J.F. Rigaud (1742–1810). In the National Maritime Museum, Greenwich, England.

Horatio Nelson was born on September 29, 1758, in Burnham Thorpe, Norfolk, the sixth of 11 children of its rector, Edmund Nelson, and his wife Catherine. The Nelsons were genteel, scholarly, but poor. The family's most important connection from which Nelson could expect preferment was that with a distant relation, Lord Walpole, the descendant of Sir Robert Walpole, who had been prime minister earlier in the century.

Decisive for Nelson's life, however, was his mother's brother, Captain Maurice Suckling, who was to become comptroller of the British navy. When Horatio's mother died, Captain Suckling agreed to relieve the widower's burden by taking the boy to sea. So, although apparently a delicate, sensitive child, the nine-year-old Horatio, after attending local school, joined his uncle's ship, the "Raisonnable," at Chatham in 1771.

First years in the navy. Nelson's first years in the navy were a mixture of routine experience and high adventure. The former was gained particularly in the Thames estuary; the latter comprised a voyage to the West Indies in a merchant ship and a dangerous and unsuccessful scientific expedition to the Arctic in 1773.

Nelson had his first taste of action in the Indian Ocean. Soon after, struck down by fever—probably malaria—he was invalided home, and while recovering from the consequent depression, Nelson experienced a dramatic surge of optimism. A vision of "a radiant orb" inspired

> a sudden glow of patriotism . . . and presented my king and country as my patron. . . . 'Well then,' I exclaimed, 'I will be a hero, and confiding in Providence, I will brave every danger.'

From that moment, Nelson's ambition, fired by patriotism tempered by the Christian compassion instilled by his father, urged him to prove himself at least the equal of his eminent kinsmen.

Service in the war against the American colonies

In 1777 Nelson passed the examination for lieutenant and sailed for the West Indies, the most active theatre in the war against the American colonies. Promoted captain in 1779, at the early age of 20, he was given command of a frigate and took part in operations against Spanish settlements in Nicaragua, which became targets once Spain joined France in alliance with the American Revolutionaries. The attack on San Juan, in which he played a leading part, was militarily successful but ultimately disastrous when the British force was almost wiped out by yellow fever; Nelson himself was lucky to survive.

He was invalided home once more, before being given command of another frigate destined for operations off North America. At New York City he met Prince William Henry, later to become duke of Clarence and King William IV, who was then a midshipman in Admiral Lord Hood's flagship. The Prince, struck by "the merest boy of a captain" found "something irresistibly pleasing in his address and conversation; and an enthusiasm, when speaking on professional subjects, that showed he was no common being." The ensuing friendship led to Nelson's reception by King George III in London.

In 1783, after the end of the American Revolution, Nelson returned to England by way of France. He longed for feminine company and fell in love with an English girl in Saint-Omer who did not reciprocate his affections. On his return to London he was cheered by the appointment, in 1784, to command a frigate bound for the West Indies.

But this was not to be a happy commission. By rigidly enforcing the Navigation Act against American ships, which were still trading with the British privileges they had officially lost, he made enemies not only among merchants and shipowners but among the resident British authorities who, in their own interest, had failed to enforce the law. Under the strain of his difficulties and of the loneliness of command, Nelson was at his most vulnerable when he visited the island of Nevis in March 1785. There he met Frances Nisbet, the widow of a doctor, and her five-year-old son Josiah. Pleasant-looking, intelligent, and poised, Fanny Nisbet, a few months older than Nelson, seemed the ideal wife for a dedicated naval officer. Nelson conducted his courtship with formality and charm, and in March 1787 the couple was married at Nevis.

Marriage

Returning with his bride to Burnham Thorpe, Nelson found himself without another appointment and on half pay. He remained unemployed for four years, observing the routines of family visits, market days, harvests, and shooting parties. He read a great deal and followed the newspapers, writing to friends and possible patrons at the Admiralty in the hope of further employment. Aware of "a prejudice at the Admiralty evidently against me, which I can neither guess at, nor in the least account for"—but which may well have been connected with his zealous enforcement of the Navigation Act—Nelson was delighted to receive a new appointment. Within a few days of the execution of King Louis XVI in January 1793, he was given command of the 64-gun "Agamemnon."

Service in the Mediterranean. From this moment, Nelson the enthusiastic professional was gradually replaced by Nelson the commander of genius and Nelson the celebrity. The coming months were probably his most tranquil emotionally. At home waited a loving wife, whose son he had taken to sea with him. His ship, fast and manoeuvrable, his crew superbly trained, pleased him. His task was to fight the Revolutionary French and support British allies in the Mediterranean. Assigned to the forlorn defense of the port of Toulon against the revolutionaries—among them a 24-year-old officer of artillery, Napoleon Bonaparte—Nelson was dispatched to Naples to collect reinforcements. King Ferdinand IV (later Ferdinand I, of the Two Sicilies), who ruled Naples and Sicily with his domineering Queen Maria Carolina, sister

of the French queen Marie-Antoinette, readily agreed to send 4,000 men; and Nelson gratefully recognized that he owed the success of his mission largely to the British minister—the adroit and scholarly Sir William Hamilton, who had lived at Naples for 30 years and whose vivacious young wife Emma was in the Queen's confidence. In Emma Hamilton he at once recognized a remarkable woman but probably not one who was to alter his life. As a young courtesan with well-connected lovers, she had acquired an attractive smattering of smart etiquette, and her beauty had been recorded by prominent artists such as George Romney and Thomas Lawrence. Her zest, generosity, and quick intelligence won response from the Neapolitans, and Nelson wrote to his wife that Lady Hamilton "does honour to the station to which she is raised." The seeds of their infatuation had been planted. Emma Hamilton's warmth was both sexual and maternal; and her sense of humour and Sir William's erudition together may have been a luxurious echo of Parson Nelson's whimsical and scholarly company in Norfolk. But the passion was to lie dormant for several years.

First meeting with Emma Hamilton

When Toulon fell, Lord Hood, Nelson's commander, moved his base to Corsica, which was in revolt against the French, whose garrisons there were under siege. Nelson and his ship's company went ashore to assist first in the capture of Bastia, then of Calvi, where a French shot flung debris into Nelson's face, injuring his right eye and leaving it almost sightless. At the end of 1794, Hood was replaced by the uninspiring Adm. William Hotham, who failed to decisively defeat French men-of-war when they attempted to recapture Corsica. In this action, Nelson himself fought and, after two hours, captured the huge "Ça Ira," which had 120 guns. Exultant, yet frustrated, he wrote, "Sure I am, had I commanded our Fleet . . . that either the whole French fleet would have graced my triumph, or I should have been in a confounded scrape."

Hotham was replaced by an officer more to Nelson's liking. At 60, Adm. Sir John Jervis was an immensely experienced seaman who quickly recognized Nelson's qualities and who regarded Nelson "more as an associate than a subordinate Officer." The arrival of Jervis coincided with an upsurge of French success by land so that the British were forced to abandon their Mediterranean bases and retreat upon Gibraltar and the Tagus.

Battle of Cape St. Vincent. Making for a rendezvous with Jervis in the Atlantic off Cape St. Vincent, Nelson found himself sailing in mist through a Spanish fleet of 27 ships on its way to join with other Spanish, French, and Dutch squadrons for an attempted invasion of the British Isles. The Spaniards were sailing in two divisions and Jervis planned to cut between the two and destroy one before the other could come to its assistance. But he had miscalculated and, having passed between the two enemy columns, it became clear that the British ships would not be able to turn quickly enough to get into action before the Spanish squadrons closed up.

Without orders from Jervis, and violating the rigid tactical doctrine of the time, Nelson hauled out of the line and threw himself at the head of the second Spanish division. While the rest of Jervis's fleet slowly turned and came up in support, Nelson held the two Spanish squadrons apart with his one ship, at one time fighting seven enemy ships. The efficiency of British gunnery was decisive; and he not only boarded and captured one enemy man-of-war but, from her deck, boarded and took a second.

The Battle of Cape St. Vincent won for Jervis the earldom of St. Vincent and for Nelson a knighthood, which coincided with his promotion by seniority to rear admiral. His first action in command of a major independent force, however, was disastrous. During an assault on Tenerife a grapeshot shattered his right elbow, and back in his flagship, the arm was amputated.

Nelson's injury and recovery

Making a slow recovery in London, nursed by his wife, Nelson was agreeably surprised to find that his failure at Tenerife had in no way dimmed the fame of his daring at St. Vincent, and as he wrote, "my reception by John Bull has been just what I wished." In the spring of 1798, he was fit enough to rejoin the Earl of St. Vincent who assigned him to watching a French fleet and some 300 transports waiting to embark an expeditionary force of 35,000 men at Toulon for an unknown destination.

Cruising off the port in his flagship, the "Vanguard," Nelson was struck by a violent northwesterly gale that blew his squadron off station and carried the French well on their way to their destination, Egypt. The British set out in pursuit, first calling at Naples. Hamilton sent a message to say he considered the French to be bound for Malta, but Nelson believed they were going either to Sicily or Egypt and that if they passed by the former, "they are going on their scheme of possessing Alexandria and getting troops to India." Hearing that they had passed Sicily, the British made their best speed for Alexandria. On arrival, they found the harbour empty so, imagining the French objective had been Sicily after all, returned there. But no French were off Sicily, and Nelson was told that they had been seen a month earlier sailing east.

Battle of the Nile, August 1, 1798. Arriving off Alexandria a second time, the British saw the harbour crowded with empty transports and, to the east, their escorting squadron of 13 ships anchored in a defensive line across Abū Qīr Bay near the mouths of the Nile. Once the signal to engage had been hoisted in the "Vanguard," the 14 British captains—Nelson's "Band of Brothers"—flung themselves at the French. With the French ships immobilized, two British could anchor and concentrate their fire on each enemy before moving on to demolish their next target.

Its outcome never in doubt from its beginning at sunset, the battle raged all night. By dawn the French squadron had been all but annihilated. The strategic consequences were immense, and Nelson, who had been wounded slightly on the forehead, took immediate steps to broadcast the news throughout the Mediterranean as well as hastening it to London.

At Naples, the most convenient port for repairs, he was given an ecstatic hero's welcome, stage-managed by Lady Hamilton, who, although she had not seen him for five years, fell half-swooning into his arm exclaiming, "Oh God, is it possible!" A prolonged British naval presence here was useful in supporting the shaky military strength of King Ferdinand, the one major ruler in Italy to be resisting the southward march of the French, who had already taken Rome and deposed the Pope.

Beginning of the liaison with Emma Hamilton

His 40th birthday party in Naples, arranged by Emma Hamilton for the hero and 1,800 guests, established Nelson as a resident celebrity, a role that obviously delighted him. It was in startling contrast to his weary years at sea and to the homely letters from his wife in Norfolk. At home, it would be Rear Adm. Sir Horatio and Lady Nelson, but in Naples it was Mars and Venus.

The love affair between Nelson and Emma Hamilton developed at a time of crisis. With Nelson's encouragement, King Ferdinand had indulged his own fantasies of glory, and openly joining the alliance of Britain, Russia, and Austria against the French, led his own insignificant army to recapture Rome. This was not only a disastrous failure, but the French counteroffensive drove him back to Naples, which itself then fell. Nelson had to evacuate the Neapolitan royal family to Sicily, and at Palermo it became obvious to all that his infatuation with Emma Hamilton was complete. She had not only proved herself indispensable company to him but had supervised the continuing treatment of his wound and convalescence, had introduced him into Neapolitan society, kept him the centre of attention and soothed his irritation at being rewarded for his victory with a barony rather than a viscountcy.

Blockade of Naples. In the summer of 1799, Nelson's squadron supported Ferdinand's successful attempt to recapture Naples with an army led by a popular prelate, Cardinal Ruffo. The French withdrew and the Parthenopean Republic, declared by Neapolitan liberals, collapsed; its supporters surrendered to Ruffo on his promise of free passage to France. On arrival off Naples, Nelson held the prisoners under his guns until the arrival of

Ferdinand, who at once rescinded Ruffo's promise. Savage reprisals began, Nelson himself handing over the Republican admiral for trial and execution. He was to be harshly criticized for these actions, but as he saw it he was loyally supporting allies and no more. His ruthlessness may have been partly induced by the emotional stress resulting from his divided loyalties between Britain and Naples and the cumulative weariness of hard living at sea and a soft life ashore.

Word of his dalliance had reached the Admiralty, and his superiors began to lose patience. Bonaparte had escaped from Egypt to France, and the French still held Malta when Lord Keith, who had replaced St. Vincent as commander in chief, decided that the enemy's next objective would be Minorca. Nelson was ordered to the island with all available ships but refused on the grounds that he expected the threat to be towards Naples. Events justified him, but to disobey orders so blatantly was unforgivable. The Admiralty, also angered by his acceptance of the dukedom of Bronte in Sicily from King Ferdinand, sent him an icy order to return home because "you will be more likely to recover your health and strength in England than in an inactive situation at a Foreign Court."

In 1800 he returned but across the continent in company with the Hamiltons. He cut a poor figure in European cities: "The Lord of the Nile seems as clumsy and dim on land as he is adroit and notable at sea," was one view. But when the curious little party landed in England, it was at once clear that he was the nation's hero and his progress to London was triumphal until the bleak reunion with his wife. Husband and wife were together for about a month when, as the caricaturists made it plain, he returned to the cuckolding of Sir William Hamilton. While the British masses worshipped him, he became a figure of fun or scandal in London society and the King publicly snubbed him.

The nation's hero

Emma was pregnant by Nelson and, in a turmoil of guilt and delight, he was appointed second in command to the elderly and timid Adm. Sir Hyde Parker, who was to command an expedition to the Baltic. This was to take preventive action against the Armed Neutrality of the North, an alliance of Russia, Prussia, Sweden, and Denmark in support of France. Shortly before the expedition sailed, Nelson heard that Emma had borne him a daughter, and he made elaborate arrangements for concealing her parentage. His intense joy at the birth was matched by his jealous frenzy over Emma who, he thought, had caught the eye of the Prince of Wales. At this time he received a touching letter from his wife, pleading for a reconciliation but this was returned, marked, "Opened by mistake by Lord Nelson, but not read."

Battle of Copenhagen. Parker's fleet sailed for the first objective, Copenhagen, early in 1801. At first Nelson's advice was not sought; then, as Danish resistance became increasingly likely, he could record, "Now we are sure of fighting, I am sent for." By the stratagem of taking the fleet's ships of shallower draught through a difficult channel, Nelson bypassed the shore batteries covering the city's northern approaches. The next morning, April 2, he led his squadron into action. There was to be no room for tactical brilliance; only superior gunnery would tell.

The Danes resisted bravely, and Parker, fearing that Nelson was suffering unacceptable losses, hoisted the signal to disengage. Nelson disregarded it, delivering his famous quip, "I have only one eye—I have a right to be blind sometimes." An hour later, victory was his, and the Danish ships lay shattered and silent, their losses amounting to some 6,000 dead and wounded, six times heavier than those of the British.

Before this success could be followed by similar attacks on the other potential enemies, Tsar Paul of Russia died, and the threat faded. Parker was recalled and succeeded by Nelson, who at last became a commander in chief. The Admiralty, well aware of his popular appeal, now made maximum use of it by giving him a home command. At once he planned an ambitious attack on the naval base of Boulogne in order to foil a possible French invasion. He did not take part himself and the operation was a gory failure. A second attempt was abandoned because of peace negotiations with France, and in March 1802 the Treaty of Amiens was signed.

Nelson as commander in chief

At last there was time to enjoy the fruits of his victories. Emma had, on Nelson's instructions, bought an elegant country house, Merton Place, near London, and transformed it into an expensive mirror for their vanities. At last her husband rebelled against his cuckolding, writing to her, "I feel that the whole attention of my wife is given to Lord N. and his interest at Merton." But it was too late for change, and he appeared reconciled to his lot when, early in 1803, he died with his wife and her lover at his side.

Victory and death at Trafalgar. Bonaparte was known to be preparing for renewed war, and two days before it broke out, Nelson, in May 1803, was given command in the Mediterranean, hoisting his flag in the "Victory." Once again he was to blockade Toulon, now with the object of preventing a rendezvous between the French ships there with those at Brest in the Atlantic and, after Spain declared war on Britain, with Spanish ships from Cartagena and Cadiz. A combined force of that size could well enable Bonaparte to invade England; and in 1805, Napoleon, who the previous year had crowned himself emperor, ordered the fleets to converge for this purpose. The French and Spanish squadrons were to burst through the British blockade; run for the West Indies; and after ravaging British possessions and trade, return across the Atlantic in a single invincible fleet to destroy the British near Ushant, an island off Brittany, and take control of the English Channel while it was crossed by an invading army of 350,000.

In March of that year, Adm. Pierre Villeneuve, who was to be in overall command, broke out of Toulon under cover of bad weather and disappeared. Nelson was convinced that Villeneuve was making for the West Indies and set off in pursuit. Villeneuve, alarmed at the speed of the pursuit, cut short his marauding to make for Ushant as planned. A British squadron had been alerted, and Villeneuve's fleet was intercepted and damaged. Failing to win control of the Channel, he ran south to Cadiz.

Preparations for battle

Nelson put into Gibraltar, made dispositions for the blockade of Cadiz, and returned to England. During his 25 days at home, he planned the strategy for the confrontation with the Franco-Spanish fleets that now seemed inevitable. Thirty-four enemy ships were blockaded in Cadiz by smaller numbers under Adm. Cuthbert Collingwood. Although Napoleon, abandoning the plan of a cross-Channel invasion, began to re-deploy the Grand Army, in Britain the danger of invasion seemed as pressing as ever, and Nelson appeared the country's hope of deliverance.

While awaiting orders to sail, Nelson by chance met the Duke of Wellington, the future victor of Waterloo, in Whitehall. At first Wellington was shocked by the admiral's conversation "all about himself, and in, really, a style so vain and silly as to surprise and almost disgust me." Then Nelson, discovering his companion's identity, changed and "talked of the state of the country and of the aspect and probabilities of affairs on the Continent with a good sense, and a knowledge . . . that surprised me. . . . I don't know that I ever had a conversation that interested me more."

When his orders came, Nelson made his emotional farewells to Emma and Horatia, his daughter by her, and, on September 15, sailed from Spithead in the "Victory." He was now at the height of his professional powers. Worshipped by his officers and sailors alike, he was confident that his captains understood his tactical thinking so well that the minimum of consultation would be required. On his 47th birthday he dined 15 captains in his flagship and outlined his plans to bring on "a pell-mell battle" in which British gunnery and offensive spirit would be decisive. He planned to advance on the Franco-Spanish fleets in two divisions to break their line and destroy them piecemeal. This was the final abandonment of the traditionally rigid tactics of fighting in line of battle. "It was new—it was singular—it was simple!" wrote Nelson. He described his original idea as "The Nelson Touch."

After receiving Napoleon's orders that he must break the blockade, Villeneuve, on October 20, sailed out of Cadiz with 18 French and 15 Spanish men-of-war. At dawn next day, the Franco-Spanish fleets were silhouetted against the sunrise off Cape Trafalgar, and the British began to form the two divisions in which they were to fight, one led by Nelson, the other by Admiral Collingwood.

Before battle, Nelson signed a document bequeathing Emma Hamilton and Horatia to the nation and wrote in his diary a prayer; perhaps the finest of much good prose that he had written. As the opposing fleets closed, he made a signal, "England expects that every man will do his duty." The battle raged at its fiercest round the "Victory." Although broadsides and small-arms fire were being exchanged at point-blank range, Nelson and his flag captain, Thomas Hardy, paced the quarterdeck, the former wearing his four stars of chivalry on his coat, so making himself an obvious target. A French sniper, firing from the mast of the "Redoutable," shot Nelson through the shoulder and chest. He was carried below to the surgeon, and it was soon clear that he was dying. The future of Emma and Horatia was much on his mind, but he listened attentively to news of the battle. When told that 15 enemy ships had been taken, he replied, "That is well, but I had bargained for 20." Hardy kissed his forehead in farewell and Nelson spoke his last words, "Now I am satisfied. Thank God, I have done my duty." At 4:30 PM he died.

"England expects . . ."

Although the victory of Trafalgar finally made Britain safe from invasion, it was, at the time, overshadowed by the news of Nelson's death. A nation racked with grief gave him a majestic funeral in St. Paul's Cathedral, and his popularity was recorded in countless monuments, streets, and inns named after him, paintings and prints, busts and plaques and, eventually, in the preservation at Portsmouth of the "Victory." Emma Hamilton and his daughter, however, were ignored. Emma died, raddled and almost destitute, in Calais nine years later. Horatia, showing her father's resilience, married a clergyman in Norfolk and mothered a large and sturdy family.

Nelson had finally broken the unimaginative strategical and tactical doctrines of the previous century and taught individual officers to think for themselves. He had introduced as normal practice a new personal quality of leadership in which subordinates were treated as individual human beings.

BIBLIOGRAPHY. A.T. MAHAN, *The Life of Nelson*, 2 vol. (1897), a biographical study by the great American naval historian, should be read in conjunction with his other works, such as *The Influence of Sea Power upon the French Revolution and Empire* and *The Influence of Sea Power upon History*. GEORGE P.B. NAISH (ed.), *Nelson's Letters to His Wife, and Other Documents, 1785–1831* (1958), is far more comprehensive than its title suggests since it includes many other letters and long explanatory passages by the editor, himself a noted British Naval historian working on the staff of the National Maritime Museum, Greenwich. SIR NICHOLAS HARRIS, *The Dispatches and Letters of Vice Admiral Lord Viscount Nelson*, 7 vol. (1844–46), remains the principal source for Nelson's biographers. CAROLA OMAN, *Nelson* (1947), a perceptive and intelligent biography, is accurate and well written but perhaps a little emotional for some tastes. TOM POCOCK, *Nelson and His World* (1968), is a short illustrated biography, laying particular stress on Nelson's early influences and their likely effects in later life. ROBERT SOUTHEY, *The Life of Nelson*, 2 vol. (1813), one of the first and the most reliable of the many biographies, treats with delicacy the more intimate aspects of Nelson's private life. OLIVER WARNER, *A Portrait of Lord Nelson* (1958), is perhaps the best all-round biography of Nelson, written with great personal and professional understanding by one of the most distinguished living British naval historians; GEOFFREY BENNETT, *Nelson the Commander* (1972), relates the Admiral's career to the naval practice of his own and later times.

(T.Po.)

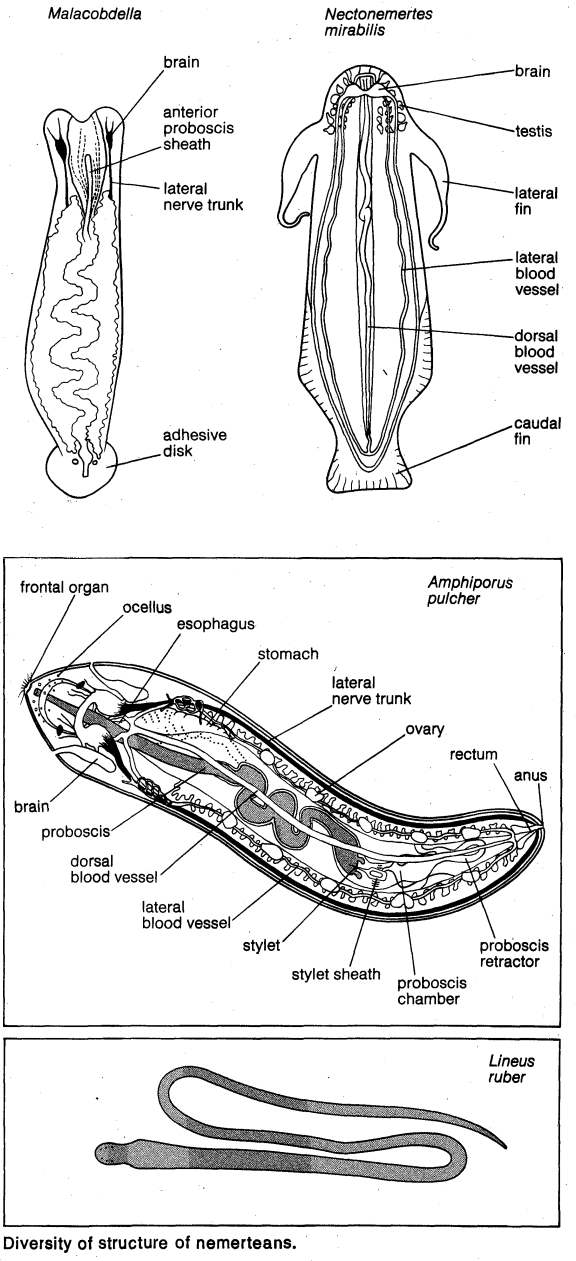
Nemertea

The ribbonworms comprise the phylum Nemertea, which sometimes is called Nemertinea or Rhynchocoela. Nemerteans are the simplest animals to have a digestive system that begins with a mouth and ends with an anus and a true vascular system for blood circulation. All of the approximately 600 species are quite similar in appearance. Typically, nemerteans are long and thin, with little distinction between head and tail. Although they range in length from less than 1 centimetre to more than 30 metres, most nemerteans are less than 20 centimetres in length. Ribbonworms, usually white, yellow, or brown in colour, may be brightly patterned.

Natural history. The majority of ribbonworms are marine and inhabit the coastal waters of temperate regions, particularly the Mediterranean and Europe. Nemerteans occupy a variety of habitats, including the undersides of stones and seaweeds and burrows in sand or mud. A few nemerteans thrive at depths greater than those of coastal waters, and some live in the open sea. Ribbonworms are uncommon in tropical and subtropical regions; a variety of species lives in the Arctic and Antarctic, however.

One genus, *Prostoma,* occurs in fresh water throughout the world; another, *Geonemertes,* is terrestrial in subtropical areas, particularly Australia and New Zealand. Ribbonworms, carnivores that feed on living or dead animals (particularly annelid worms), actively pursue prey, entan-

From *Invertebrate Zoology* by Paul A. Meglitsch. Copyright © 1967 by Oxford University Press, Inc. Reprinted by permission



Diversity of structure of nemerteans.

gle it with their proboscis, and sometimes inject a poisonous substance into it.

Reproduction

Male and female nemerteans occur in most species, with annual reproduction typical. Eggs and sperm are released separately, and fertilization takes place externally. The fertilized egg develops by a process similar to that of flatworms (phylum Platyhelminthes), annelids (phylum Annelida), and mollusks (phylum Mollusca). Ribbonworms can develop in one of two ways: the most common is the direct method—without a larval stage; the other is indirect and includes a larva, which may be one of two types. One kind, a free-swimming larva of complicated pattern known as the pilidium, is more common; the other type, similar to an adult, is called Desor's larva. Larvae change into young ribbonworms. Within the genera *Prostoma* and *Geonemertes*, the species may be either dioecious (*i.e.*, separate male and female animals) or hermaphroditic (*i.e.*, male and female reproductive organs in one animal). All ribbonworms have the ability to regenerate lost or damaged parts of their bodies; some species actually break up and form a number of fragments, which then grow into complete individuals. This mechanism provides asexual reproduction.

Form and function. The long, flat ribbonworm is characterized by a proboscis that lies in a sheath, which extends down the body dorsal to the gut. The wall of the sheath is continuous with the layers of the body wall. The proboscis, which is eversible (*i.e.*, can be turned outward), is attached to the posterior end of the proboscis sheath by a retractor muscle and may have one or more piercing stylets at its tip. The muscular body wall consists of two or three layers of circular or longitudinal muscles; the arrangement of the layers varies in different groups. The head often bears a number of eyes and a mouth that leads to a straight gut, which ends in an anus at the posterior end of the animal. The nervous system consists of an anterior brain; typically, a single dorsal and two lateral nerves extend the length of the body. The closed blood vascular system usually consists of a mid-dorsal vessel and two lateral ones.

Evolution and classification. The affinities of the nemerteans lie with the flatworms (phylum Platyhelminthes); both have similar types of embryonic development and the same basic body plan. Unlike the flatworms, however, the nemerteans have a complete gut and a circulatory system. The ribbonworms generally are regarded as a distinct phylum at the highest point of acoelomate (without a body cavity) development. The important features used to classify the nemerteans include the position of the brain relative to the mouth, the presence or absence of a stylet (or stylets) on the proboscis, and the position of the lateral nerves relative to the muscle layers. The classification that is outlined below has been in use for some time.

PHYLUM NEMERTEA (ribbonworms)
Elongated, muscular, nonsegmented worms.

Class Anopla
Brain anterior to the mouth, proboscis without stylets.
Order Palaeonemertini
Body wall with 2 or 3 muscle layers, the outer one circular.
Order Heteronemertini
Body wall with 3 muscle layers, the outer one longitudinal.

Class Enopla
Brain posterior to, or at same level as, mouth.
Order Hoplonemertini
Proboscis with 1 or more stylets.
Order Bdellonemertini
Proboscis with no stylets; single genus with less than 10 species, commensal in mollusks.

BIBLIOGRAPHY. L.H. HYMAN, "Platyhelminthes and Rhynchocoela," in *The Invertebrates*, vol. 2 (1951), the standard text; R.D. BARNES, *Invertebrate Zoology*, 2nd ed. (1968), a brief but excellent general account of the group based largely on Hyman's major text; P.P. GRASSE, "Platyhelminthes, Mésozoaires, Acanthocéphales, Némertiens," in *Traité de zoologie*, vol. 4, pt. 1 (1961), a beautifully illustrated French volume.

(F.E.G.C.)

Nepal

The Kingdom of Nepal, lying along the southern slopes of the Himalayan mountain ranges, is a landlocked country located between Sikkim and India to the east, India to the south and west, and the Tibetan Autonomous Region of China to the north. Its territory, which has an area of 54,362 square miles (140,798 square kilometres), extends roughly 500 miles from east to west and 90 to 150 miles from north to south. Nepal, for long under the rule of hereditary prime ministers favouring a policy of isolation, remained closed to the outside world until a palace revolt in 1950 restored the king to his position of authority in 1951; the country gained admission to the United Nations in 1955. In 1960, King Mahendra assumed direct rule, and on his death in 1972, King Birendra succeeded his father to the throne. Wedged in between two giants, India and China, Nepal seeks to keep a balance between the two countries in its foreign policy—and thus to remain independent. A factor that contributes immensely to the geopolitical importance of Nepal is the fact that a strong Nepal can deny China access to the rich Gangetic Plain; Nepal thus marks the southern boundary of the Chinese sphere north of the Himalayas in Asia.

As a result of its years of geographic and self-imposed isolation, Nepal is one of the least developed nations of the world. In recent years India, China, the Soviet Union, the United States, the United Kingdom, Israel, New Zealand, and Switzerland have provided economic assistance to Nepal. The extent of foreign aid to Nepal has been influenced to a considerable degree by the strategic position of the country between India and China. (For associated physical features, see EVEREST, MOUNT and HIMALAYAN MOUNTAIN RANGES; for historical aspects, see NEPAL, HISTORY OF.)

THE NATURAL ENVIRONMENT

Relief and drainage patterns. Nepal contains some of the most rugged and difficult mountainous terrain in the world. Roughly 75 percent of the country is covered by mountains; along the southern border a belt of lowland comprises the only extensive area of flat, level land. From the south to the north, Nepal can be divided into four main physical belts, each of which extends east to west across the country, These are, firstly, the Tarai, a low, flat, fertile land adjacent to the border of India; secondly, the forested Churia foothills and the Inner Tarai zone, rising from the Tarai plain to the rugged Mahābhārat Lekh range; thirdly, the mid-mountain region between the Mahābhārat Lekh range and the Great Himalayas; and fourthly, the Great Himalaya Range, rising to more than 29,000 feet.

The Tarai forms the northern extension of the Gangetic Plain and varies in width from less than 16 to more than 20 miles, narrowing considerably in the Rāpti and Nārāyani administrative zones. A ten-mile-wide belt of rich agricultural land stretches along the southern part of the Tarai; the northern section, adjoining the foothills, is a marshy region that is inhabited by wild animals and in which malaria is endemic.

The Churia Hills, which are sparsely populated, rise in almost perpendicular escarpments to an altitude of more than 4,000 feet. Between the Churia Hills to the south and the Mahābhārat Lekh range to the north, there are some broad basins from 2,000 to 3,000 feet high, about ten miles wide, and 20 to 40 miles long; these basins are often referred to as the Inner Tarai. In many places they are being cleared of the forests and savanna grass to provide timber and areas for cultivation.

A complex system of mountain ranges some 50 miles in width and varying in elevation from 8,000 feet to 14,000 feet, lie between the Mahābhārat Lekh range and the Great Himalayas. The ridges of the Mahābhārat Lekh range present a steep escarpment toward the south and a relatively gentle slope to the north. To the north of the Mahābhārat Lekh range, which encloses the valley of Kāthmāndu, are the more lofty ranges of the Inner Himalaya (Lesser Himalaya), rising to perpetually snow-covered peaks. The Kāthmāndu and the Pokhara valleys

lying within this mid-mountain region are flat basins, formerly covered with lakes, formed by the deposition of fluvial and fluvio-glacial material brought down by rivers and glaciers from the enclosing ranges during the four glacial and intervening warm phases of the Pleistocene Epoch (from about 2,500,000 to 10,000 years ago).

The Kāthmāndu Valley

The Kāthmāndu Valley, the political and cultural hub of the nation, is drained by the Bāghmati River, flowing southward, which washes the steps of the sacred temple of Paśupatināth (Pashupatinath) and rushes out of the valley through the deeply cut Chhobar gorge. The Kāthmāndu International Airport (unofficially known as Gauchar Airport) is situated on the flat top of lacustrine deposits. Some sandy layers of the lacustrine beds act as aquifers (water-bearing strata of permeable rock, sand, or gravel), and springs occur in the Kāthmāndu Valley where the sands outcrop. The springwater often gushes out of dragon-shaped mouths of stone made by the Nepalese; it is then collected in tanks for drinking and washing and also for raising paddy nurseries in May, before the advent of the monsoon. Drained by the Seti River, the Pokhara Valley, 96 miles west of Kāthmāndu, is also a flat lacustrine basin. There are a few remnant lakes in the Pokhara basin, the largest being the Phewa Tāl, which is about two miles long and nearly a mile wide. North of the basin lies the Annapūrna massif of the Great Himalaya Range.

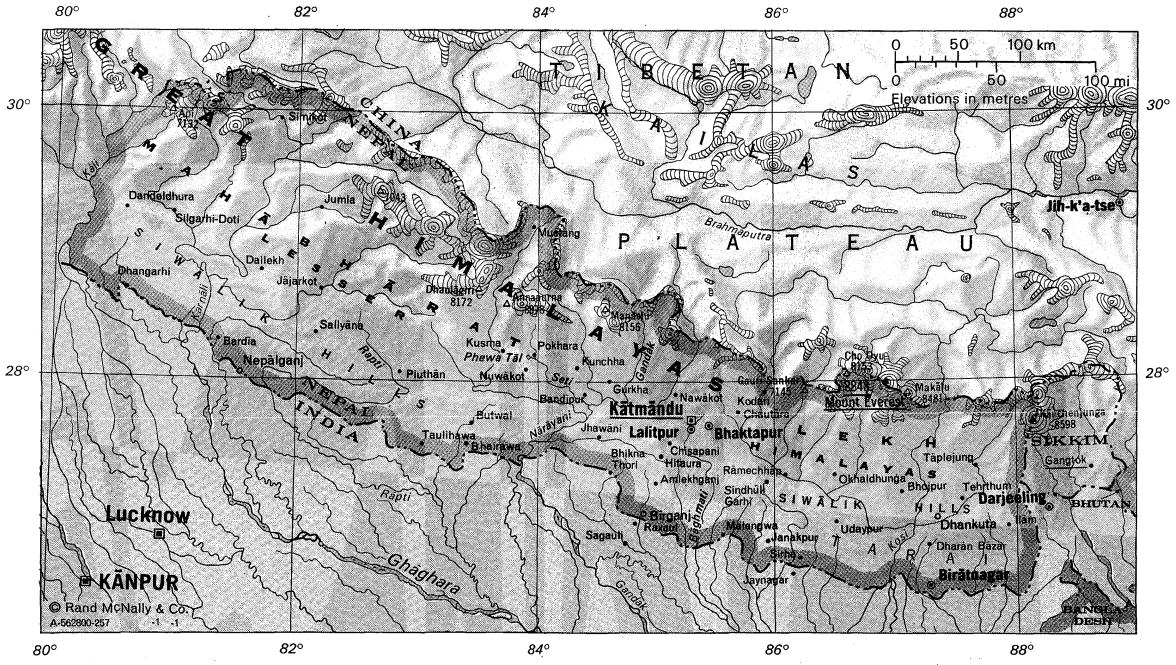
The Great Himalaya Range, ranging in elevation from 14,000 feet to more than 29,000 feet, contains 5 of the world's 13 highest peaks—Everest, Kānchenjunga, Makālu, Dhaulāgiri, and Annapūrna—that are above 26,400 feet. Except for scattered settlements in high mountain valleys, this entire area is uninhabited, and all areas over 16,000 feet in elevation are permanently covered with snow.

The major rivers of Nepal—the Kosi, Gandak, and Karnāli, running southward across the strike of the Himalayan ranges—form transverse valleys with deep gorges, which are generally several thousand feet in depth from the crest of the bordering ranges. The watershed of these rivers lies not along the line of highest peaks in the Himalayas but to the north of it, usually in Tibet.

In the upper courses of all Nepalese rivers, which run through mountain regions, there are little or no flood problems. In low-lying areas of the Tarai plain, however, floods are a serious problem.

The rivers and small streams of Tarai, especially those in which the dry season discharge is small, are polluted by large quantities of domestic waste thrown into them. Although no records exist to determine the rate of increase in pollution, it is certain that pollution of Nepal's rivers is increasing. Towns and villages are expanding without proper provision for sewage disposal facilities, and more industries are being established at selected centres in the Tarai. The polluted surface water in the Kāthmāndu and Pokhara valleys, as well as in the Tarai, are unacceptable for drinking purposes.

Climate, vegetation, and animal life. Nepal's climate, influenced by elevation as well as by its location in a subtropical latitude, ranges from subtropical monsoon conditions in the Tarai, through a warm temperate climate between 4,000 to 7,000 feet in the mid-mountain



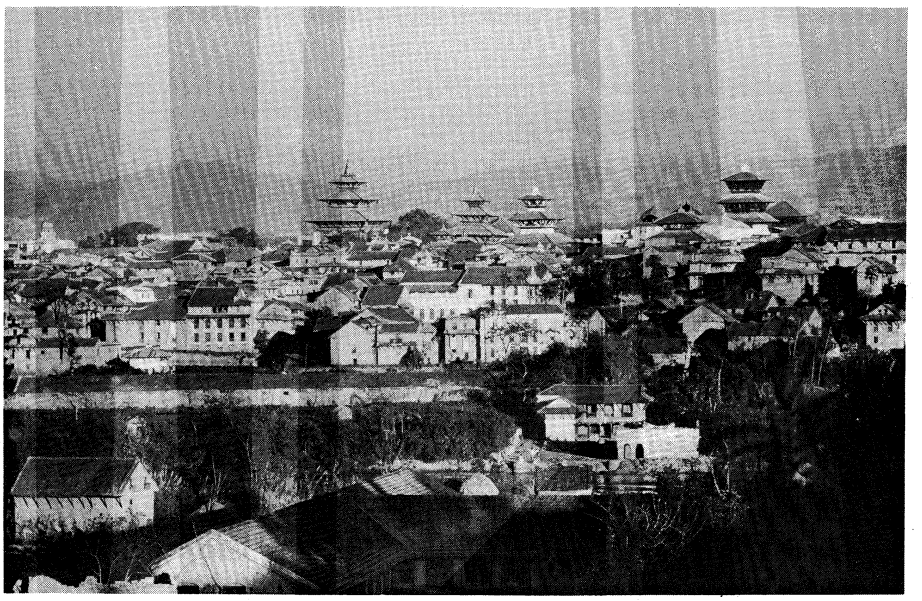
NEPAL
MAP INDEX

Cities and towns

Amlekhganj.....27·17n 85·00e
Bandīpur.....27·56n 84·25e
Bardīa.....28·18n 81·23e
Bhairawa.....27·31n 83·24e
Bhaktapur.....27·42n 85·27e
Bhikna Thori....27·21n 84·37e
Bhojpur.....27·11n 87·02e
Birātnagar.....26·18n 87·17e
Bīrganj.....27·01n 84·54e
Butwal.....27·42n 83·28e
Chautāra.....27·47n 85·43e
Chisapani.....27·34n 85·08e
Dailekh.....28·50n 81·43e
Dandeldhura....29·17n 80·36e
Dhangarhi.....28·41n 80·38e
Dhankuta.....26·59n 87·21e
Dharān Bāzār....26·49n 87·17e
Gurkha.....28·01n 84·38e
Hitaura.....27·25n 85·05e
Ilām.....26·55n 87·56e
Jājarkot.....28·42n 82·14e
Janakpur.....26·44n 85·56e
Jhawāni.....27·35n 84·38e
Jumla.....29·17n 82·13e
Kātmāndu.....27·42n 85·20e
Kodari.....27·56n 85·56e
Kunchha.....28·08n 84·22e
Kusma.....28·13n 83·41e
Lalitpur.....27·41n 85·20e
Malangwa.....26·52n 85·34e
Mustāng.....29·11n 83·57e
Nawākot.....27·55n 85·10e
Nepālganj.....28·03n 81·38e
Nuwākot.....28·08n 83·53e
Okhaldhunga....27·19n 86·31e
Piuthān.....28·06n 82·54e
Pokhara.....28·12n 83·59e
Rāmechhāp.....27·20n 86·05e
Sallyāna.....28·22n 82·12e
Silgarhi-Doti....29·16n 80·58e
Simikot.....29·58n 81·51e
Sindhūli Garhi...27·16n 85·58e
Sirha.....26·39n 86·12e
Tāplejung.....27·21n 87·40e
Taulihawa.....27·32n 83·05e
Tehrthum.....27·07n 87·32e
Udaypur.....26·54n 86·32e

Physical features and points of interest

Annapūrna, *mountain*.....28·34n 83·50e
Api, *mountain*...30·01n 80·56e
Bāghmati, *river*..27·00n 85·20e
Cho Oyu, *mountain*.....28·06n 86·39e
Churia Range, see Siwālik Hills
Dhaulāgiri, *mountain*.....28·42n 83·31e
Everest, Mount, *mountain*.....27·59n 86·56e
Gandak, *river*....27·30n 84·00e
Gauri Sankar, *mountain*.....27·57n 86·21e
Great Himalayas, *mountain range*.....28·30n 84·30e
Kāli, *river*.....29·25n 80·15e
Kānchenjunga, *mountain*.....27·42n 88·09e
Karnāli, *river*.....28·25n 81·10e
Kosi, *river*.....26·30n 86·55e
Lesser Himalayas, see Mahābhārat Lekh
Mahābhārat Lekh (Lesser Himalayas), *mountains*.....28·20n 84·00e
Makālu, *mountain*.....27·54n 87·06e
Manāslu, *mountain*.....28·33n 84·35e
Nārāyani, *river*...27·30n 84·10e
Phewa Tāl, *lake*.....28·13n 83·57e
Seti, *river*.....28·00n 84·10e
Siwālik Hills.....27·40n 83·40e
Tarai, *physical region*.....26·30n 88·00e



Kāthmāndu, capital of Nepal, with several of its famous pagoda-shaped temples. The Himalayas are in the background.
By courtesy of the National Geographic Society, © National Geographic Society

region, to cool temperate conditions in the higher parts of mountains between 7,000 and 11,000 feet, to an Alpine climate at altitudes between 14,000 feet and 16,000 feet along the lower slopes of the Himalaya mountains. At altitudes above 16,000 feet the temperature is always below freezing and the surface covered by snow and ice. Rainfall is adequate in the eastern portion of the Tarai (which receives from 70 to 75 inches a year at Morang) and mountains, but the western portion of Nepal (where from 30 to 35 inches a year fall at Kanchanpur) is drier.

Temperatures

In Kāthmāndu Valley, average temperatures range from 50° F (10° C) in January to 78° F (26° C) in July, and the lowest and highest temperatures recorded have been 27° and 99° F (−3° and 37° C). The average annual rainfall is about 58 inches, most of which falls in the period from June to September. At Pokhara temperature ranges from 40° F (4° C) in January to approximately 100° F (38° C) in June, just before the monsoon. In winter, temperature during the day will rise to 70° F (21° C), creating pleasant climatic conditions with cool nights and warm days. Rainfall is quite heavy (about 100 inches) in Pokhara valley because warm rain-bearing monsoon winds discharge most of their moisture as they encounter the Annapūrna range.

Natural vegetation of Nepal follows the pattern of climate and altitude. A tropical, moist zone of deciduous vegetation occurs in southern Nepal, in the Tarai and the Churia Hills. These forests consist mainly of khair (*Acacia catechu*), a spring tree with yellow flowers and flat pods; sissoo (*Dalbergia sissoo*), an East Indian tree yielding dark-brown durable timber; and sal (*Shorea robusta*), an East Indian timber tree with foliage furnishing food for lac insects (which deposit lac, a resinous substance used for the manufacture of shellac and varnishes, on the tree's twigs). On the Mahābhārat Lekh Range, at elevations between 5,000 to 10,000 feet, vegetation consists of a mixture of many species, chiefly pines, oaks, rhododendrons, poplars, walnuts and larch. Between 10,000 and 12,000 feet, fir mixed with birch, as well as rhododendron, abound. In the mid-mountain region of Nepal a fairly dense population has cleared all but the most inaccessible and remote parts of the forest, which are restricted to areas of steep slopes and rocky terrain. Similarly all readily accessible parts of valuable sal forest in the Terai have been devasted by overcutting and deplete practices. The vast forested area below the timber line in the Great Himalaya Range bears some of the most valuable forests in Nepal, containing spruce, fir, cypress, juniper, and birch. Alpine vegetation occupies higher parts of the Great Himalaya Range. Just below the snow line, grassy vegetation affords favourable grazing ground in summer between 14,000 and 15,000 feet. (P.P.K.)

Animal life

The forested areas of the Tarai are the home of tigers and leopards, gaurs (wild ox), occasional elephants and buffalo, and many deer; the deer include chital, or axis, deer (which have white-spotted bodies), sambar (a large Asiatic deer with coarse hair on the throat and strong antlers), and swamp deer. The Lesser Rāpti Valley, in the Chitawan district, is one of the last homes of the great Indian rhinoceros (*Rhinoceros unicornis*), of which only about 700 are believed to exist, 300 of these being in Nepal. Much poaching has gone on, as the horn of the rhinoceros is reputed to be valuable as an aphrodisiac, but in the 1960s the Nepal government was organizing protective measures.

There are few wild animals in the central zone, because of the clearing of forests. Occasional leopards, bears, and the smaller carnivores inhabit the forests and ravines, and muntjacs (a kind of small deer, also called the barking deer) are found in the woods. In the Alpine zone are musk deer, much persecuted for the musk pods they carry, the tahr (a Himalayan beardless wild goat), the goral (any of several goat antelopes, closely related to the Rocky Mountain goat), and wild sheep, which are preyed upon by wolves and snow leopards. Pheasant are common. The Yeti (bear-man, or Abominable Snowman) is said by the Sherpas to inhabit the high snow mountains but has eluded discovery by several expeditions that have set out to solve the mystery of this fable. Strange tracks are often found in the snow, but opinion inclines to the belief that they were probably made by bears. River wildlife includes the mahseer, a large Indian freshwater food and sport fish. (R.R.Pr./M.Zu.)

THE PEOPLE

Groups historically associated with the contemporary country. The large-scale migrations of Mongoloid groups from Tibet and Indo-Aryan people from northern India, which accompanied the early settlement of Nepal, have produced a diverse linguistic, ethnic, and religious pattern. Nepalese of Indo-Aryan ancestry comprise the people of the Tarai, the Pahari, the Newars, and the Tharus, numbering nearly 7,800,000 persons, or nearly 80 percent of the total population. Indo-Aryan ancestry has been a source of prestige in Nepal for centuries, and the ruling families have been of Indo-Aryan and Hindu background. Most of the Tibeto-Nepalese groups—the Tamang, Rai, Limbu, Bhote, and Sunwar—live in the north and east, while the Magar and Gurung inhabit west central Nepal. The bulk of the famous Gurkha contin-

gents in the British army have come from the Magar, Gurung, and Rai groups.

The main and official language of Nepal is Nepali (Gorkhali), spoken in the Tarai and the mid-mountain region. Nepali, a derivative of Sanskrit, belongs to the Indo-Aryan branch of the Indo-European family. There are a number of regional dialects found in the Tarai and mountain areas. The languages of the north and east belong predominantly to the Tibeto-Burman family. These include Magar, Gurung, Rai, Limbu, Sunwar, Tamang, Newari, and a number of Bhote dialects, including Sherpa and Thakali. Although Newari is commonly placed in the Tibeto-Burman family, it was influenced by both Tibeto-Burman and Indo-European languages.

Nepal, Area and Population*

	area		population	
	sq mi	sq km	1961 census†	1971 census‡
Geographic regions				
Himalaya Mountains				
Census regions				
Eastern Hills	10,114	26,195	1,887,000	2,770,000
Far Western Hills	18,902	48,956	1,698,000	1,954,000
Kāthmāndu	218	565	460,000	333,000
Western Hills	11,144	28,863	1,947,000	1,570,000
Inner Tarai				
Census regions				
Central Inner Tarai	2,176	5,636	244,000	324,000
Eastern Inner Tarai	1,829	4,737	194,000	112,000
Western Inner Tarai	714	1,849	99,000	474,000
Tarai				
Census regions				
Eastern Tarai	5,115	13,248	2,213,000	2,900,000
Far Western Tarai	2,843	7,363	272,000	413,000
Western Tarai	1,307	3,385	400,000	439,000
Total Nepal	54,362	140,798§	9,413,000‖	11,289,000

*Because area and population data were not available according to administrative subdivisions, this table is organized according to the geographic census regions of the 1961 census. The actual administrative unit is the zone (anchal). There are 14 zones, namely Bāgmati, Bheri, Dhaulagiri, Gandaki, Janakpur, Karnali, Kosi, Lumbini, Mahakali, Mechi, Nārāyani, Rāpti, Sagarmatha, and Seti. †De jure. ‡Preliminary figures. §Converted area figures do not add to total given because of rounding. ‖Figures do not add to total given because of rounding.
Source: Official government figures.

Demography. Although population enumerations have been made approximately every ten years since 1911, the early census data of Nepal are not reliable. It was not until the early 1950s that the first systematic enumeration was made, and even then the data were limited in reliability and scope. The 1961 census recorded a population of 9,413,000. At the 1971 census the population was nearly 11,300,000; the rate of growth is estimated at 2 percent, which is expected to result in a total population figure of more than 12,000,000 by 1975.

Population density

The average population density of Nepal is 208 persons per square mile. If the area under snow and Alpine meadow (about 22 percent), timber (about 31 percent), and other unproductive land (about 12 percent) is subtracted from the total land area, the density on the area under cultivation or capable of being cultivated rises to 509 per square mile. Some 56 percent of the population lives in the mid-mountain region including the high Himalayan country (where 27 percent of the cultivated land is located mostly in valleys and terraced hill slopes); 5 percent lives in the Kāthmāndu Valley (with 2 percent of the cultivated land); and 42 percent lives in the Tarai (with 71 percent of the cultivated land).

The 1971 census recorded 26,478 more males than females in the total population, making the ratio of males to females about equal. Accurate data on mortality and birth rates are not available, but the death rate is estimated at 23 per 1000 and the birth rate about 45 per 1000. The high mortality rate will probably decrease as pure drinking water facilities become increasingly available, in addition to the mass-inoculation programs. In the early 1970s it appeared that Nepal's rate of population growth would remain high and that pressure of population on land would become more intense as the population increased in areas that were already overcrowded.

Almost all Nepalese live in villages or in small market centres. According to a 1970 projection, there were about 518,000 urban dwellers, representing less than 5 percent of the total population. Outside of Kāthmāndu itself (population 153,000), there are no major cities, though there are several smaller urban centres (Birātnagar, 45,000; Nepālganj, 36,000; Bīrganj, 20,000), located in the Tarai along the Indian border, in addition to a few townships—such as Hitaura, Butwal, and Dharān Bāzār —which are beginning to emerge in the foothills and hill areas, where economic activity is developing.

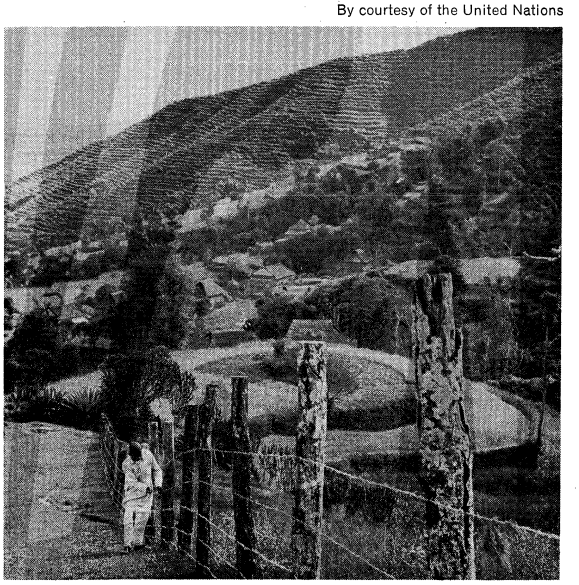
THE NATIONAL ECONOMY

Landlocked, lacking substantial resources for economic development, heavily dependent for basic materials and the market for its export, and with an inadequate transportation network, Nepal is one of the least developed nations in the world. Slow changes have occurred in the structure of the economy in the 1950s and 1960s, following the 1951 overthrow of the Rana regime. The lack of change in the political and administrative structure is largely responsible for the slow rate of economic growth. A tentative official estimate placed the 1961 to 1962 gross domestic product (based on market prices) at Nepalese rupees (NRs.) 3,680,000,000, giving a per capita income of about NRs. 390, or $51 U.S. In 1969, the gross domestic product was estimated at NRs. 9,449,000,000 (10.12 NRs = $1U.S.; 26.43NRs. = £1 sterling, on December 1, 1970), giving a per capita income of $80. The rapid increase in gross domestic product between 1961–62 and 1969 was the result of an increase in prices in the period. The increase in real output over the period is estimated at between 6 and 7 percent.

Resources and resource exploitation. *Agriculture.* Agriculture engages about 90 percent of Nepal's population and accounts for nearly 70 percent of the gross domestic product and over three-fifths of the country's export earnings. Yet agricultural productivity is very low. The average yield of paddy (rice), the main crop in Nepal, is about 1,713 pounds per acre; of maize (corn), the chief food item of the people in the mid-mountain region, 1,427 pounds per acre; and of wheat 892 pounds per acre.

Crop yields

By courtesy of the United Nations



Boundary of a reforestation project near Kāthmāndu sponsored by the government of Nepal. Over 240 acres of eucalyptus and acacia trees were planted here to help maintain one of the country's primary resources.

The low yield results from poor quality of seed, low fertilizer utilization, and use of inefficient techniques. Also, since only 4 percent of Nepal's cultivated land area of about 5,000,000 acres is under irrigation, output depends upon the vagaries of weather.

On the whole Nepal has a small surplus in food grains. There are, however, major dislocations in supply and demand. Periods of shortage between harvests of various crops occur in the mountain areas. At the same time, substantial amounts of food grain are moved to India from the Tarai. Because of the lack of adequate transportation, surplus food grain from the Tarai does not move north into the food deficit areas of the mid-mountain region. Some food grains move northward from the Tarai and the mountain areas into Tibet, despite a shortage in the mountain regions.

The greatest potential for future increases in agricultural production is in the Tarai. In the mid-mountain region the potential for increasing agricultural production is limited. Because of the high population concentration in this region, almost all land capable of cultivation is tilled. Increasing the cultivated land area by cutting into standing forests would aggravate erosion, resulting in reduced yields and land losses by landslides. Thus, in the early 1970s, the shortage of food in this region was expected to become even more severe in the near future.

Forests. About one-third of Nepal's total area is under forest; most of this area is state owned. In spite of overcutting and poor management, timber represents one of the country's most valuable resources and is a major source of potential revenue. Exports of forest products constitute an important source of Indian rupees. Almost all timber is exported to India. The saw mills of the Timber Corporation of Nepal, a government-owned lumber-processing concern, supply Kāthmāndu Valley with construction and furniture wood.

Minerals. Nepal's mineral resources are small, scattered, and barely developed. There are known deposits of coal, iron ore, pyrite (used for making sulfuric acid), limestone, and mica, but they are too small to warrant major investment for exploitation. Nepal's great river systems provide immense potential for hydroelectric development. If developed and utilized within the country and exported to India (the principal market for power generated in Nepal), it could become a mainstay of the country's economy.

Industry and trade. Industrial production represents a small but growing segment of economic activity. Most industries are small localized operations based on the processing of agricultural products. The jute industry, centred in Birātnagar, is Nepal's principal earner of foreign exchange. Sugar factories are located in Birātnagar, Bīrganj, Bhairawa, and Nepālganj. There is a cigarette factory at Janakpur. There are several saw mills in Hitaura, and a number of rice and oil mills are located in the Tarai. In general, there are more industrial enterprises in the private than in the public sector. In the early 1970s a pattern of selective industrial development was expected to continue, based on the limited capital and managerial resources available. The main areas of manufacturing concentration were Birātnagar, the Bīrganj-Hitaura corridor, and the Kāthmāndu Valley.

Tourism

Tourism represents a small but rapidly expanding industry. At present, foreign tourism is primarily confined to the Kāthmāndu Valley, which is the only area equipped with the necessary hotels, food supplies, roads, and international transport services. There are, however, many areas outside the Kāthmāndu Valley ripe for the development of tourism; these include Pokhara, the Mt. Everest area, and the Nārāyani area (where big game exists).

For geographical and historical reasons, nearly all of Nepal's trade is with India. Attempts have been made to diversify trade by new trade agreements with Pakistan, the United States, Poland, China, and the Soviet Union. The state trading agency, National Trading Limited, is expanding its activities by fostering the development of commercial entrepreneurial activity, which has been limited. Large-scale commercial activity has hitherto been in the hands of foreigners, primarily Indians.

Transportation. Transport facilities in Nepal are very limited; there are few independent nations in the world with such a low density of road mileage (11 miles of permanent road per 100,000 inhabitants) or such a high proportion of persons per vehicle (2,000 persons per vehicle). Nepal has only about 5,000 registered motor vehicles and a paved road mileage of only 424 miles to serve nearly 11,300,000 people. The main means of transportation has been the network of footpaths, which interlace the mountain terrain as well as the valleys. Trails have evolved into main trade routes, which tend to follow the river systems. Lack of adequate bridges across these rivers, however, has reduced the usefulness of many of the trails. In 1956, the Indian-made road from the border town of Raxaul to Kāthmāndu was completed; the Chinese have also built a 65-mile highway from Kāthmāndu to Kodari on the Tibetan frontier. In 1968, the Indian-built 130-mile highway from Sunauli, on the Indian border, to Pokhara was opened to traffic. In the early 1970s, the Chinese were assisting in the construction of a highway from Kāthmāndu to Pokhara. Furthermore, the East–West highway, a long cherished dream to unite the various sectors of the Tarai and thus help unify the nation, was under construction with foreign aid.

The meagre road-transport facilities in Nepal are supplemented by few railway- and air-transport links. Increased use of road transport has reduced the significance of the two narrow-gauge railroads that run from Amlekhganj to Raxaul (India) and from Janakpur to Jaynagar (India). The Royal Nepal Airline Corporation, an autonomous government agency, is the only commercial airline. Together with Indian Airlines, it operates flights from Kāthmāndu to Patna, Calcutta, Vārānasi, and Delhi in India. Domestic air service within the country is being expanded. The United States-built Kāthmāndu–Hitaura ropeway, although operational, is not used extensively for transporting food and supplies into the capital.

ADMINISTRATION, SOCIAL CONDITIONS, AND CULTURAL LIFE

Government and administration. Nepal is ruled by the King through a council of ministers and a four-tier system of panchayats (local bodies). The Basic units of government are the panchayats at the village level (more than 3,500 in all), where representatives are elected by a show of hands. These are followed by 75 district and 14 zonal panchayats, with a National Panchayat that resembles a parliament at the apex. The panchayats are supported by nonpolitical interest groups of workers and farmers. Political parties have been banned in Nepal since 1960 when King Mahendra suspended parliamentary democracy and imprisoned political leaders, including the prime minister.

The panchayat system of government

The decision to dismiss parliamentary democracy and assume complete powers, abolishing political parties, has been unpopular in Nepal. In 1963, the restoration of all fundamental rights (except the right to form political parties) brought a return to some political stability. The palace secretariat, comprising the immediate aides and assistants to the king, is of central importance in the prevailing system of government, in which royal initiative is the activating force. King Mahendra made several changes in the judicial system. In 1961, a Supreme Court act designed to bring the judicial appointment process under firm royal control was promulgated. Regional and district courts have been reorganized and there is now a district court in each of 75 development districts and zonal courts in each of the 14 zones.

Armed forces and police. Nepal's armed forces consist of the Royal Nepalese Army, predominantly an infantry force of about 14,000 men. Except for a few simple weapons, all military supplies are imported. Nepal is famous for the fighting qualities of its Gurkha soldiers; nearly 10,000 of these serve in British Gurkha units, and 50,000 in Indian Gurkha units. The British maintain a recruiting centre at Dharān Bāzār. Gurkha veterans represent a valuable human resource of Nepal.

For police purposes the country is divided into three zones: eastern, central, and western, with headquarters at Birātnagar, Kāthmāndu, and Nepālganj, respectively.

Each zonal headquarters, under a deputy inspector general of police, is responsible for several subsections composed of four to five police districts operating under a superintendent of police. A district superintendent is in charge of police stations in his area. Each station normally is supervised by a head constable who performs basic police functions.

Health and education. The administration of public-health services is the function of the Department of Health. It is responsible for the support and administration of more than 50 hospitals and 93 health clinic centres. The government has taken steps to improve existing health centres and to establish new ones. The welfare program is aimed at stimulating local initiative for a program of village development and is administered through the panchayat system. A small-scale relief operation for Tibetan refugees, who fled to Nepal after the Chinese invasion of Tibet in 1950, is supported by international agencies, including the Red Cross.

Between 1951 and the early 1970s tremendous gains took place in educational administration, supervision, and planning. The literacy rate rose from about 4 percent in the early 1950s to almost 12 percent in the early 1970s. The Director of Public Instruction is responsible for administration and supervision of all elementary and secondary education. Higher education has developed relatively recently. The first college was established in 1918, and Tribhuvan University in Kāthmāndu, with faculties of arts, sciences, commerce, and education, was charactered in 1959. The University Senate has sole legal responsibility for higher education and authority to grant academic recognition to colleges but is largely dependent upon the Education Ministry for funds.

Higher education

The cultural milieu. The relaxation of strict censorship, following the overthrow of Rana rule in 1951, encouraged a revival of artistic and intellectual expression during the 1950s and 1960s. In literature and poetry, Nepali works emphasize the cultural renaissance and national patriotism. In 1963, the government took steps to foster literary and intellectual development by subsidizing a number of young writers and by supporting and coordinating literary activities. King Mahendra, a poet whose Nepali lyrics have been published in English translation under the name of M.B.B. Shah (for Mahendra Bir Bikram Shah), did much to promote the revival of arts and literature.

The state of the arts. The cultural heritage of Nepal, particularly contributions made by Newars of Kāthmāndu Valley to sculpture, painting, and architecture, is a source of great pride. Hindu and Buddhist religious values have provided the basic source of inspiration to Newari artisans in their artistic expression. The themes of most artistic work have been primarily religious: the lives of the gods, saints, and heroes; the relationship of man to society and to the universe are expounded in sculpture, architecture, and drama. In the Kāthmāndu Valley, some 2,500 temples and shrines display the skill and highly developed aesthetic sense of Newar artisans.

Music and dance are favourite pastimes among the Nepalese. Religious ceremonies require the use of drums and wind instruments that have been preserved from ancient times. Devotional songs are an important feature on most religious and family occasions; these songs have elements of both classical and folk music and have been used by some contemporary musical revivalists in their attempt to bridge the gap between the two. The increasing urban demand for folk music is a manifestation of the new cultural nationalism; in response, Radio Nepal schedules programs featuring folk-music concerts.

The press. Newspapers and periodicals are published in Nepali and in English. The quality of the press is inferior and investigative reporting is not common. Newspapers are frequently sensational in tone and are poorly staffed and financed. *Gorkhapatra,* published by the government, occupies a commanding position in the Nepalese press. After 1960, King Mahendra required newspapers to obtain official clearance for all reports of political activity. Nepalese newspaper readers rely on the foreign press, particularly the Indian newspapers, which are daily flown into Kāthmāndu, for more sophisticated coverage of world and national news.

PROSPECTS FOR THE FUTURE

Upon assuming all power, King Mahendra promised a Nepalese form of democracy that would make the country bloom. By the early 1970s, King Mahendra's panchayat democracy had lasted over ten years, but it had still produced no substantial sign of democracy and little economic blooming. Decision making was centralized in the palace, where the palace secretariat acts as a counter to the entrenched bureaucrats of the central secretariat. The King maintained direct control over the army. The death of King Mahendra in 1972 strengthened pressures for social and political change, and some change appeared inevitable under the rule of King Birendra.

Situated between India's Gangetic Basin and the Tibetan Plateau, Nepal forms a stepping stone to the Indian subcontinent. As India becomes increasingly concerned with the security of South Asia, it has a big stake in assuring that Nepal remains politically stable and at least nonaligned. At present, India and China seem reasonably satisfied with the regime in Nepal or at least not so dissatisfied that they would be prepared to interfere. Should either nation feel that events in Nepal imperilled its interests, however, it would probably take action, for neither can afford to have the other in full control of this Himalayan state.

Historic, religious, cultural, social, and economic ties all bind Nepal more closely to India than to China. The only Hindu kingdom in the world, it naturally feels affinity with India. Yet precisely the strength of these ties stirs resentment, and since 1960 Nepal has attempted to balance India's influences by seeking a close relationship with China. Balance has been sought through nonalignment in the disputes between India and China that characterized the 1960s and early 1970s and through a process of political and economic diversification that was intended to modify Nepal's close economic and historic relationship to India. Neither policy has been a total success. Nepal's capacity to act independently remains limited, although the range of freedom allowed to Nepal has expanded.

Ties with India and China

Change is, nevertheless, slowly taking place, as the efforts made during the 1950s and 1960s begin to generate a process of modernization. Nepal's strategic location, as already mentioned, has prompted aid from many countries. India has a substantial aid program that supplies about US $13,000,000 annually; China's aid program, on the other hand, has been modest. The United States, for its part, has provided heavy financial support to encourage both economic development and the gradual growth of democratic institutions. Much of the foreign assistance, however, has not left an imprint on the national level; there is little evidence that the total of more than US $200,000,000 of aid from the United States, extended between the 1950s and 1970, has produced any significant improvement in the standard of living of any but a small minority of Nepalese people. **(P.P.K.)**

BIBLIOGRAPHY. PRADYUMNA P. KARAN, *Nepal: A Cultural and Physical Geography* (1960), provides a summary of basic cultural and physical patterns in Nepal. LEO E. ROSE and MARGARET W. FISHER, *The Politics of Nepal: Persistence and Change in an Asian Monarchy* (1970), offers a systematic study of the politics of Nepal, and discusses the country's efforts to maintain its national identity in view of its geopolitical situation between India and China. EUGENE BRAMER MIHALY, *Foreign Aid and Politics in Nepal: A Case Study* (1965), is a detailed study of aid programs and their effects in Nepal. BADRI PRASAD SHRESHTHA, *The Economy of Nepal* (1967), is a pioneer publication on economic development by a Nepali economist. HORACE B. and MARY J. REED, *Nepal in Transition* (1968), describes the setting within which Nepal's educational system operates, and relates to overall national development. DOR BAHADUR BISTA, *The People of Nepal* (1967), provides a good account of the various ethnic groups of Nepal. ROBERT M. WORTH and NARAYAN K. SHAH, *Nepal Health Survey 1965–1966* (1969), is an account of the first public health survey made in Nepal.

(P.P.K./R.R.Pr./M.Zu.)

Nepal, History of

Nepal's name is derived from the Valley of Nepal, or Kāthmāndu. The country's rich prehistory consists mainly of legendary traditions of the Newar community, which inhabits the valley area.

Prehistory and early history. There are usually both Buddhist and Brahmanic Hindu versions of these various legends. Both are accepted indiscriminately in the festivals associated with these events—a tribute to the remarkable synthesis that has been achieved in Nepal between two related but divergent value systems.

References to Nepal Valley and Nepal's lower hill areas are found in the ancient Hindu classics, suggesting that the central Himalayan hills were closely related culturally and politically to the Gangetic Plains at least 2,500 years ago. Lumbinī, Gautama Buddha's birthplace in Nepal, and Nepal Valley also figure prominently in Buddhist accounts. There is substantial archaeological evidence of an early Buddhist influence in Nepal, including a famous column inscribed by Aśoka (emperor of India, 3rd century BC) at Lumbinī and several shrines in the valley.

A coherent dynastic history for Nepal Valley becomes possible, although with large gaps, with the rise of the Licchavi dynasty in the 4th or 5th centuries AD. Although the Kirati dynasty, which had ruled the valley for the preceding 800 years, had claimed Kṣatriya (ruler and warrior) caste status, the Licchavis were probably the first ruling family in that area of Indian (*i.e.*, plains) origin. This set a precedent for what became the normal pattern thereafter—Hindu kings claiming high-caste Indian origin, ruling over a population that was neither Indo-Aryan nor Hindu in ethnic origin.

The Licchavi dynastic chronicles, supplemented by numerous stone inscriptions, are particularly full from AD 500 to 700; a powerful, unified kingdom emerged in Tibet during this period, and the Himalayan passes to the north of the valley were opened. Extensive cultural, trade, and political relations developed across the Himalayas, transforming the valley from a relatively remote backwater into the major intellectual and commercial centre between South and Central Asia. Nepal's contacts with China began in the mid-7th century with the exchange of several missions. But intermittent warfare between Tibet and China terminated this relationship; and while there were briefly renewed contacts in subsequent centuries, these were re-established on a continuing basis only in the late 18th century.

Medieval period. The medieval period in Nepali history is usually considered coterminous with the rule of the Malla dynasty (10th–18th centuries). Although most of the Licchavi kings had been devout Hindus, they had not imposed Brahmanic social codes or values on their non-Hindu subjects; the Mallas perceived their responsibilities differently, however, and the great Malla ruler Jaya Sthiti, who reigned early in the 15th century, introduced the first legal and social code strongly influenced by contemporary Hindu principles.

The Malla dynasty

Jaya Sthiti's successor, Yakṣa Malla (reigned *c.* 1429–*c.* 1482), divided his kingdom among his three sons, thus creating the independent principalities of Kāthmāndu, Patan, and Bhatgaon in the valley. Each of these states controlled territory in the surrounding hill areas, with particular importance attached to the trade routes northward to Tibet and southward to India—these were vital to the valley's economy. There were also numerous small principalities in the western and eastern hill areas, whose independence was sustained through a delicate balance of power based upon traditional interrelationships and, in some cases, common ancestral origins (or claims thereto) among the ruling families. By the 16th century, virtually all of these principalities were ruled by dynasties claiming high-caste Indian origin whose members had fled to the hills in the wake of Muslim invasions of northern India.

In the early 18th century, one of these principalities—Gorkha, ruled by the Shah family—began to assert a predominant role in the hills and even to pose a challenge to Nepal Valley. The Mallas, weakened by familial dissension and widespread social and economic discontent, were no match for the great Gorkha ruler Prithvi Narayan Shah. The Gorkha ruler conquered the valley in 1769 and moved his capital to Kāthmāndu shortly thereafter, providing the foundation for the modern state of Nepal.

Modern period. The Shah rulers faced tremendous and persistent problems in trying to centralize an area long characterized by extreme diversity and ethnic and regional parochialism. They established a centralized political system by absorbing dominant regional and local elites into the central administration at Kāthmāndu. This action neutralized potentially disintegrative political forces and involved them in national politics; but it also severely limited the centre's authority in outlying areas because local administration was based upon a compromise division of responsibilities between these local elites and the central administration.

The Shah dynasty

Throughout most of the modern period, Nepali politics has been characterized by a confrontation between the royal family and several noble families. The position of the Shah dynasty was weakened by the fact that from 1775 to 1832 three minor kings sat on the throne in succession. The regents and the nobility competed for political power, using the young rulers as puppets; both factions wanted a monopoly of political offices and power for their families, with their rivals exterminated, exiled to India, or placed in a subordinate status. This was achieved by the Thapa family (1806–37) and, on an even more extensive basis, by the Rana family (1846–1951). In these periods the Shah ruler was relegated to an honorary but powerless position, while effective authority was concentrated in the hands of the leading members of the dominant family. Although intrafamilial arrangements on such questions as the succession and the distribution of responsibilities and spoils were achieved, no viable national political institutions were created. The excluded noble families thus had only two alternatives—to accept inferior posts in the administration and army or to conspire for the overthrow of the dominant family. Until 1950, and to some extent thereafter, Nepali politics was basically conspiratorial in character, with familial loyalty taking precedence over loyalty to the crown or nation.

External relations. Prithvi Narayan Shah (reigned 1742–75) and his successors established a unified state in the central Himalayas and launched an ambitious and remarkably vigorous program of expansion, seeking to bring the entire hill area, from Bhutan to Kashmir, under their authority. They made considerable progress, but successive setbacks in wars with China and Tibet (1788–92), with the Sikh kingdom in the Punjāb (1809), with British India (1814–16), and again with Tibet (1854–56), frustrated Nepal and set the present boundaries of the kingdom.

The British conquest of India in the 19th century posed a serious threat to Nepal—which expected to be another victim—and left Kāthmāndu with no real alternative but to seek an accommodation with the British to preserve Nepal's independence. This was accomplished by the Rana family regime after 1860 on terms that were mutually acceptable, if occasionally irritating, to both. Under this de facto alliance, Kāthmāndu permitted the recruitment of Nepalis for the highly valued "Gurkha" units in the British Indian Army and also accepted British "guidance" on foreign policy; in exchange, the British guaranteed the Rana regime against both foreign and domestic enemies and allowed it virtual autonomy in domestic affairs.

The Rana family

Contemporary politics. The British withdrawal from India in 1947 deprived the Ranas of a vital external source of support and exposed the regime to new dangers. The Ranas quickly evolved a new foreign policy, under which Nepal's traditional isolation policy—vigorously enforced until then—was abandoned and relations were established for the first time with several other governments as a counterbalance to the new democratic Indian regime. Meanwhile anti-Rana forces, composed mainly of Nepali residents in India who had served their political apprenticeship in the Indian nationalist movement,

formed an alliance with the Nepali royal family, led by King Tribhuvan, and launched a revolution in November 1950. With strong diplomatic support from New Delhi, the rebels accepted a settlement with the Ranas, under which the sovereignty of the crown was restored and the revolutionary forces, led by the Nepali Congress, gained an ascendant position in the administration.

The introduction of a democratic political system in Nepal, a country accustomed to autocracy and with no deep democratic tradition or experience, proved a formidable task. A constitution was finally approved in 1959, under which general elections for a parliament were held. The Nepali Congress won an overwhelming victory and was entrusted with the formation of Nepal's first popular government. But persistent controversy between the Cabinet and King Mahendra (whose reign began in 1955) led to the dismissal of the Nepali Congress government in December 1960. The 1959 constitution was abolished in 1962, and a new constitution was promulgated, making the crown the real source of authority. King Mahendra obtained both Indian and Chinese acceptance of his regime, and the internal opposition was weak, disorganized, and discouraged. Mahendra died in January 1972 and was succeeded by his son, Birenda.

The reign of King Mahendra

In the early 1970s there was substantial dissatisfaction with some aspects of the country's political and economic system, and Katmandu was the scene of anti-government rioting in August 1972.

BIBLIOGRAPHY. DOR BAHADUR BISTA, *The People of Nepal* (1967), the best descriptive account of the various tribes and ethnic communities in Nepal; BHOLA CHATTERJI, *A Study of Recent Nepalese Politics* (1967), a personal account of the 1950–51 revolution and of the role played by the Indian Socialist Party; ANIRUDHA GUPTA, *Politics in Nepal: A Study of Post-Rana Political Developments and Party Politics* (1964), a detailed account of Nepali political parties in the 1947–60 period; TONI HAGEN, *Nepal: The Kingdom in the Himalayas* (Eng. trans. 1961), detailed topographical and geological data, based on seven years' work in the hill areas; BHUWAN LAL JOSHI and L. E. ROSE, *Democratic Innovations in Nepal: A Case Study of Political Acculturation* (1966), a thorough study of political developments in Nepal since 1951 with background historical material; SATISH KUMAR, *Rana Polity in Nepal: Origin and Growth* (1967), a study of the Rana political system covering the period from 1847 to 1951; SYLVAIN LEVI, *Le Népal: Étude historique d'un Royaume Hindou* (1905–08), the best of the older Western-language histories of Nepal, although now outdated; LUCIANO PETECH, *Medieval History of Nepal, c. 750–1480* (1958), a historical account of this period based on Chinese, Tibetan, and Nepali sources; L.E. ROSE, *Nepal: Strategy for Survival* (1971), the first thorough study of Nepal's relations with India and China in the modern period (after 1770), and with M.W. FISHER, *The Politics of Nepal: Persistence and Change in an Asian Monarchy* (1970), a general analysis of Nepali politics intended as a textbook for introductory politics courses; GIUSEPPI TUCCI, *Nepal: The Discovery of the Malla* (Eng. trans. 1962), a brief historical account of one of the major kingdoms in the Western hill area of Nepal in the early medieval period.

(L.E.R.)

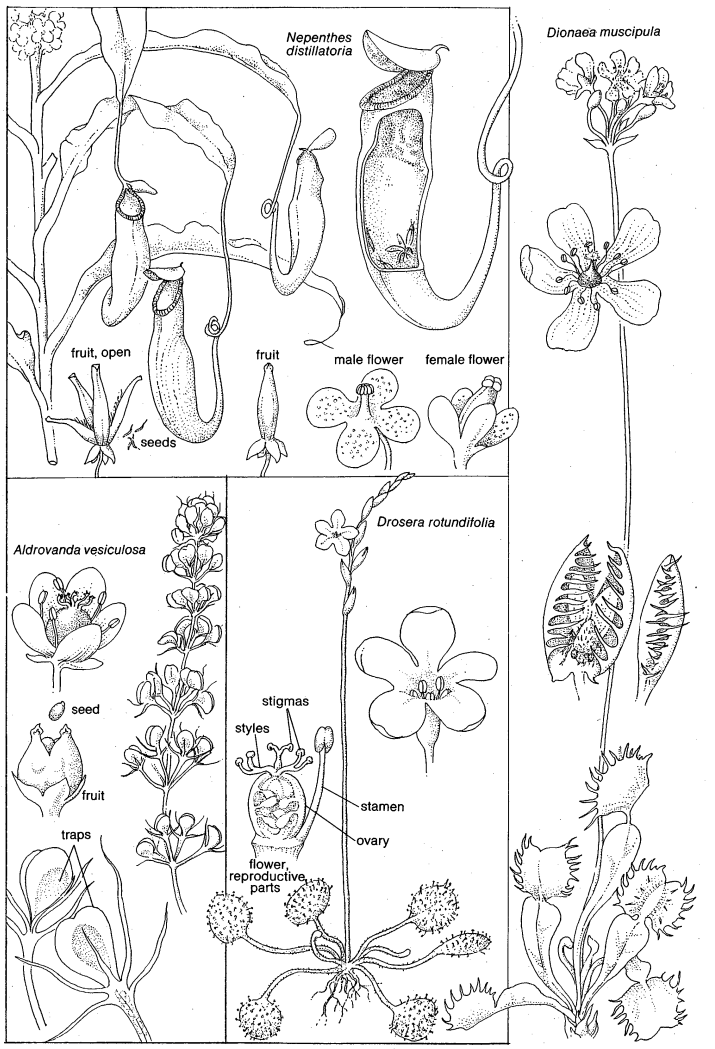
Nepenthales

The flowering plant order Nepenthales consists of two families, the Old World pitcher plants or monkey cups (family Nepenthaceae) and the sundews and Venus's-flytrap (family Droseraceae). All of the approximately 180 species have leaves that are ingeniously adapted as pitfalls and traps to capture insects and other tiny animals, digest them, and absorb the nutrients.

The family Nepenthaceae contains the single but large pitcher-plant genus *Nepenthes*. Its approximately 80 species are found in wet tropical forests from Ceylon and Madagascar to New Guinea and northern Australia. All of the species capture insects by means of pitcher-shaped leaves that function as pitfalls. The prey is attracted to the large showy leaves by nectar, and it slips from a treacherous lip into the pitcher-like hollow leaf, which in reality is a cistern of digestive liquid from which there is no escape.

The family Droseraceae contains species with leaves modified as real traps, active mechanisms that clamp shut or coil about victims lured by nectar secretions and delicate odours. The principal genus is *Drosera* (the sundew), with about 100 species, which together have a distribution in virtually all parts of the world. Three other genera each contain a single species. *Dionaea muscipula*, the Venus's-flytrap, is a narrowly distributed Atlantic North American endemic (a native species restricted to one region). *Drosophyllum lusitanicum* is an unusual sundew narrowly endemic to the Atlantic European area. *Aldrovanda vesiculosa* is a wide-ranging aquatic plant in the Old World.

Drawing by M. Pahl



Representative plants of the order Nepenthales. (Upper left) Family Nepenthaceae. (All others) Family Droseraceae (see text).

GENERAL FEATURES

Size range and diversity of structure. The species of *Nepenthes* are quite uniform in general appearance and, by comparison with other known plants, stand clearly by themselves. The sizes and shapes of the pitchers are characteristic for each species, although there may be as many as three distinct pitcher forms in a single species. The pitchers are colourful and are often mistaken for flowers. The plant typically is a coarse vine that grows over vegetation, sometimes climbing to the crowns of 50- to 60-foot (15- to 20-metre) trees.

The Venus's-flytrap has jawlike leaf blades, which dominate the vegetative aspect of the plant and are capable of clamping shut with amazing speed when properly stimulated.

The most common sundews are low, herbaceous (nonwoody) plants consisting of a slender rhizome, or rootlike stem, crowned by a rosette of sticky leaves that snare insects. The small, often tiny leaves are densely covered

The sundew plants

with slender, sensitive tentacles capped by globular red glands that secrete a highly viscid, clear mucilage. The droplets of mucilage glisten in sunlight, giving the plant a dewy appearance. This explains the generic name *Drosera*, derived from the Greek word for "dewy," and the common name, sundew. Because the clear mucilage magnifies the red gland, the whole plant may appear bright red. Species of *Drosera* vary in height from a few inches (*D. rotundifolia*) to three or more feet (*D. gigantea*).

Drosophyllum lusitanicum is a shrubby plant up to five feet (1.5 metres) tall with linear leaves that remain attached to the plant after they die, forming skirtlike layers around the base of the stem. The leaves are covered with numerous stalked glands bearing large droplets of clear mucilage that give the plant a dewy appearance responsible for the common name dewy pine. These glands are not capable of movement as in *Drosera*, but vast numbers of gnats and flies adhere to them, and, because of this, the plant is sometimes called flycatcher.

Although widespread and unusual, the inconspicuous *Aldrovanda vesiculosa* is not well-known except among the few botanists who specialize in carnivorous plants. *Aldrovanda* is a small freshwater plant that floats just below the surface of quiet waters, and its tiny steel-trap leaves function as in the Venus's-flytrap (*Dionaea*) except that the lobes are not hinged. In the "set" position, these leaf traps suggest a widely opened clam or mussel. The whole plant, however, is only four to six inches (ten to 15 centimetres) long and barely an inch across, including the spreading leaves.

Distribution and abundance. The geographic range of the pitcher plants radiates from Borneo, particularly around Mt. Kinabalu in Sabah (East Malaysia), where the most striking species occur. Some 20 species have been reported from the island. About 12 species occur on the adjacent island of Sumatra; eight to ten species are found in the Philippines and in northern Australia. The extremely light seeds seem to be well adapted for wind dispersal, which may explain the distribution in radial manner from Borneo as the focal point. Many of the species, such as those in Madagascar and Ceylon, are relatively local endemics. *Nepenthes phyllamphora*, on the other hand, ranges from southeastern China to New Guinea and has close relatives in northern Australia. Several other species (*N. ampullaria*, *N. gracilis*, *N. rafflesiana*) are less widespread but relatively wide-ranging.

The 100 or more species of *Drosera* are distributed primarily in the Southern Hemisphere. Half of them are found in Australia and New Zealand, the centre of greatest diversity of the genus. Southern Africa is second as a centre of speciation. Only a few species occur in the north Temperate Zone, among them two circumboreal ones (*D. anglica* and *D. rotundifolia*) and one nearly circumboreal species (*D. intermedia*). Only these three species are found in Europe and Soviet Asia, although in temperate North America there are four additional well-marked species (*D. brevifolia*, *D. capillaris*, and *D. filiformis* of the southeastern states and *D. linearis* of the northern states and Canada).

Distribution of the Venus's-flytrap

Dionaea, the Venus's-flytrap, is endemic to the Atlantic Coastal Plain from Beaufort County, North Carolina, to the south bank of the South Santee River, Charleston County, South Carolina. Within this highly restricted range, the flytrap tends to occur in widely scattered colonies but may be abundant locally. *Drosophyllum* is native only to northern Morocco and southern Spain and Portugal. By contrast, *Aldrovanda* ranges from Atlantic Europe to India and Japan and south to Mozambique (East Africa) and Australia.

Economic importance. The pitcher plants, sundews, and especially the Venus's-flytrap have always been valued as objects of curiosity, but they have never accounted for more than an insignificant element of the horticultural trade. Otherwise, the order Nepenthales is of little or no economic importance except on a local basis.

NATURAL HISTORY

Life cycle. Except for some annual species of *Drosera*, the plants of this order are perennial (*i.e.*, they grow for several years). In *Nepenthes* most species are caulescent (stemmed) and shrubby; in the Droseraceae the acaulescent (stemless) herbaceous habit predominates. Apart from a new crop of fine fibrous roots produced each year from the rhizome, few if any true roots are present. In some species of *Drosera*, the seedling has a short-lived taproot that functions while the earliest leaves are forming but never elongates. *Aldrovanda* is rootless and floating, and some species of *Nepenthes* are epiphytes, plants that grow upon the branches of other plants.

The stem is rudimentary or relatively short in most species of the Droseraceae, but in *Nepenthes* it may elongate considerably into a vine. The young *Nepenthes* stem is densely clothed with leaves, but the internodes (stem sections between leaf attachment points) lengthen with increasing age. When lying on the ground the stems of some species (*e.g.*, *N. ampullaria*) may root abundantly at the nodes (points of leaf attachment) and produce dense clusters of leaves with reduced blades and richly coloured pitchers. Two or three types of pitchers may be formed on the same plant at different heights on the stem and seem to function like flowers in attracting insects. Some species of *Drosera* have climbing stems as in *Nepenthes*, and some form tubers (enlarged underground stem structures) by which they overwinter to produce new shoots the following season. The tuber-bearing species have no roots. In the common sundew (*D. rotundifolia*) the stem dies off behind as the plant grows, and in winter the rosette is reduced to a tight bud that may have no measurable stem or root; successively, a series of dead rosettes forms. *Aldrovanda* produces new shoots each season from a persistent terminal bud. The temperate perennial species of the Droseraceae all go through a period of winter dormancy.

The flowering period of *Nepenthes* in the wild, though extended and inconstant for many species, is mainly from March to September. Under cultivation in the Northern Hemisphere *Nepenthes* blooms from August to October. The inflorescences (flower clusters) attract clouds of insects to sip on the nectar secreted abundantly by the flowers, which are cross-pollinated by insects. The male flowers give off a fetid odour that probably attracts insects. After pollination the fruits, containing 100 to 500 seeds per capsule, mature in six to eight weeks.

The common sundew, which is representative of north temperate species of *Drosera*, sends up a slender, arching flower stalk in early summer or later. The flowers open one at a time in sunlight, for one morning each. They are bisexual and pollinate themselves after closing if cross-pollination has not taken place while they are open. The single flower stalk of the Venus's-flytrap is produced in late May, June, or early July. Each flower remains open about three days, and the ripe pollen is shed before the stigmas (pollen receptive surfaces of the female flower parts) are receptive. The seeds are shed as early as mid-July.

Modes of reproduction

The seeds of *Nepenthes* germinate readily within five to six weeks after ripening in loose, moist habitats, such as humus or sphagnum moss. *Drosera* reproduces easily by seed, exhibits remarkable powers of vegetative regeneration, and in some species also reproduces prolifically by asexual means. Several species, including *D. pygmaea*, reproduce by brood bodies, or gemmae (small clusters of vegetative cells that break off and grow into new plants); in Australia, reproduction by tubers is common. The underground tubers are important as stored food and water reserves for the plant to use during the dormant season when the rosette leaves are not present. Some species (*e.g.*, *D. rotundifolia*) have strong axillary buds (buds located in the angles between leafstalks and stems) that form secondary rosettes below the main one; as the stem decays, these are separated and propagate the plant. The formation of plantlets by adventitious budding, especially from the leaves but also from the roots and inflorescence, rivals seed reproduction in this genus.

In *Dionaea* seeds germinate immediately in the summer, and they are viable in nature only for a relatively short period of time. High humidity and loose organic matter on a sandy surface are required for germination. As in

Drosera, regeneration from vegetative parts or the immature inflorescence is possible.

In *Nepenthes* hybridization, or cross-fertilization between different species, is common in cultivation and also takes place in the wild. All of the cultivated species have been hybridized, and the hybrids are highly fertile. As a result, many interesting artificial hybrids now exist in cultivation.

Carnivorous adaptations. *Trapping mechanisms.* The pitchers of *Nepenthes* are conspicuously supplied with glands that attract prey. Insects, particularly ants, are lured to the mouth of the pitcher by numerous nectar glands scattered over the stems and leaves, including the tendrils and the inner surface of the lid. An insect that successfully negotiates the peristome (corrugated lip of the pitcher) loses footing on the crumbly, waxy surface lining the upper half of the interior of the pitcher and drops into the pitcher. The surface of the digestive region in the lower part of the pitcher is crowded with more than 6,000 digestive glands per square centimetre.

All species of *Drosera* capture their prey by means of flexible tentacles on the upper surface of the leaf. As many as 400 tentacles have been counted on a leaf of the common sundew, which is about one centimetre in diameter. These tentacles really are stalked glands with globular heads that secrete a clear, sticky mucilage. When a midge, gnat, or mosquito lands on the leaf, the alluring mucilage becomes the first snare, smearing and smothering the insect. The struggling insect stimulates further secretion. Within minutes the inner tentacles bend over and gradually pin down the insect. Soon the outer tentacles, which become successively longer toward the edge of the leaf, also begin to curl inward. In a matter of minutes or hours, depending on the species and amount of stimulation, the insect is hopelessly engulfed by a web of sticky stalks.

Tentacle movement in the sundew

Bending and unbending of the tentacles are caused by differential rates of rapid growth. When the outer cells of the tentacles grow faster than the inner ones, incurving results. The tentacles straighten again when the inner cells grow faster than the outer ones. As a growth response, the action of the tentacles has limits; the bending and unbending can be repeated only about three times by any one tentacle. The action of the tentacles is slow compared to the trapping mechanism of the Venus's-flytrap, and the sundew can catch only small or very weak prey; flies and even ants are able to escape. After a few days to a week, digestion is completed, the leaf reopens, resetting the trap, and the remains of the previous victim fall off or are washed or blown away.

The trapping mechanism of *Drosophyllum* involves glandular tentacles like those of *Drosera*, but the tentacles do not move. The insect becomes caught among the copious sticky glands in much the same way as it might be stuck by flypaper.

In the Venus's-flytrap, *Dionaea*, the blade of the leaf consists of two lobes hinged to each other like the jaws of a steel trap and provided with spinelike teeth along the margins. On the face of the lobes are numerous nectar and digestive glands. The digestive glands, which secrete very actively when prey is captured, contain a red sap giving the entire surface a bright red and flowerlike appearance. In the middle of each lobe are three pressure-sensitive hairs that trigger trap closure when properly stimulated. Two or more stimuli are required to spring the trap, the number depending on the time intervals between the stimuli. Two stimuli are sufficient if the second comes not less than 1.5 seconds nor more than 20 seconds after the first. The same trigger hair may be struck twice or two hairs may be struck once each. If the stimuli are spaced over a period of five to six hours, as many as 18 to 20 separate stimuli may be required. When stimulated, the trap springs shut with amazing speed, usually in about a half second at normal daytime temperatures. The minimum closing time recorded is 80 milliseconds, less than one-tenth of a second. If prey is captured, the clamping movement continues slowly until the surfaces of the lobes press tightly against the body and the marginal teeth enmesh. The more times the hairs are stimulated, the tighter the leaf closes; thus, a struggling insect inside the trap causes the trap to close more tightly. A fly caught one afternoon may still be struggling the next morning; but, in the meantime, the trap will have contracted so tightly that the fly's form shows through the trap lobes. If no insect is caught, or if the insect is tiny and is able to escape between the teeth, the leaf reopens in five to ten hours. Otherwise the trap remains closed for seven to ten days or even longer. After having captured three or four victims, the trap has matured completely and dies.

In the aquatic *Aldrovanda* the trapping mechanism is similar to that of the Venus's-flytrap, although the trap is tiny by comparison. The number of stimuli required to trigger the trap depends on the age of the leaf. Small insects, larvae, copepods, and other tiny animals that spring the trap are locked into an ovate hollow formed when the two lobes snap together, a movement requiring only 1/50 of a second.

Digestive mechanisms. The pitchers of *Nepenthes* secrete a quantity of enzymatic fluid before opening, and, after opening, secretion is stimulated greatly by the presence of animal victims. The fluid contains a weak acid in addition to proteinase, the main enzyme, which breaks down protein. Bacteria live in the mature pitchers and help to break down the animal compounds. Under high temperatures digestion is accomplished in two to eight hours. The nitrogenous end products are then absorbed through the walls of the pitcher.

Bacteria and enzymes in digestive fluid

When an insect becomes stuck on the leaf of a sundew and struggles to get free, the tentacles are stimulated to secrete a digestive enzyme that breaks down the proteins in the insect body. In *Drosophyllum*, enzyme secretion is continuous, and the digestive cycle does not have a well-marked beginning and end.

Most of the upper surface of the leaf in *Dionaea* (*i.e.*, the inside of the trap) is covered with up to 50 sessile (stalkless) glands per square millimetre; the glands secrete a digestive fluid containing proteinase and formic acid after the trap has closed on an insect or other digestible victim. Digestion begins within 36 hours after trap closure and continues for about ten days.

Aldrovanda has glands lining the lobes of the trap that release digestive juices after the trap springs shut. The end products are absorbed by the plant, and the trap may reopen several times after feeding if the catches have been meagre.

Animals that exploit leaf traps. The pitchered leaf of *Nepenthes* constitutes a special habitat with both wet and dry ecological niches, and certain insects and other small animals have adapted to these niches. They may exploit the pitcher for food, shelter, or as a place to pass part of their life cycle. In some instances, the animal takes advantage of the trap by robbing the plant of its own victims; in other instances, the plant tissue itself is parasitized.

Adaptation to the pitcher-plant environment has become so specific in some cases that the animals have lost their ability to complete a life cycle elsewhere. Several species of spiders occur only in the pitchers of *Nepenthes*, in which they live in the mouth above the water level and prey on insects attracted to the pitchers. More than two dozen species of the insect order Diptera (flies) are known to occur exclusively in the aquatic environment of the pitchers; two-thirds of these species are mosquitos.

Many animals have been recorded as accidental or occasional inhabitants of *Nepenthes* pitchers, especially in the creeping species (notably *N. ampullaria*), in which the pitchers stand half buried in the forest floor. These include protozoans, slime molds, desmids (a group of green algae), diatoms, rotatoria, oligochaetes (worms), crustaceans, larvae of various flies, and even an occasional tadpole. In Sabah the small rodent *Tarsius spectrum* visits the pitchers of various species, perches on the margin, bends its head and neck into the opening, and scoops out and devours the insects caught there.

A few animals are able to inhabit the sundews (*Drosera*). A moth larva of the family Noctuidae feeds on the glandless portions of the plant, and certain Australian bugs of the order Heteroptera (family Capsidae) move

about the sticky leaves without apparent harm or difficulty, sucking the juices of ensnared victims. In Florida, the larva of the moth *Trichoptilus parvulus* (plume moth) has been discovered on the leaves of *Drosera capillaris*. It eats the stalked glands with their sticky secretion, the leaf blades, and even trapped dead insects.

Ecology. *Nepenthes* is a vinelike jungle plant (liana) that creeps over the ground or climbs on the trees of tropical swamp or rain forests. The genus ranges in altitude from sea level to 10,000 feet (3,000 metres), and only rarely does it grow in other than moist or very moist habitats. Around Mt. Kinabalu in Sabah, where the distribution of the group is centred, the species show a distinct zonation in their altitudinal distribution. Some species (*e.g.*, *N. gracilis*) grow at low elevations on the hot humid plains; others (*e.g.*, *N. ampullaria*) grow commonly at 2,000 to 2,300 feet (600 to 700 metres); still others (*e.g.*, *N. rajah*) reach the 5,000- to 10,000-foot (1,500- to 3,000-metre) zone where clouds or rain are always present, and they are exposed almost continuously to a cool, dripping wet environment. No species of pitcher plants occur between the 10,000-foot level and the 13,700-foot (4,200-metre) summit of Mt. Kinabalu. At least one species inhabits wet savannas (grassland areas with scattered trees), where it climbs on shrubs; the genus is even known to occur on dry, sandy hillsides.

Altitudinal distributions of pitcher plants

The species of the family Droseraceae typically inhabit damp heaths, bogs, swamps, wet savannas, or sandy shores, showing a marked preference for wet habitats with acid soil in which nitrogen and other nutrients are deficient or unavailable. The flycatcher (*Drosophyllum*) of the western Mediterranean region is an exception to the wet-habitat rule; this coarse, wiry sundew grows on dry, gravelly (quartz) hills, where, nevertheless, the soil is acidic.

In North America, the sundew (*Drosera*) is one of the first plants to invade newly disturbed habitats such as roadside ditches and abandoned logging trails. The bright red rosettes often cover the ground. Many of the species thrive best in habitats where sphagnum is growing. Several species of *Drosera* are often found in the same habitat. *Drosera anglica* and *D. rotundifolia* occupy swamps but always in the parts of the swamp most deficient in salts. Most *Drosera* species grow in poor soils. The soil of a pine forest, for example, contains some 27 times more nitrogen than a *Drosera* habitat. The sundew is among the earliest appearing inhabitants of raw peat surfaces, which are a nutrient-poor habitat. Too much calcium in the soil may be toxic to *Drosera* plants. The poor-soil generalization does not apply necessarily to all of the many species of *Drosera* in the Southern Hemisphere; in fact, at least one species of southern Australia (*D. whittakeri*) grows on wooded slopes among general vegetation.

Venus's-flytrap thrives in a habitat transitional between wet evergreen-shrub bog (pocosin) and dry sandy longleaf-pine savanna on the Atlantic Coastal Plain of the United States. The plants tend to grow along the edges of wet depressions and pools, often in company with pitcher plants of the genus *Sarracenia*, sundews, and butterworts (*Pinguicula*). Occasional burning helps to maintain this limited pineland in the scrubby, semi-savanna condition. Thus fires, far from being the enemy of the flytrap, actually may be required for its continued existence.

Aldrovanda grows in shallow quiet water between the stems of such marsh plants as cattails (*Typha*), reeds (*Phragmites*), and wild rice (*Zizania*).

Carnivory, the adaptation for digesting animal prey, is the most important aspect of the ecology of the species in this order; it has enabled the plants to occupy otherwise unavailable niches.

FORM AND FUNCTION

The *Nepenthes* plant consists of a creeping rhizome from which arises a coarse, leafy stem that climbs over trees and other vegetation by means of its modified leaves, which may attain a length of three to ten feet (one to three metres). The expanded, tongue-shaped or round blade of the thick, glossy leaf narrows to a long tendril that twines around other plants and is always coiled except in the short upright or creeping species of the genus. The tendril has enormous strength and, in some species (*e.g.*, *N. rafflesiana*), may support as much as 13 pounds (six kilograms) of weight without breaking. The tendril finally becomes transformed at its tip into a pitcher that always is held upright and cannot function otherwise. The pitchers vary considerably in form even on the same plant; they may be tubular, cylindrical, urn-shaped, funnel-shaped, ovoid, goblet-shaped, or globose. In most species, but noticeably in the cylindrical ones, the upper one-third of the pitcher, corresponding to the inner waxy zone, is somewhat constricted. Pitchers range from the size of a thimble to that of a pint jar or larger (up to 20 inches, or 50 centimetres), but commonly they are four to six inches (ten to 15 centimetres) long. The colour usually varies, even on the same plant, from yellow green to dark green with all degrees of red marking, depending on age, light exposure, and soil conditions; sometimes, however, the pitchers are a uniform pale green or deep red. When red occurs on the rim (lip), it forms stripes. The rim of the pitcher is fluted, and the corrugations strengthen the lip and help to keep the mouth from collapsing. This corrugated structure, known as the peristome, is a conspicuous feature of nearly all species. An immobile lid (operculum), formed by an outgrowth of the leaf, may overhang the mouth and help to keep out rain, or it may be small, narrow, and turned back. The under surface of the lid and the corrugations of the lip are supplied copiously with nectar glands except in a few species.

Specialized structure of pitcher plant leaves

The small greenish or claret-coloured flowers of *Nepenthes* are borne in clusters at the tips of stems opposite leaves; as few as 15 to 20 and as many as several hundred flowers may be produced in one inflorescence. The flowers are noteworthy chiefly because they are unisexual, and the male and female flowers, unlike those of other carnivorous plants, are found on different plants (*i.e.*, *Nepenthes* is dioecious).

Drosera species vary from a few inches to three or more feet in height. The smallest often are hidden among the mosses in a sphagnum bog. Among the North American species, which are herbaceous and stemless, there are three main leaf forms: oval or round (*D. rotundifolia*); spoon-, club-, or strap-shaped (*D. intermedia*); and threadlike or wirelike (*D. filiformis*). The leaves of the wide-ranging common sundew resemble tiny mittens or baseball catcher's gloves. The well-known sundews of the Northern Hemisphere usually produce one to 20 erect flowers on a single, leafless flower stalk that is six to eight inches (15 to 20 centimetres) long or occasionally up to a foot (30 centimetres) or more long. The flower, measuring an inch or less across, is bisexual and regular (radially symmetrical); its five petals are white to pink or even deep red, the colour depending on the species. Some species of *Drosera* from the Southern Hemisphere bear their flowers on leafy stems.

Flowers of the sundew

The Venus's-flytrap has a single flowering scape (leafless stalk) up to a foot high. The two to 15 bisexual white flowers measure about one inch across and are similar to those of *Drosera*. The leaf is divided into a blade and a prominently winged petiole (stalk), narrowing at the base of the blade. The blade consists of two identical semicircular lobes united along the middle by a thick midrib and standing at an angle of about 50° to each other; the lobes constitute the hinged trap. The outer margin of each lobe is provided with a comb of long bristles, about 13 in number, that interlock like fingers when the trap closes. The leaf is from two to six inches (five to 15 centimetres) long and about an inch or two across. The plants are bright red when growing in strong sunlight, but green if in shade.

EVOLUTION AND CLASSIFICATION

Evolution. The fossil record of the order is extremely limited and does little to illuminate the evolutionary history. Speculations on the phylogenetic relationships, which have been based on comparative studies of the modern species of the order, have resulted in divergent

views. Although there is no common view on the origin of carnivory among the flowering plants, authorities now agree that the carnivorous plants represent several more or less unrelated lines of evolution.

Phylogenists who accept the concept of the order Nepenthales as defined here, comprising the families Droseraceae and Nepenthaceae, believe that it probably evolved from the order Saxifragales. Most authorities now are convinced on the basis of floral morphology that the family Droseraceae is closely related to the Saxifragaceae, especially that part of the family sometimes segregated as the separate family Parnassiaceae.

Traditionally, the families Droseraceae and Nepenthaceae were thought, on the basis of leaf morphology, to form a natural group with the Sarraceniaceae (New World pitcher plants), because of the obvious vegetative transitions between the two pitcher-plant families (Sarraceniaceae and Nepenthaceae). Some authorities still hold to this view, choosing to place more weight on the evidence from leaf morphology than from floral morphology. At issue, primarily, is the evolutionary position of the Nepenthaceae, about which there is little agreement. The family Droseraceae clearly is more advanced than the Sarraceniaceae, and the two should not be placed in the same order (see SARRACENIALES). The members of the family Nepenthaceae are closely knit and clearly have evolved from a common ancestor. One line of phylogenetic reasoning places the Nepenthaceae in an order by itself that stands between the evolutionary positions of the Droseraceae and Sarraceniaceae but somewhat closer to the latter; another interpretation has placed Nepenthaceae in the order Aristolochiales (*q.v.*; a group that includes the pitcher-like Dutchman's pipe). Evidence from pollen and ovule morphology, however, now indicates a close affinity between the Nepenthaceae and the Droseraceae and at the same time casts doubt on the supposed affinities of the Old World pitcher plants with the Aristolochiales or Sarraceniales.

Differing opinions on the evolutionary position of pitcher plants

Classification. *Distinguishing taxonomic features.* The Nepenthales is distinguished from other orders chiefly by the characteristics of the distinctive insect-trapping leaves and, to a lesser extent, by features of the inflorescence and flower, particularly pollen features. The main characteristics for separating the two quite distinct families are the habit and size of plant; form and arrangement of leaves; presence of unisexual or bisexual flowers; presence or absence of differentiated petals; number of carpels (ovule-bearing segments of the pistil, or female structure); number, form, and fusion of styles (narrow, upper part of pistil); and number and fusion of stamens (male pollen-producing structures).

Annotated classification. The classification presented here departs from the traditional one and places the Droseraceae and the Nepenthaceae in an order by themselves, apart from the Sarraceniales. Whatever the degree of their affinity may be, these two families superficially do not show much morphological similarity. To the casual observer their chief similarity is functional—the carnivory of their leaves.

ORDER NEPENTHALES

Plants stemmed or stemless, annual or perennial, herbaceous or somewhat shrubby, with alternate leaves specialized in various ways for trapping insects and other tiny animals. Flowers unisexual or bisexual, radially symmetrical, with sepals only or sepals and petals. Stamens few to many, free or more or less fused. Pistil 1, positioned above the other flower parts (superior), with 3 to 5 fused carpels (ovule-bearing segments), the ovules few to numerous. Fruit usually a capsule. Seeds small, few to numerous. Two families, 5 genera, and about 180 species distributed worldwide; concentrations of species occur in Borneo (*Nepenthes*) and Australia (*Drosera*).

Family Droseraceae (sundew family)

Annual or usually perennial, glandular-hairy, stemless herbs or rarely subshrubs. Leaves usually in basal rosettes, both surfaces (except in *Dionaea*) generally covered with sticky, gland-tipped hairs and sensitive tentacles that enmesh small insects. Flowers bisexual, mostly with 5 sepals and 5 petals. Stamens 5 to many, free or nearly so, in whorls of 5. Ovary 1-chambered (occasionally 3- to 5-chambered), with 3 to 5 carpels; styles 3 to 5, free, often divided. Fruit a capsule. Seeds 3 to many, spindle-shaped. Four genera (*Aldrovanda, Dionaea, Drosera,* and *Drosophyllum*) and about 100 species, ranging throughout the world from temperate to tropical regions.

Family Nepenthaceae (pitcher-plant family)

Perennial, herbaceous or usually shrubby, stemmed plants, often climbing by their leaves, anchored in the soil or growing as epiphytes upon other plants, often forming vinelike lianas. Leaves with a winged or expanded portion followed by a constricted, often coiled tendril terminated by hanging but upright, cylindrical or urn-shaped, often highly coloured, insect-trapping, pitcher-shaped structures with a lid (operculum). Flowers unisexual, with usually 4 distinct, nectar-producing sepals (in 2 whorls of 2 each). Petals absent. Male flowers with 4 to 24 stamens fused into a column. Female flowers with single pistil that has a 3- or 4-carpelled and 3- or 4-chambered ovary; style 1, short to absent, with 4 stigmas. Capsule leathery. Seeds numerous, with long terminal wings, very small. One genus (*Nepenthes*) with about 80 species distributed in the tropics from northeast India to Australia and New Caledonia; concentrated in Borneo and Sumatra and reaching also to the Seychelles (1 species) and Madagascar (1 species).

Critical appraisal. The many species of the family Nepenthaceae, despite their sometimes striking variation, are, by comparison with other plants, so morphologically homogeneous that they have always been placed in the single genus *Nepenthes*; the cohesiveness of the family never has been challenged seriously. The family Droseraceae, with several strikingly different types of traps, is by contrast a heterogeneous family of four vegetatively distinct genera, which, however, show a close affinity to each other on the basis of floral morphology. Some authorities have treated the bizarre Venus's-flytrap (*Dionaea muscipula*) as a separate, monotypic (single species) family, Dionaeaceae, although this view is defensible only if vegetative morphology is stressed at the expense of flower structure.

The question of the closeness of relationship between the Old World pitcher plants (family Nepenthaceae) and the New World pitcher plants (order Sarraceniales) will remain unresolved until more is known about the evolution of the pitchered leaves in *Nepenthes*, on the one hand, and *Heliamphora* (family Sarraceniaceae), on the other hand. New worldwide monographic studies are needed to determine the number of species in the large genera *Drosera* and *Nepenthes*.

BIBLIOGRAPHY. CHARLES DARWIN, *Insectivorous Plants* (1875), the classic study of the carnivorous flowering plants, especially the family Droseraceae and particularly *Drosera rotundifolia*, common sundew; FRANCIS ERNEST LLOYD, *The Carnivorous Plants* (1942), the definitive modern treatment of all groups of carnivorous plants on a worldwide basis; CARROLL E. WOOD, JR., "The Genera of Sarraceniaceae and Droseraceae in the Southeastern United States," *J. Arnold Arbor.*, 41:152–163 (1960), an excellent review of the morphological and biological features of the family Droseraceae; FRANCES E. WYNNE, "*Drosera* in Eastern North America," *Bull. Torrey Bot. Club*, 71:166–174 (1944), latest revision of North American species, with key, descriptions, maps, and illustrations; PATRICIA R. ROBERTS and H.J. OOSTING, "Responses of Venus Fly Trap (*Dionaea Muscipula*) to Factors Involved in Its Endemism," *Ecol. Monogr.*, 28:193–218 (1958), the most comprehensive study of the ecology of this species ever made; WILLIAM R. WEST, "Carnivorous Plants," pt. 2–4, *Carolina Tips*, 28:35–36, 39–40, 29:1–2 (1965–66), brief general accounts of *Dionaea* and *Drosera* and on the culturing of carnivorous plants; STANWYN G. SHETLER and FLORENCE MONTGOMERY, "Insectivorous Plants," *Inf. Leafl. Smithson. Instn. 447* (1965), a brief, popular, well-illustrated survey of the major groups of carnivorous plants in the world; TERRY ASHLEY and JOSEPH F. GENNARO, JR., "Fly in the Sundew," *Nat. Hist.*, 80:80–85, 102 (1971), a readable account of modern experimentation with radioactive marking to study digestion; FRANK L. LAMBRECHT, "The Pitcher-Plant: Carnivorous Incubator and Tool of Natural Selection," *Jl. N.Y. Bot. Gdn.*, 21:100–105 (1971), fascinating narrative of research on animals that exploit the pitchers of *Nepenthes pervillei* on Ceylon; SOPHIA PRIOR, "Carnivorous Plants and 'The Man-Eating Tree'," *Bot. Leafl. Field Mus. Nat. Hist. 23* (1939), a brief account of popular lore on carnivorous plants.

(S.G.S.)

Neptune

Neptune, symbol ♆ in astronomy, is the eighth planet from the Sun and, until Pluto was discovered in 1930, was the outermost known planet in the solar system. Its mean distance from the Sun is almost exactly 30 astronomical units (1 a.u. = the mean distance of the Earth from the Sun). Neptune will, in fact, be the most distant planet from the Sun for about two decades after 1979, when Pluto will be near its point of closest approach to the Sun, inside Neptune's orbit.

Neptune, being only as bright as a star of the eighth magnitude, is invisible without optical aid. The image it forms in a telescope of moderate aperture shows a sensible disk of a slightly greenish hue and about 2″.3 in diameter. Physical data for Neptune are given in Table 1.

Table 1: Physical Data for Neptune

Mean distance from Sun	$4{,}486.10^6$ km (30.1 a.u.)
Eccentricity of orbit	0.011857
Inclination of orbit to ecliptic	1° 46′ 34″
Sidereal period of revolution	$60190^d.0$ (164.8 years)
Rotation period	$15^h.8$
Albedo	0.84
Mean synodic period	$367^d.49$
Mean orbital velocity	5.44 km/sec
Mean daily motion	0° .005981
Inclination of equator to ecliptic	28° 48′
Mass (Earth = 1)	17.46
Diameter (including atmosphere)	50950 km
Density	1.54 gm/cc
Apparent magnitude	+7.8
Satellites	2

Discovery. The detection of Neptune through its action upon Uranus is one of the fascinating stories in the history of astronomy, and a striking example of the precision of the theory of planetary motions. Before its actual discovery by the English astronomer Sir William Herschel in 1781, Uranus had been observed as a fixed star on at least 17 other occasions, first by the English astronomer John Flamsteed in 1690. The French astronomer Alexis Bouvard, in 1820, constructed tables of the motions of Jupiter, Saturn, and Uranus; the observed positions of Jupiter and Saturn were represented satisfactorily but he found it impossible to do this for Uranus. During the next two decades there was continued divergence between the theory of perturbations, describing how the known planets should alter slightly one another's motions, and the observations. The idea of an ultra-Uranian planet occurred to Bouvard, the German astronomer and mathematician Friedrich Wilhelm Bessel, the U.S. astronomer William Joseph Hussey, and others. At Mannheim, the German astronomer Friedrich Bernhard Gottfried Nicolai decided that a planet beyond Uranus could account for the apparent inconsistency between older and more recent observations of Halley's Comet.

Bessel, in 1842, announced his intention of investigating the problem of the motion of Uranus. Unfortunately, he died before much had been done. Meanwhile an English astronomer, John Couch Adams and a French astronomer, Urbain-Jean-Joseph Le Verrier, were also tackling the problem energetically.

Adams, at St. John's College, Cambridge, began serious work in 1843 and by September, 1845, had communicated his results to the English astronomer James Challis at Cambridge, and the following month to the Astronomer Royal, Sir George Biddell Airy. Adams said that his calculations would account for the irregularities in the motion of Uranus by the action of an exterior planet whose orbital elements were tabulated. The position of the exterior planet, as derived from Adams' work, was later found to be so close to the true value that two evenings of assiduous searching should have been sufficient to locate it. Adams also demonstrated that the observations of Uranus from 1780 onwards were in good agreement with his theory. Instead of looking for the unknown planet, however, Airy did no more than to ask, in a reply, whether the assumed perturbations would also explain the known error in the calculated distance of Uranus from the Sun.

Meanwhile Le Verrier in France, working on similar lines had published the results of his thorough mathematical investigations in a communication to the French Academy of Sciences on November 10, 1845. A second memoir was presented by Le Verrier on June 1, 1846.

The close agreement between the elements found by the two investigators is clear from the tabulated values (see Table 2) and led Airy to suggest to Challis, on July 9, 1846, that a search for the unknown planet should be made with the Northumberland telescope, then at Cambridge Observatory. Challis commenced on July 29, sweeping each region twice, so that a comparison of the two sets of observations would then reveal the unknown planet by its motion. It subsequently transpired that Challis observed the planet on August 4 but failed to compare his observations with those of the previous night.

Table 2: Orbital Data According to Le Verrier and Adams

elements	Le Verrier	Adams	
		hypothesis I	hypothesis II
Semimajor axis (a.u.)	36.154	38.38	37.27
Eccentricity	0.1076	0.16103	0.12062
Longitude of perihelion	284° 45′	315° 57′	299° 11′
Mean longitude	318° 47′	325° 08′	323° 02′
Epoch	1847, Jan. 1	1846, Oct. 1	1846, Oct. 1
True longitude	326° 32′	328°	329°

On August 31, Le Verrier's final paper was presented to the French Academy and on September 18 he wrote to the German astronomer Johann Gottfried Galle at Berlin suggesting that he should search for the planet, with the hope of distinguishing it from a star by its disklike appearance. On September 23 the planet was located at the first attempt by Galle and the German astronomer Heinrich Louis d'Arrest, not by its disklike appearance but by the absence of an observed starlike body from one of the new star charts prepared at the observatory.

Controversy over the credit for Neptune's discovery

The question whether Le Verrier should receive the sole credit was warmly discussed by French and British astronomers; Adams and Le Verrier took no part in the controversy. Arago went so far as to suggest that the new planet should be named Le Verrier, but this was received with such little favour outside France that he withdrew it, proposing Neptune instead.

The mathematical methods employed in the solution of the problem and, in particular, the choice of the value of the semimajor axis of the orbit have been questioned. At the date of its discovery Neptune was only 1° from the place predicted by Le Verrier and about 2½° from the place predicted by Adams. However, it has since been shown that it was fortunate that the search was made around the middle of the 19th century, because if the orbit worked out by Adams had been used for a search in 1770, then Neptune would have been nearly 30° from its predicted place, and not much less if Le Verrier's orbit had been used. Figure 1 shows how sharply the true and predicted places would have diverged after 1850–60. Recent work on the orbit of Neptune has provided a new value for the mass of Pluto; the revised values for six defining quantities (elements) of the closest elliptic orbit are given in Table 3.

Appearance. The great distance and relative faintness of Neptune make it exceedingly difficult for any details to be observed, even with large telescopes. Limb darkening, or dimming of the image near its edge, is pronounced. A few markings have been noted, but these are so vague that it has not been possible to determine a rotation period from such observations.

Neptune is seen by light reflected from the Sun and its slight greenish hue is due to absorption of the red, orange, and yellow solar rays by its extensive atmosphere. Strong absorption bands in the infrared region of the spectrum were identified in 1932 as being due to methane (CH_4); some unidentified bands at a wavelength of about 7500

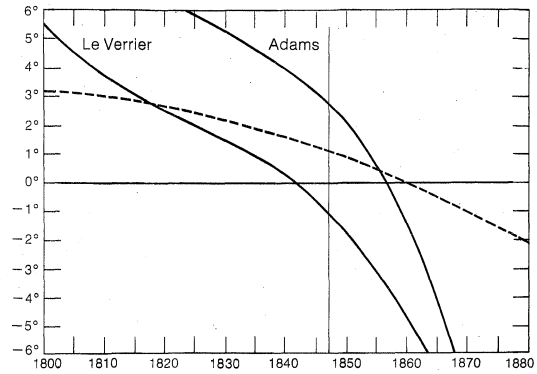


Figure 1: Curves of amounts by which orbits of Adams, Le Verrier, and circular-orbit solution (dashed curve) differ from true position of Neptune. The vertical marks the date of discovery, 1846 (see text).
From R.A. Lyttleton, *Mysteries of the Solar System* (1968); Clarendon Press, Oxford

angstroms, near the deep red end of the visible spectrum, were found by the U.S. astronomer Gerard Peter Kuiper in 1949. It is inferred, though not observed, that the most abundant gases in the atmosphere are hydrogen and helium. A temperature of −205° C has been derived for the atmosphere, and this is in fair agreement with measurements at the Nançay Radio Observatory in France, which gave an upper limit to the disk temperature of −223° C. Observations with the 43-metre (140-foot) radio telescope at the National Radio Astronomy Observatory, Green Bank, West Virginia, indicate that Neptune emits radio waves as if its black-body disk temperature were −193° C ±40° C. For such measurements the planets are considered black bodies in the physicist's sense of the term, meaning they freely absorb and emit radiation of all wavelengths. As the indicated temperature for Neptune is higher than that expected for a planet at Neptune's distance from the Sun, these measures suggest the retention of solar heat under some layer of Neptune's atmosphere (as by the glass panels of a greenhouse), or, another possibility, a strong source of internal heating.

Temperature of Neptune

Table 3: Elements of Neptune's Orbit

Mean anomaly	133° 44' 09".783
Perihelion	270° 03' 30".833
Ascending node	131° 16' 41".893
Inclination	1° 46' 33".651
Eccentricity	0.0118570458
Semimajor axis	29.9871290465

Constitution. The constitution of the planet can only be obtained from theoretical considerations, since the extensive cloudy atmosphere precludes direct observation. Its physical constitution is probably similar to those of Jupiter, Saturn, and Uranus. It has been calculated that there is a rocky core about 20,000 kilometres (12,000 miles) in diameter enveloped in an ice layer 10,000 kilometres (roughly 6,000 miles) thick, with a gaseous envelope 3,000 kilometres (roughly 2,000 miles) deep.

Rotation. In 1928, a rotation period now generally accepted of $15^h{\cdot}8 \pm 1^h{\cdot}0$ was obtained from spectroscopic observations. Light from the portion of the rotating planet moving toward the observer is shifted toward the blue end of the spectrum, while light from the receding part of the surface is reddened. Photographic and visual observations indicated periodic variations in brightness, attributed to rotation, but photoelectric observations have shown no evidence of variation of brightness. The degree of flattening at the poles, caused by the planet's rapid spin, is estimated to be $\frac{1}{50}$. This means the equatorial diameter is about 1,000 kilometres longer than the polar diameter.

Diameter. The U.S. astronomer Edward Emerson Barnard, in 1899, used a filar micrometer—in which two wires or threads are aligned with the edges of the planet's image—to measure the angular diameter directly. The planet's distance being known, the linear diameter was determined as 52,900 kilometres (32,900 miles). The value is still used in astronomical ephemerides today. G.P. Kuiper at the McDonald Observatory in 1949, obtained a value of 44,600 kilometres (27,700 miles). In 1953, a double-image micrometer measurement by the U.S. astronomer Henri Camichel yielded a value of 45,000 kilometres (28,000 miles). These two latter results were in good agreement but differ noticeably from the value given by Barnard. Limb darkening may introduce some uncertainty in the angular measures and a change of only 0".1 in angular measure is equal to a linear change of about 2,000 kilometres (1,200 miles).

Measurements of Neptune's occultation

Successful efforts have been made to derive the diameter from observation of the occultation of a star (its disappearance and reappearance behind the moving planet) from several different places on the Earth. Because of the slow motion and small angular size of the planet such occultations are quite rare. A comparison of the ephemeris of the planet with star catalogs was performed annually from 1952 onward, but the first predicted occultation did not occur until April 7, 1968. Fortunately, the event was successfully observed photoelectrically at observatories in Australia, New Zealand, and Japan (see Figure 2).

From G. Taylor, *Monthly Notices of the Royal Astronomical Society* (1970)

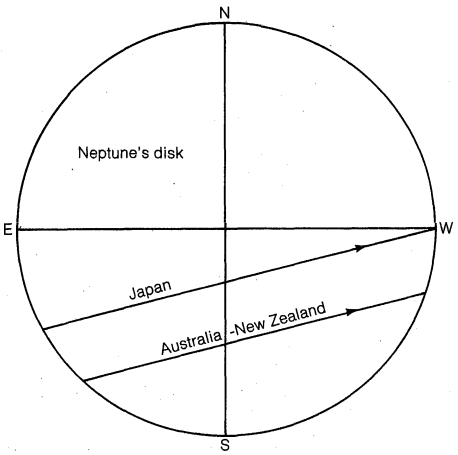


Figure 2: Tracks of star BD-17° 4388 behind disk of Neptune on April 7, 1968. The star was obscured for about 45 minutes for Japanese observers, about 40 as seen from Australia and New Zealand. At the time, the planet's distance from Earth was 2,700,000,000 miles (4,300,000,000 kilometres).

Observations from these countries were highly consistent with each other and an analysis of them yielded two main results: (1) an extremely accurate position of the centre of Neptune relative to the occulted star (the difference in declination at the time of conjunction being determined to an accuracy of better than 0".01); and (2) a new value for the diameter of the planet. Two solutions were made, one assuming Neptune to be a sphere, the other assuming it to be an oblate spheroid. There was little to choose between the two solutions, the value using an oblate spheroid having only a slightly smaller standard deviation. The value quoted for the equatorial diameter of the oblate spheroid was 50,950 kilometres ± 170 kilometres (which included an extra 110 kilometres as a correction for the deflection of light in a gravitational field). This value corresponded to the height at which the first photoelectrically detectable dimming of light occurred and in this sense could be taken as the upper limit of the atmosphere.

With this value for the diameter, the density is 1.54 g/cm^3, close to the value currently given for Uranus, namely 1.56 g/cm^3.

Satellites. Neptune has two satellites, Triton and Nereid. Triton, the larger and brighter of the two was discovered visually by W. Lassell, using a 60-centimetre (24-inch) telescope at Liverpool, England, in 1846, within a month of the discovery of the planet itself. It was originally assumed that Triton revolved around Neptune in the same direction as the planet's rotation, implying that Nep-

tune's north pole pointed south of the ecliptic, like that of Uranus (in those days no retrograde satellites were known). In 1928, however, careful spectroscopic work at Lick Observatory, California, showed that Neptune rotated in the direct sense unlike Uranus; so Triton moved round Neptune in the retrograde direction, opposite to that of Neptune's spin. The orbit of Triton is circular and was seen edge-on from the Earth in 1953. The apparent diameter of Triton has been measured as 0″.173 at a distance of 30·1 a.u. Nereid was discovered photographically by G.P. Kuiper in 1949, using the 208-centimetre (82-inch) McDonald telescope in Texas. The value of the diameter is only an estimate based on its apparent magnitude and an assumed albedo (the fraction of light reflected from the surface).

Discovery of Nereid

Table 4: Satellite Data

	Triton	Nereid
Apparent magnitude	13.5	18.7
Mean distance from Neptune (km)	355,000	5,562,000
Mean sidereal period	$5^d.877$	$359^d.881$
Inclination of orbit to Neptune's equator	159° .9	27° .7
Eccentricity of orbit	0.000	0.749
Diameter (km)	3,700	300
Reciprocal mass (Neptune = 1)	750	?
Density (water = 1)	5.1	?

The orbital plane of a satellite around a spherical planet would be fixed in direction, apart from an extremely slow change produced by the action of the Sun. As the plane of Triton's orbit is changing noticeably with time it may be deduced that Neptune is not spherical but is an oblate spheroid, flattened at the poles, indicating rapid rotation. The degree of flattening thus obtained is $\frac{1}{50}$.

A mathematical investigation of the orbits of the satellites has led to the suggestion that Triton could have been drawn in from a highly elliptical or even parabolic orbit, and that this may have caused the high eccentricity of Nereid's orbit. According to these calculations, Triton is getting nearer to Neptune and is expected to approach and collide with the planet or, more likely, to disintegrate at a close approach and form a ring system.

Searches for other satellites indicate that any other satellite in the range 16″–10″ from the planet would have to be fainter than a photographic magnitude of 22·5 to remain undetected. Farther out the limit is altered to magnitude 19 or 20. Closer than 10″ to the planet the limit is brighter because of the glare of Neptune itself.

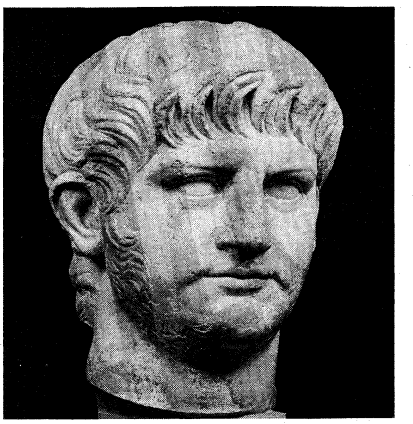
BIBLIOGRAPHY. B.A. GOULD, *Report on the History of the Discovery of Neptune* (1850), a classic work providing detailed accounts of the history of the discovery of the planet; S. NEWCOMB, *Investigation of the Orbit of Neptune* (1866); R.L. DUNCOMBE, W.J. KLEPCZYNSKI, and P.K. SEIDELMANN, "Orbit of Neptune and the Mass of Pluto," *Astron. J.*, 73: 830–835 (1968); R.A. LYTTLETON, *Mysteries of the Solar System* (1968), with a chapter on the history of the discovery and a discussion of the orbits; G.E. TAYLOR, "The Occultation of B.D.-17° 4388 by Neptune on 1968 April 7," *Mon. Not. R. Astr. Soc.*, 147:27–33 (1970), gives a new value for the diameter of the planet.

(G.E.T.)

Nero

History has branded the Roman emperor Nero with the reputation of being a monster of cruelty and has also credited him, on doubtful evidence, with the burning of Rome and the first persecution of the Christians. His appalling family background, stained as it was with incest and murder, no doubt accounted for his instability of character. This displayed itself in the unbounded vanity that caused him to nurse artistic and histrionic pretensions that seriously undermined his imperial dignity and helped to bring about his early death and, with it, the extinction of the Julio-Claudian line, which had descended from Julius Caesar. Proclaimed when he was not yet 17, he was the first boy emperor.

Nero was born Lucius Domitius Ahenobarbus in AD 37 in Rome. His father died in about AD 40, and Nero was



Nero, portrait bust by an unknown artist. In the Museo Nazionale Romano, Rome.
Anderson—Alinari

brought up by his mother, Agrippina the Younger, a great-granddaughter of the emperor Augustus. After poisoning her second husband, Agrippina incestuously became the wife of her uncle, the emperor Claudius, and persuaded him to favour Nero for the succession, over the rightful claim of his own son Britannicus, and to marry his daughter Octavia to Nero. Having already helped to bring about the murder of Valeria Messalina, her predecessor as the wife of Claudius, in 48 and ceaselessly pursuing her intrigues to bring Nero to power, Agrippina eliminated her opponents among Claudius' palace advisers, probably had Claudius himself poisoned in 54, and completed her work with the poisoning of Britannicus in 55. Earlier, Nero had passed his childhood under the menace of the mentally deranged emperor Caligula (Gaius Caesar), his mother's brother, until Caligula was murdered and succeeded by Claudius.

Agrippina's intrigues

Promising start. Brought up in this kind of atmosphere, Nero might well have begun to behave like a monster upon his accession as emperor in 54 but, in fact, behaved quite otherwise. When he was required to sign a death warrant for the first time, he cried "Why did they teach me how to write?" The testimony of his contemporaries depicts Nero at this time as a handsome young man of fine presence but with soft, weak features and a restless spirit, entirely under the influence of his terrifying mother, whom he censured and feared while he adored her. His first speech to the Senate was universally praised as heralding a new Golden Age. Up to the year 59 his biographers cite only acts of generosity and clemency on his account. Indeed, the measures he instigated were those of a wise and virtuous man: he forbade contests in the circus involving bloodshed, banned capital punishment, reduced taxes, and accorded permission to slaves to bring civil complaints against unjust masters. He pardoned writers of epigrams against him and even those who plotted against him, and secret trials were few. More independence was granted to the Senate. The law of treason was dormant: Claudius had put 40 senators to death, but, between the murders instigated by Agrippina in 54 and the year 62, there were no like incidents in Nero's reign. He inaugurated competitions in poetry, in the theatre, and in athletics as counter-attractions to gladiatorial combats. He saw to it that assistance was provided to cities that had suffered disaster and at the request of the Jewish historian Josephus gave aid to the Jews. In all this, he was pursuing an intention that the philosopher Seneca, his old tutor and now his adviser, considered as mad—to remain innocent of any crime.

Artistic pretensions and religious mania. Although it must be admitted that his nocturnal rioting in the streets was a scandal as early as 56, the emergence of real brutality in Nero can be fixed in the 35-month period between the putting to death of his mother at his orders in 59 and his similar treatment of his wife Octavia in June 62. He was led to the murder of Agrippina by her insanity and her fury at seeing her son slip out of her control, to the murder of Octavia by his having fallen in love with Pop-

Affair with Poppaea

paea Sabina, the young wife of the senator (and later emperor) Otho, and by his fear that his repudiated wife was fomenting disaffection at court and among the populace. He married Poppaea in 62, but she died in 65, and he subsequently married the patrician lady Statilia Messalina.

Nero gave rein to inordinate artistic pretensions. He fancied himself not only as a poet but also as a charioteer and lyre player, and in 59 or 60 began to give public performances; later he appeared on the stage, and the theatre furnished him with the pretext to assume every kind of role. He dreamed of abandoning the empire in order to fulfill his poetical and musical gifts, saying "They would adore in me what I am." But this puerile ambition did not suffice, and another, more dangerous inspiration took hold of him: "No one has known all that a prince could do," he said, desiring now to create God. This form of insanity was nevertheless in keeping with the messianic spirit of the age: all peoples—Jews, Greeks, Syrians, Romans, and even Indians and Chinese—were waiting for the new god. From about 63 or 64 Nero was drawn to preachers of novel cults, to Zoroastrian mages, esoteric Judaists, to such men as the Gnostic magician Simon Magus, the Neopythagorean sage Apollonius of Tyana, and perhaps even to St. Paul, who is depicted in a fresco from the Palatine Chapel conversing with the Emperor. His predecessors Caligula and Claudius had shown interest in Eastern religions, but Nero was to go further than they did. Obsessed by the mystery-religion symbols of the sea and the fish, he became a devotee of the Syrian virgin-mother Atargatis, to whom the fish was sacred, then of the virgin-mother Juno-Canathos; the ecstasy that seized him in the Temple of Vesta in 64 showed that there was no limit to his bizarre delirium. Taking advantage of the fire that ravaged Rome in that year, he had the city reconstructed in the Greek style and built the most prodigious palace that a man had ever consecrated to the new god—the fabulous Golden House, which, had it been finished, would have covered a third of Rome. During the fire Nero was at his villa at Antium 35 miles (56 kilometres) from Rome and therefore cannot be held responsible for the burning of the city. Nor, despite the apocryphal additions made centuries later to the *Annals* of the Roman historian Tacitus and to the *Nero* of Suetonius, the Roman biographer and antiquarian, should he be charged with the first persecution of the Christians, because there were then so few Christians in Rome. Paul had just arrived in the city, and the Jews were anxious to learn from the Apostle the nature of this new sect, proof enough that no one had yet told them about it. Nero corresponded with Tiridates, the Armenian prince, and learned the religion of Mithra from him. He gave himself up to debauches and vice but also imposed the wearing of the white toga and had the floral festivals of the Field of Mars transformed into sacred proceedings.

The Golden House

The approaching end. The imperial government had had some success in the east. Their general, Gnaeus Domitius Corbulo, was able to impose an advantageous settlement on Armenia, which had long chafed under Roman suzerainty. Rome now recognized the Armenian ruler Tiridates, but he was compelled to come to Rome to receive his crown from Nero. This secured Armenia as a buffer state against Parthia, Rome's implacable foe in the east. Nevertheless, the provinces were increasingly uneasy, for they were oppressed by exactions to cover the imperial court's extravagance. A revolt in Britain was headed by Queen Boudicca (Boadicea) in 60 or 61, and an insurrection in Judaea lasted from 66 to 70. Nero had many antagonists by this time. The great conspiracy of the Piso family in 65 reveals the diversity of his enemies —senators, knights, officers, and philosophers. The Emperor did not give way to panic; the slaves kept him out of danger by warning him of plots hatching among their masters. And he did not altogether abandon his lenient attitude. Out of 41 conspirators, only 18 died (including Seneca and the poet Lucan), either by order or from fear; the others were exiled or pardoned. At the end of the year 66 he undertook a long visit to Greece, the "country of the gods," which was to keep him away from Rome for 15 months, and during his absence he entrusted the consulate to one of his freedmen. He freed Greek cities in honour of their glorious past, suppressed the pagan gods, and went everywhere garbed as an ascetic, barefoot and with flowing hair. He returned to Rome to establish there the worship of Apollo in his manifestation as the dolphin god, to free slaves, and in his own way to instruct the young nobles, to distribute money among the people, and to surround himself with "priests of the sea" (followers of Neptune).

Visit to Greece

In the four months following his return (from February to June of 68) these delirious actions aroused the enmity not only of the Christian Jews, who called him the Antichrist, but also of the Senate and the patricians, whom he insulted openly, and especially the great Roman families who had been dispossessed by him. Even the soldiers of the legions were scandalized to see the descendant of Caesar publicly act the parts of ancient Greek heroes and re-create in Greek fashion the legend of a saviour born of a virgin-mother (Canathos). "I have seen him on stage," Gaius Julius Vindex, the legate who rebelled against him, was to say, "playing pregnant women and slaves about to be executed."

At the news of revolts brewing throughout the empire —that of the provincial governor Servius Sulpieius Galba in Spain, the rebellion of Vindex at Lyons in Gaul (France), and others on the eastern frontier—Nero only laughed. Instead of taking action, he composed hymns and had hydraulic organs made. Arming his priestesses with light shields, he said "I have only to appear and sing to have peace once more in Gaul." Meanwhile, the legions made Galba emperor, and the Senate condemned Nero to die a slave's death: on a cross and under the whip. The Praetorian Guard, his palace guard, abandoned him, and his freedmen left to embark on the ships he kept in readiness at Ostia, the port of Rome. He was obliged to flee the city. According to Suetonius, he stabbed himself in the throat with a dagger, and his Christian mistress Acte buried him piously in a white sheet in conformity with his wishes. According to another version (recounted by Tacitus and almost certainly fiction) he reached the Greek islands, where the following year (69) the governor of Cythnos (modern Kíthnos) recognized him in the guise of a red-haired prophet and leader of the poor, had him arrested, and executed the sentence passed by the Senate.

Abandonment and suicide

The emperors Otho and Aulus Vitellius honoured his memory, and the people adored images of Nero for 20 years. The historians Plutarch and Josephus protested against the "calumnies that soiled his memory," but all in vain. Trajan and subsequent Roman emperors execrated his memory, destroyed his works, including the Golden House, as well as the many works in which Lucan, Plutarch, Rusticus, and many other writers commemorated the period of his reign. Under these emperors, Tacitus, Suetonius, and later Dio Cassius wrote the biographies of Nero known today. Lastly, the Christian Church, triumphant under the emperor Constantine I the Great, furthered the legend of Nero as a monster and even added to it in the 15th century by drawing upon the apocrypha of Sulpicius Severus, a Christian ascetic and historian of the early 5th century, to transform Nero into Antichrist, the first executioner of Christians.

BIBLIOGRAPHY. Up to the last century the only accepted sources for Neronian history were the *Annals* and the *Histories* (Book I) of TACITUS, the *Lives of the Caesars* by SUETONIUS, and the *Roman History* of DIO CASSIUS. The monstrous image given of the Roman emperor by these tendentious works was the only one known to all of the dramatists and novelists, from RACINE'S *Brittannicus* to DUMAS' *Acté* and the celebrated *Quo Vadis* of SIENKIEWICZ. The discovery of Nero's oration announcing to the Greek cities their liberation opened the first breach in the legend: MAURICE HOLLEAUX, *Discours prononcé par Néron à Corinthe en rendant aux Grecs la liberté, 28 November 67 J.C.* (1889). A number of Italian studies then appeared questioning the traditional stories about Nero and the burning of Rome, and the first persecution of the Christians. ARTHUR WEIGALL in *Nero, Emperor of Rome* (1930, reprinted 1950), proposed a less sanguinary—and almost amiable—portrait of the Roman emperor. Among the most recent contributions to the rehabilitation of the fifth

emperor of Rome are GEORGES ROUX, *Néron* (1962); GILBERT CHARLES-PICARD, *Auguste et Néron* (1962; Eng. trans., 1966); and JEAN-CHARLES PICHON, *Saint Neron* (1961), republished with commentary, under the title *Néron et le mystère des origines chrétiennes* (1971).

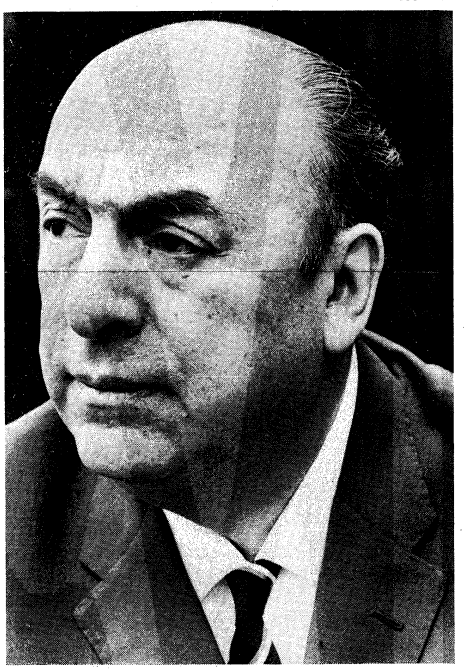
(J.-C.P.)

Neruda, Pablo

One of the most original and fertile poets to write in Spanish this century, Pablo Neruda successfully combined a literary career with a constant concern for the social problems of his time. He has been referred to as "the poet of enslaved humanity." He was awarded the Lenin Prize for Peace, and the Nobel Prize for Literature in 1971, during which time he served as the Chilean ambassador in France.

Neruda was born on July 12, 1904, in Parral, a small town in central Chile, the son of José del Carmen Reyes, a railway worker, and Rosa Basoalto. He was christened Neftalí Ricardo Reyes. His mother died within a month of Neruda's birth, and two years later the family moved to the "new lands" of the south and settled in Temuco, where his father remarried. The boy came to love his stepmother as if she had been his real mother. "In this frontier—or 'far west'—of my country, I was born to life, land, poetry, and rain," Neruda later wrote.

Camera Press—Pix



Neruda.

In 1910 he entered the Temuco Boys' School and completed his secondary schooling in 1920. He began writing poetry when he was ten; some of it was later published (under a pseudonym in order not to annoy his father) in student publications. In 1920 he began using the name Pablo Neruda, which he was legally to adopt in 1946.

When he was 12, Neruda met Gabriela Mistral, the famous Chilean poet who introduced him to the great classic writers who were to influence his choice of career and his political development. He became a militant anarchist and translated the work of Jean Grave, the outstanding theorist of Peter Kropotkin's "anarchist communism."

In 1921 Neruda left Temuco for the capital, Santiago: his immediate goal was to receive his degree in French from the teachers' college, but he found himself devoting more time to his poetry than to his studies. The romantic provincial student invaded the literary life of the Chilean capital with his poet's cape (which had been his father's railroad cape) over his shoulders and wearing a wide sombrero. In October of that year, he won first prize in the Federation of Chilean Students poetry contest with his "La canción de la fiesta," and he began publication of his literary works in the federation's magazine, *Claridad*.

Early poetry

In 1923 his first book, *Crepusculario*, was published at his own expense and financed by selling what furniture he had and the watch his father had given him. The following year he found a publisher for what was to become his most widely read book: *Veinte poemas de amor y una canción desesperada* (1924; *Twenty Love Poems and a Song of Despair*, 1969). The work achieved immediate success and has retained its popularity over the years. Neruda himself was surprised that these poems, written with such pain, have brought consolation to new lovers. "By some miracle which I do not understand, this tormented book has opened the road to happiness for many, many people."

At 20, and with two books published, Neruda became one of the best known Chilean poets. He abandoned his French studies and devoted himself entirely to poetry. He wrote *Tentativa del hombre infinito* ("Attempt of the Infinite Man"); *Anillos* ("Rings"), in collaboration with Tomás Lago; and *El hondero entusiasta* ("The Enthusiastic Slingshooter").

In 1927 he was named honorary consul in Rangoon, Burma, and for the next five years he represented his country in Asia. From Rangoon he went to Colombo, Ceylon, Batavia (modern Jakarta), Java—where he married his first wife, a Dutch girl, Maria Haagenar—and Singapore. He attended the "Pan Hindu" Congress in Calcutta. His books recalling these trips are full of pleasant anecdotes—moving, but sometimes pathetic; his life at the time was one of loneliness, relieved only by his romance with a Burmese girl, Jossie Bliss. It was during these years in South Asia that he wrote *Residencia en la tierra* (*Residence on Earth*). In 1933 he was named Chilean consul in Buenos Aires, where he began a friendship with the Spanish poet, Federico García Lorca, who was visiting the city. The following year he was transferred to Barcelona, and later to Madrid, where he was married again, this time to Delia del Carril.

Friendship with García Lorca

With the same impact he had made on his own country's literary life, he entered the European world in which he and his poetry quickly became known. The Spanish poets welcomed him in *Homenaje a Pablo Neruda* ("Homage to Pablo Neruda"), carrying the signatures of representatives of Spain's intelligentsia.

These years of poetic development were interrupted suddenly by the outbreak of the Spanish Civil War in 1936. The execution of his friend García Lorca, the imprisonment of Miguel Hernández, "the blood in the streets," all contributed to the maturing of the Chilean poet's political attitudes. He was to say, "The world has changed, and my poetry has changed. One drop of blood falling on these lines will remain alive in them, indelible like love." He wrote *España en el corazón* ("Spain in the Heart"), which was published during the Civil War in the Republican front lines.

Neruda returned to Chile in 1938 with a group of Spanish refugees. Shortly thereafter, his government sent him to Mexico, where he entered a phase of intense poetic production—much of it determined or inspired by World War II in Europe, and especially by the heroic defense of Stalingrad against the German onslaught. Mexican mural painting also influenced his work at this time.

In 1943 he returned to Chile by sea, receiving a tremendous welcome in every country along the Pacific coast. In 1945 he was elected senator and for the next three years devoted as much time to the problems of his country as he did to his literary career. His political activity was halted, however, when the Chilean government moved to the right. Neruda, a Communist, was forced along with other leftists into hiding. These years of clandestine living proved to be fruitful ones for his writings. He wrote *Canto general* ("General Song"), one of the greatest epic poems written on the American continent. In February 1948 he left Chile, crossing the southern part of the Andes Mountains on horseback and in April attended the Congress of Supporters of Peace in Paris. In June 1949 he visited the Soviet Union for the first time to attend the 150th birthday celebration of Aleksan-

Diplomatic appointment

dr Pushkin. He then visited other parts of Europe and once again went to Mexico. In 1952, after the order to arrest leftist writers and political figures had been withdrawn, he returned to Chile. He married his third wife, a Chilean, Matilde Urrutia.

Although his permanent residence was the Isla Negra on the Pacific, he travelled constantly, visiting, among other places, Cuba in 1960 and the United States in 1966. His poetry was translated into almost every language. He built a house called La Chascona in Santiago on the side of San Cristóbal hill and another in Valparaíso called La Sebastiana. The three houses were notable for their taste and for their collections of ships' figureheads and varied objects collected during his travels.

Political activities

In essence, the poetry of Neruda represents constant change, the kind any man experiences during his lifetime. In his youth, the lyric and erotic *Crepusculario* and *Veinte poemas de amor y una canción desesperada* gave way to the more spiritual *Tentativa del hombre infinito* and *El hondero entusiasta*. Withdrawing from the external world and its creative stimuli into the domain of his inner self, he produced the towering and enigmatic work *Residencia en la tierra*. With this work, he broke with traditional form and created a highly personalized poetic technique, which has become known as *nerudismo*. The poetry he wrote during the Spanish Civil War moved toward a greater realism and awareness of the affairs of the outside world. This accusatory and reality-oriented poetry was followed by the imposing epic *Canto general*, but this grander style was in turn rejected for the more simple themes of a poetry of social criticism.

The variety of his themes and techniques did not decrease *as* he grew older. Love was once again central to *Cien sonetos de amor*, humour to *Estravagario*, and *Memorial de Isla Negra* was a nostalgic autobiography.

Neruda himself confessed that "I will keep on writing a large cyclical poem which has not yet ended because it will only end with the word at the final instant of my life." That life came to an end in Santiago on September 23, 1973.

MAJOR WORKS

POEMS: "La canción de la fiesta" (1921); *Crepusculario* (1923); *Veinte poemas de amor y una canción desesperada* (1924); *El habitante y su esperanza* (1925); *Tentativa del hombre infinito* (1926); *Anillos* (1926); *El hondero entusiasta* (1933); *Residencia en la tierra, 1925–1931* (1933); *Residencia en la tierra, 1925–1935* (1935); *Tercera residencia, 1935–1945* (1947); *Canto general* (1950); *Los versos del Capitán* (1952); *Las uvas y el viento* (1954); *Odas elementales* (1954); *Nuevas odas elementales* (1956); *Tercer libro de las odas* (1957); *Estravagario* (1958); *Navegaciones y regresos* (1959); *Cien sonetos de amor* (1959); *Cantos ceremoniales* (1961); *Plenos poderes* (1962); *Memorial de Isla Negra* (1964); *La barcarola* (1967); *Las manos del dia* (1968); *La espada encendida* (1970); *Las piedras del cielo* (1970).

TRANSLATIONS: *Residence on Earth and Other Poems* (1946); *The Man Who Told His Love* (1959); *Selected Poems of Pablo Neruda* (1961); *The Elementary Odes of Pablo Neruda* (1961), a selection; *The Heights of Macchu Picchu* (1966); *We Are Many* (1967), a selection; *Twenty Poems of Pablo Neruda* (1967); *Pablo Neruda: A New Decade*, ed. by Ben Belitt (1969); *Twenty Love Poems and a Song of Despair* (1969); *Selected Poems*, ed. by N. Tarn (1970).

BIBLIOGRAPHY. The first critical survey of the work of Pablo Neruda was that of AMADO ALONSO, *Poesía y estilo en Pablo Neruda* (1940), which covers his early work and deals in particular with *Residencia en la tierra*. In 1957, two critical works on Neruda appeared—one, by the poet MARIO JORGE DE LLELIS, *Pablo Neruda*; and the other by ROBERTO SALAMA, *Para una crítica a Pablo Neruda*. The Chilean critic RAUL SILVA CASTRO, who deals with the works of Neruda from his childhood on, published a study of Neruda's work, *Pablo Neruda* (1964). MARGARITA AGUIRRE, *Genio y figura de Pablo Neruda* (1964), is the first biography of the man, complete with chronology and details of his life; also published as *Las vidas de Pablo Neruda* (1967), an expanded version with photographs. The excellent Uruguayan critic EMIR RODRIGUEZ MONEGAL has published one of the most interesting books about Neruda: *El viajero inmóvil* (1966). HERMAN LOYOLA, *Ser y morir en Pablo Neruda, 1918–1945* (1967), contains good material relevant to Neruda's childhood and adolescence. Neruda himself published an autobiography in the magazine *O Cruceiro Internacional* (1962); it appeared in 10 consecutive issues and has never been reproduced, even though it is of tremendous value. The second edition of the *Obras completas* of Neruda, under the direction of MARGARITA AGUIRRE, was published in 1968 in two volumes. It contains two of the best collections of studies of Neruda's work, that of HERNAN LOYOLA and that of ALFONSO ESCUDERO; these constitute the best of all that has been published about the author. The Pablo Neruda Foundation, located at the University of Chile, holds both Neruda's library (valuable for its incunabula, first editions, and manuscripts of famous writers) and his collection of shells, which he donated.

(M.Ag.)

Nerve Impulse

Nerve impulses are electrical units of information; carried rapidly by nerve fibres throughout the bodies of multicellular animals, nerve impulses consist of brief (about one-thousandth of a second) electrical events, called action potentials, which occur across the membrane of a nerve fibre in response to a stimulus. Because an action potential travels as a wave of excitation along a nerve fibre, it also may be called a propagated potential. An action potential is conducted along a nerve fibre at speeds that range from one to 100 metres per second, depending on the properties of the fibre and its environment. The characteristics of an action potential are independent of the way in which it is initiated, although the frequency of action potentials conducted along a nerve fibre may vary. A nerve impulse is transmitted from one nerve cell to another at a junction called a synapse; nerve fibres that innervate muscles transmit impulses at regions of contact called neuromuscular junctions.

Impulses normally travel in only one direction, sometimes called the orthodromic direction, along a fibre. If, however, a nerve fibre is stimulated artificially by an electric shock, the fibre will conduct impulses in the opposite, or antidromic, direction. Both orthodromic and antidromic impulses are conducted at the same velocity. If electric shocks are used to stimulate two nerve impulses conducted in opposite directions along a nerve fibre, both impulses stop at the point at which they collide; nerve fibres, therefore, are not two-way communication lines. One-way conduction of impulses along nerve fibres is the basis for their classification according to function. Impulses that arise in sense organs (*e.g.*, eyes) and provide sensory information to the central nervous system (brain and spinal cord) are said to travel in afferent, or sensory, nerve fibres. Impulses that arise in the central nervous system and travel to effector organs (*e.g.*, muscle fibres) are said to travel in efferent, or motor, nerve fibres. Impulses in sensory nerve fibres, therefore, are conducted toward the central nervous system; those in motor nerve fibres, away from it. Action potentials, detectable in nerve fibres when they are conducting information (Figure 1), are not peculiar to nerve tissue; they play an important role in many muscle fibres and also occur in certain excitable plant cells.

This article includes the following topics: the structure

From (A–C) E.D. Adrian in *The Journal of Physiology* (1933), (D–G) Barron in J.F. Fulton, *Textbook of Physiology*, 17th ed. (1955), W.B. Saunders Company, Philadelphia

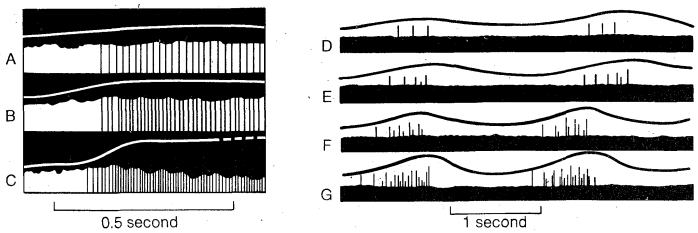


Figure 1: *Electrical impulses in the respiratory nerves of a cat.* (A-C) Sensory impulses traveling from the lungs to the brain in the vagus nerve. White lines indicate the degree of inflation of the lungs. The impulses, recorded with external electrodes from a strand of the nerve trunk, appear as black vertical lines. The deflections produced by the impulses are shown downward and the frequency of the impulses rises as the distension of the lungs is increased. (D-G) Motor impulses traveling to the diaphragm in the phrenic nerve. Upper tracing indicates the respiratory movement; impulses are shown below as upward deflections. One fibre carries the impulses in (D), two in (E) and (F), and three in (G).

of nerve cells, the general features of a nerve impulse, the way in which an action potential originates and is conducted as a result of electric currents it generates, the mechanisms involved in transmitting impulses between cells, and ways by which the nerve impulse is studied.

THE NEURON

Nerve fibres, which are extensions of nerve cells, or neurons, may vary in length from a fraction of a millimetre to a metre or more in a large mammal. Many fibres generally comprise the large nerves found outside the central nervous system and the fibre tracts inside. At the termination of fibre tracts, individual fibres separate and ramify, either to supply effector organs (*e.g.*, muscle fibres) or sense organs (*e.g.*, eyes) or to contact other nerve cells (see NERVES AND NERVOUS SYSTEMS).

Structural features. A neuron (Figure 2A) is a specialized cell consisting of a cell body and fibres. The internal medium (cytoplasm) of a cell body, which may range in diameter from five to 100 microns (one micron equals $\frac{1}{1,000}$ millimetre), contains a nucleus, within which is a smaller structure called a nucleolus, and various other components; *e.g.*, mitochondria, lysosomes, a Golgi apparatus, and Nissl's granules. The nerve fibres extending from nerve-cell bodies are either axons or dendrites; an axon carries nerve impulses away from the cell body; dendrites carry information to it. Both may contain very small fibres called neurofibrils and neurotubules. A nerve cell usually has a single axon, which extends from a raised region known as the axon hillock, and numerous dendrites, which extend from many points on the surface of the cell body. The axon, often the largest nerve-cell fibre, may branch many times, transmitting impulses onward to other nerve cells or conducting impulses to effector cells. In certain nerve cells, particularly sensory neurons, the principal fibre is a dendrite, which transmits impulses from distant sense organs to a cell body located near the central nervous system. The number of fibres extending from nerve cells is used to classify them; *i.e.*, a cell with many dendrites is termed multipolar, a cell with but one dendrite and one axon is termed bipolar (*e.g.*, as found in the retina of the eye, in ganglia of the middle ear, in the olfactory mucous membranes of the nose), and a cell with a single fibre is termed unipolar (*e.g.*, sensory neurons of spinal ganglia).

Axons and dendrites

From (A) Sir J. Eccles, "The Synapse," copyright © 1965 by Scientific American, Inc., all rights reserved; (B) B. Katz, "How Cells Communicate," copyright © 1961 by Scientific American, Inc., all rights reserved; (C,D) F.O. Schmitt, *Reviews of Modern Physics*, vol. 31, p. 461 (1959)

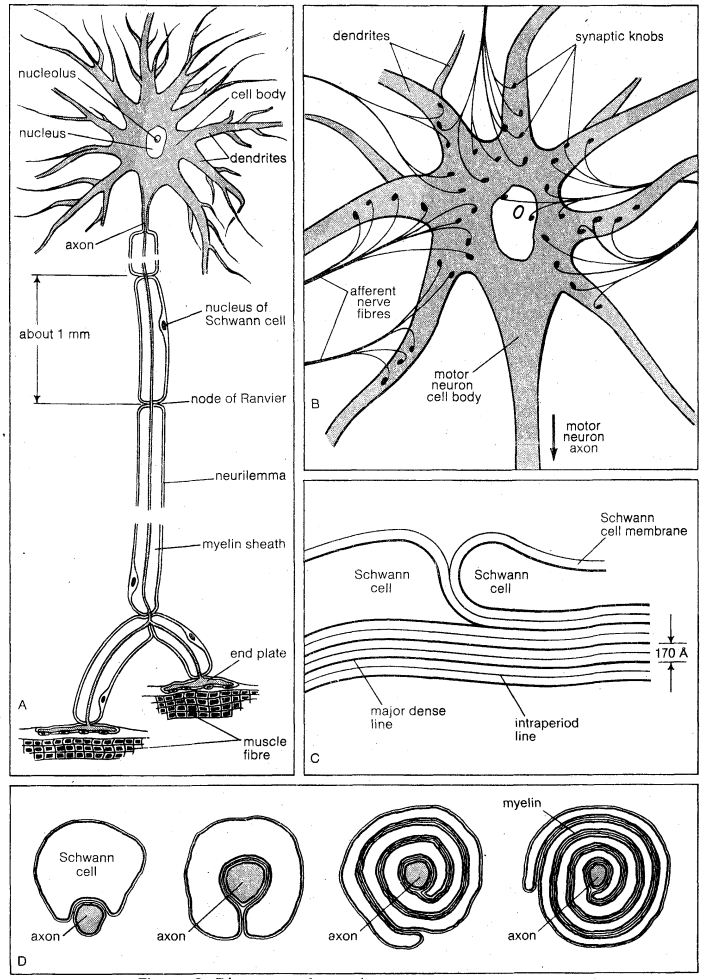


Figure 2: *Diagrams of vertebrate neurons.*
(A) Main features of a myelinated motor neuron. (B) Termination of afferent nerve fibre axons. (C) Outer edge of myelin sheath in transverse section. (D) Stages in formation of myelin sheath.

Myelin. Nerve fibres generally are surrounded by special cells, called Schwann cells, which form a membranous sheath, the neurilemma. The specialized Schwann cells of the large nerve fibres of vertebrates (also prawns and shrimps) wind many times around the fibre to form a multilayered deposit, called myelin, which consists of fatty, or lipid, membranes. Nerve fibres with these lipid-rich membranes are said to be myelinated; those without them, unmyelinated. The myelin in myelinated nerve fibres acts as an insulator, thereby hastening conduction of nerve impulses along them. Gaps occur in the Schwann-cell layer of myelinated fibres at intervals of approximately a millimetre; the gaps, called nodes of Ranvier, result in direct exposure of the nerve-fibre surface to the extracellular space. Although the small nerve fibres of vertebrates (*i.e.*, those with a diameter of less than about one micron) and the nerve fibres in most invertebrates have Schwann cells surrounding the fibres, they lack multiple membranous layers and, thus, are unmyelinated fibres. Nerve-cell bodies lack myelin but are surrounded by supporting cells, known as glia, in the central nervous systems of vertebrates; in invertebrates, the nerve fibres may lack obvious association with specialized supporting cells.

Nodes of Ranvier

The electron microscope reveals the myelin sheath as a repetitive array of dark lines, known as major dense lines, which alternate with less dense lines, called intraperiod lines (Figure 2C). Various stages involved in the formation of the myelin sheath from Schwann cells may be seen in Figure 2D. Myelin is composed of a variety of fatty materials (*e.g.*, cholesterol, cerebrosides, sphingomyelin, and sphingophospholipids), protein, and water. The layer-like character and the orientation of the molecules in myelin enable it to affect the passage of light unequally in different directions; *i.e.*, myelin is birefringent. The myelin sheath is lost at the site of nerve-fibre termination; that is, at sense organs, at effector organs, or against cell bodies and dendrites of other nerve cells (Figure 2B).

Synaptic vesicles. A functional connection between two nerve cells, which allows nerve impulses to be transmitted onward from one cell to the next, is known as a synapse. At a synapse, a nerve-fibre ending swells to form a small, often buttonlike structure covered with the cell membrane; it contains large numbers of small vesicles, called presynaptic vesicles, each about 500 angstrom units in diameter (one angstrom unit [Å] is 10^{-7} millimetre). The fibre, called a presynaptic fibre, is separated from the next fibre, called a postsynaptic fibre, by a space (synaptic cleft) 150 to 200 Å wide. Presynaptic and postsynaptic fibres so close to each other that their membranes appear to fuse are known as tight junctions.

Regeneration. The ability to conduct impulses does not depend on the integrity of a whole nerve cell. Nerve fibres separated from their cell bodies, for example, may retain the ability to be excited for hours or days, even though they begin to degenerate within a short time. The myelin sheath and the nerve fibre eventually swell and disintegrate. The cell body may survive after removal of its largest axon or dendrite, although the Nissl granules disappear for a time. Fibres of neurons outside the central nervous system may regenerate by growth outward from the cell body; this regeneration process, which may restore nerve function, proceeds in man at the rate of about one to two millimetres per day.

CHARACTERISTICS OF ARTIFICIALLY STIMULATED NERVE FIBRES

Effects of stimulation. *All-or-none law.* Although nerve fibres may be excited artificially by any of several methods (*e.g.*, chemicals, mechanical methods), electrical stimulation is the most convenient, the most easily controlled, and the least damaging. A brief electrical shock applied as a stimulus to a nerve fibre must be of a certain strength, called the threshold, in order to excite the fibre and to initiate an action potential. Increases in shock strength above the threshold do not affect either the size or the time course of an action potential; this phenomenon, called the all-or-none law, occurs because the nerve impulse, like the firing of a line of gunpowder, derives its energy locally and not from the initial stimulatory process. For the same reason the action potential does not diminish in size as it is conducted along a nerve fibre; *i.e.*, conduction of a nerve impulse is without decrement. If something is applied at one region of a nerve fibre to modify, but not extinguish, an action potential, the modification affects the action potential only in the region of application, not along the entire fibre.

Concept of threshold

In nerve trunks containing many individual fibres, the compound action potential is graded, because the thresholds of the individual fibres vary; that is, a minimal shock stimulates one fibre, and increases in shock strength cause additional fibres to respond, until all are stimulated, after which the response is independent of increasing shock strength.

Tetanus. Repeated stimuli applied to a nerve fibre, called a tetanus, results in repeated action potentials and the fibre is able to respond to such stimuli for long time intervals without becoming fatigued. The nerve is unlike the junction between a nerve fibre and a muscle fibre in this respect; *i.e.*, the junction generally fatigues more readily than do nerve fibres. If a nerve fibre is stimulated rapidly and repeatedly, the threshold of excitation increases after the first shock; with extremely frequent stimuli (about 1,000 per second), however, some stimuli fail to cause an action potential, indicating that there is a limit to the rate at which nerve fibres can conduct information.

Refractory periods. After an effective stimulus, a nerve fibre is briefly unable to respond to a second stimulus, regardless of its strength; termed the absolute refractory period, this interval lasts about half a millisecond in large mammalian nerve fibres. The absolute refractory period is followed by the relative refractory period, during which the threshold of excitation is abnormally high. Strong electrical shocks are needed to stimulate the nerve during the relative refractory period, and the action potential of a nerve fibre stimulated at this time is small and conducted less rapidly than usual. Later excitability changes are related to small electrical changes (afterpotentials) that occur in a fibre following the action potential.

Latent addition. Although a subthreshold shock applied to a nerve fibre does not stimulate a conducted action potential, it affects the fibre; that is, two stimuli, inadequate if given independently, may, if given in succession, stimulate the fibre by the process known as latent addition. Subthreshold shocks closely approximating the threshold strength cause localized electrical changes in the fibre that are difficult to distinguish from the stimulating current. The indication of the stimulating current on graphic recordings of an artificially induced nerve impulse is known as shock artifact; useful as an initial time marker, it is caused by a slight leakage of charge between the stimulating apparatus and the recording device (see Figure 4C, D). A large shock artifact occurs close to the stimulating electrodes.

Shock artifact

Strength–duration curve. The excitability of a nerve fibre in relation to previously applied stimuli reveals limitations of the rate at which information can be conducted. The conditions required for excitation of a nerve fibre by a single shock, on the other hand, have been studied in attempts to clarify the mechanisms involved in excitation. Usually, the action potential originates at the cathode (negative electrode) as a shock is applied; occasionally, particularly with nervous tissue in poor condition, the action potential begins when current flow ceases, and the anode (positive electrode) is the site of excitation, the effect being called anodal break excitation.

When very brief shocks are applied to a nerve fibre, a minimal quantity of charge must pass to stimulate the fibre. Its amount depends on conditions such as the properties of the fibre and the position of the stimulating electrodes. Indeed, very high frequency currents, in which little charge flows, can be applied to tissues to heat them and to stop bleeding. The minimal current strength needed to stimulate a fibre, regardless of the time it flows, is called the rheobase. When a current stronger than the rheobase passes through a fibre, it flows for a certain time interval, known as the utilization time, before the fibre is excited. A graphic representation of steady current against the time required for that current to excite a fibre is called a strength-duration curve (Figure 3A). Such curves are used to calculate the time required for a current twice as large as the rheobase to cause excitation. This time, known as the chronaxie, has been determined for many types of nervous tissue. Strength-duration curves have the form expected if excitation of a nerve fibre requires a specific change in voltage across a leaky condenser; that is, a change in voltage to a fixed value must be reached in a fibre that accumulates and holds electrical charge but is not perfectly insulated, so that some charge is lost through leakage. Other evidence indicates that the cell membrane acts as a leaky condenser in nerve fibres.

Chronaxie

Adapted from B. Katz, *Electric Excitation of Nerve*, published by Oxford University Press as an Oxford Medical Publication

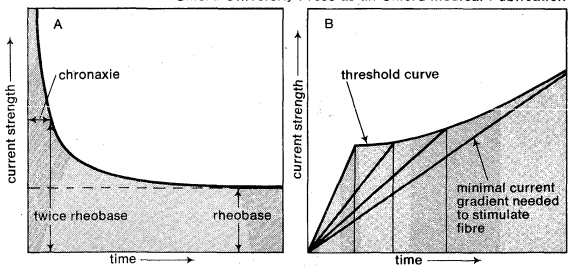


Figure 3: *Requirements for electrical stimulation.*
(A) Strength–duration curve showing threshold conditions.
(B) Illustrates threshold requirements for currents that increase steadily with time. Ordinate is current strength, abscissa is time. Triangular constructions show time course of some currents that achieve threshold conditions (see text).

When slowly increasing, instead of steady, currents are used to excite a nerve fibre, large currents are required before the fibre responds; indeed, if the rate of increase of current strength is sufficiently low, the nerve fibre will never be stimulated, regardless of the final current strength (Figure 3B). The lack of responsiveness of excitable tissue to slowly increasing currents occurs because it adapts to the current, a characteristic called accommodation.

NATURE OF THE NERVE IMPULSE

The nerve impulse or action potential can be detected as it passes electrical connections placed outside the nerve fibre. It then appears as a brief, small, electrically negative wave passing the electrical connections. A much larger electrical charge is noted if one connection is in-

side the fibre; it is then found that the inside of the fibre is electrically negative when it is resting but briefly becomes electrically positive as the action potential passes the electrical connection.

These potentials principally originate from the unequal distribution of ions (charged particles) across the cell membrane. Positively charged potassium ions occur in high concentration within the nerve cell; positively charged sodium ions are in high concentration outside. The ions tend to diffuse away from the regions where they are concentrated.

In the resting state, the membrane allows potassium ions to move out from the cell fairly readily; the membrane is said to be selectively permeable to potassium ions. Because positively charged potassium ions move out of the cell, the contents of the resting cell are electrically negative. The outward movement of potassium ions stops when the electrical potential of the cell is sufficiently negative to overcome the tendency of the remaining potassium to diffuse outward. The size of the electrical potentials generated in this way can be calculated when the concentrations of the ions on either side of the membrane are known (see also MEMBRANE, BIOLOGICAL).

When very small steady changes in the resting potential of a membrane are measured, the membrane behaves as a simple resistance to the flow of current. The nerve fibre acts as an electric circuit equivalent to that of a submarine cable or core conductor; that is, the nerve fibre may be considered as having a core of cytoplasm, which is a good conductor of electricity, surrounded by a membrane, which is a poor conductor and spreads the current along the fibre. Mathematical descriptions of the distribution of weak currents in nerve fibres and their associated voltage displacements are based on these assumptions and give an account of such phenomena.

The core-conductor theory. Although a core-conductor circuit, in which the membrane acts as a simple resistance, can not explain the events that occur during an action potential, it does provide the basis for a well-established theory of the way in which an action potential is conducted. Where the action potential occurs in a nerve fibre, the inside of the fibre is positively charged, and currents flow outward from this zone. The properties of an unmyelinated fibre (*i.e.*, a well-conducting cytoplasm, or core, surrounded by a poorly conducting membrane), ensure the spread of current from the excited region into neighbouring regions of the fibre; the current spreads outward through the membrane in advance of and behind the zone occupied by the action potential. Immediately behind the action potential is a region in which the membrane is in a refractory condition; that is, it cannot be stimulated. The theory, known as the local-circuit theory, states that the local currents ahead of the action potential excite the membrane (Figure 4A), causing the continuous, forward spread of the action potential in unmyelinated nerve fibres.

Saltatory theory

Although the local-circuit theory also applies to myelinated nerve fibres (those surrounded by layers of fatty membranes containing myelin), the process, according to the saltatory theory, is slightly modified. Myelin, a good insulator, covers an entire fibre, with the exception of the gaps (nodes of Ranvier); moreover, only small quantities of charge are stored in the myelinated regions between nodes. Local currents through the fibre membrane occur ahead of an action potential only at a node (Figure 4B). Conduction of an action potential in myelinated fibres, therefore, is a discontinuous process, the active region jumping from one node to the next, thereby hastening conduction.

It follows from the local-circuit theory that the velocity of conduction of an impulse increases if electrical resistance outside or inside the fibre decreases. In unmyelinated fibres, for example, conduction velocity varies directly with the square root of the fibre diameter; in myelinated fibres, it varies with the diameter of the fibre. The relationship between fibre size and velocity of conduction causes a dispersion of impulses when a stimulus is given to a nerve trunk containing fibres with different diameters. As the action potentials are recorded progressively farther from the point of stimulation, fibre types become recognizable; *i.e.*, the action potentials of larger and, therefore, faster conducting fibres reach a given point in the nerve trunk sooner than those of smaller, hence slower conducting, ones.

Adapted from (A,B) *Proceedings of the Royal Society* (1958); (C,D), *Nature* (1961)

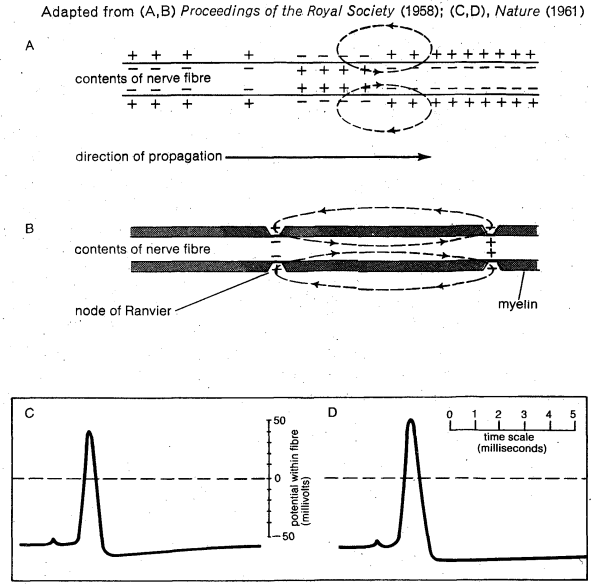


Figure 4: *Action potentials and the local circuit theory for conduction of action potential.*
(A, B) Local circuit theory for conduction of an action potential in an unmyelinated nerve fibre (A) and a myelinated nerve fibre (B). Arrows show direction of electric currents at a given moment in time. (C, D) Two action potentials recorded across the membrane of a giant squid nerve fibre. In D, the fibre contents were replaced by a solution containing potassium sulfate.

Electrical events. *The resting potential.* Electrodes, usually silver wires coated with silver chloride, arranged to make indirect contact (through a salt solution) with nervous tissue, are used to detect both the action potential associated with excitation and the potential, called a resting potential, that occurs between the inside and the outside of the resting cell. The resting potential is most easily detected after a nerve has been crushed. The crushed region of a nerve is electrically negative with respect to an intact zone, and this, the injury potential, is a representation of the resting potential across the membrane of a nerve fibre. Crushing damages the fibre membranes, exposing the fibre contents to the electrode. The injury potential is less than the value of the resting potential because the fluid in the region between the electrodes reduces the measurement. Techniques other than crushing are used to measure accurately the resting potential of nerves. In one technique, used with giant nerve fibres, long electrodes, consisting of fine glass tubes containing salt solution and wires, are manipulated inside the fibre. In another technique, small electrodes, made from glass capillaries with a very small tip diameter and filled with salt solution, are pushed inward through the side of a fibre or nerve-cell body without causing excessive damage.

The resting potential for nerve fibres and several other types of electrically excitable cells lies in the range of 60 to 95 millivolts, with the inside of the cell negatively charged. If the inside of a cell becomes more electronegative (*i.e.*, if the potential is increased over that of the resting potential), the membrane or the cell is said to be hyperpolarized. If the inside of the cell becomes less negative (*i.e.*, the potential is decreased below that of the resting potential), the process is called depolarization. The action potential occurs as a brief depolarization when the inside of the fibre becomes positively charged. Small potential changes (afterpotentials) often occur after an action potential, before the resting potential recurs; the slow decrease in potential, which sometimes occurs at

the end of the action potential, is termed a negative afterpotential, and subsequent hyperpolarization of the membrane is termed a positive afterpotential. Both positive and negative afterpotentials depend upon previous activity in the fibre and, at least in vertebrate nerve fibres, may cause the increase in threshold that follows the relative refractory period.

The ionic theory. This theory accounts for the origins of the resting and action potentials. The resting cell membrane is relatively permeable to positively charged potassium ions, which tend to diffuse outward from the cell contents, where they occur in rather high concentration. They leave the cell contents negatively charged, thus giving rise to the resting potential. During an action potential (Figure 4C), the inside of a cell becomes electrically positive to the extent of 30–60 millivolts, so that the charge is reversed from that found in the resting nerve cell; an excited membrane does not become permeable to all ions, however, even though a decrease in membrane resistance accompanies the action potential.

The sodium hypothesis

The observation that excitable tissues generally require sodium ions led two nerve cell physiologists, A.L. Hodgkin and Sir Bernard Katz, to postulate that an excited membrane becomes highly permeable to positively charged sodium ions. These are found in small concentration inside fibres; the positive charge inside a cell during an action potential is caused by the entry of sodium ions. Calculations similar to those used to determine resting potentials, but dependent on sodium-ion concentrations rather than those of potassium, can be used to predict fairly accurately the potential (in millivolts) across a membrane at the peak of an action potential. The selectivity of the membrane for sodium ions is not perfect, however, and nerve fibres do not become quite so positive internally as calculations indicate. In accord with the sodium hypothesis, nerve fibres commonly cannot be excited if sodium is removed from the region outside the membrane; however, certain ions (*e.g.*, lithium ions), most of which do not normally occur in significant concentrations in tissues, may substitute for sodium ions, another indication that the increase in membrane permeability is not completely selective for sodium ions.

The sodium hypothesis has been used to predict the increase in the sodium-ion content of nerve fibres following stimulation. The charge required to alter the voltage across a cell membrane by the amount observed during an action potential, about 120 millivolts, can be determined, and the numbers of sodium ions required to carry this charge can then be calculated. In the squid nerve, the number of sodium ions that enter the cell exceeds calculated values by a factor of three, so that sodium ions can account for the change in potential across a stimulated membrane.

Role of potassium ions

If the entry of sodium ions into a fibre were not balanced by the exit of some other ion of similar charge, an action potential could not decline from its peak value and return to the resting potential. The falling phase of the action potential is caused by movement of potassium ions out of the nerve fibre. In squid nerve fibres, stimulation results in a loss of potassium ions from the cell approximately equal to the entry of sodium ions. For a short time after the peak of an action potential the permeability of the membrane to potassium ions is very high, and they are able to cross the membrane barrier and to leave the cell more easily than are those in a resting fibre. In an isolated giant nerve fibre from a squid, the period of high potassium permeability allows the membrane potential to enter a positive phase, or positive afterpotential. The increase in membrane permeability to potassium ions, which occurs whenever a nerve fibre is depolarized (*i.e.*, the resting potential decreased), develops after a brief delay; it is responsible for the phenomenon known as delayed rectification, and depolarization maintained at any point in a nerve fibre causes a large outward flow of charge, which is almost entirely potassium ions.

The nerve-cell membrane has the dominant role in the potentials observed in nerve fibres. Schwann cells, which play a minor role, may provide a weak barrier to ions traversing the membrane; *e.g.*, the potassium ions leaving a stimulated nerve fibre. In repeatedly stimulated nerve fibres, potassium ions can temporarily accumulate in a cleft between the Schwann cells and the fibre membrane, thereby reducing the potassium-ion concentration gradient across the membrane. This depolarizing effect of external potassium ions can contribute to afterpotentials in these nerve fibres.

According to the ionic theory, therefore, the changes that occur in nerve fibres during an action potential involve a transient increase in the membrane permeability to sodium ions, which allows them to enter the fibre and, thus, reverse the charge of the inside of the cell (from negative to positive), followed by an increase in the membrane permeability to potassium ions, which allows a charge approximately equivalent to that brought into the cell (as sodium ions) to leave in the form of potassium ions.

Maintenance of ionic concentration gradients. The ionic theory requires that the sodium and potassium ions, which move in response to concentration gradients (*i.e.*, from regions of high concentration to regions of low concentration) during an action potential, must be restored to their original locations if the nerve fibre is to function continuously. The mechanism by which the concentration gradients are restored, called the sodium pump, or the sodium transport system, maintains ionic concentration gradients in many types of cells and indirectly maintains the resting potential across nerve-cell membranes.

Role of the sodium pump

The function of the sodium pump is energy-dependent; *i.e.*, it utilizes energy-rich compounds formed during metabolism, especially adenosine triphosphate (ATP). The energy released from an ATP molecule is used to transport several sodium ions across the cell membrane and out of the cell. The removal of sodium ions is tightly coupled to an uptake of potassium ions into the fibre. The ratio of the number of sodium ions removed from a fibre to the number of potassium ions taken up is probably less than 2:1. Because the number of sodium ions that leave a nerve fibre is greater than the number of potassium ions that enter, the sodium pump establishes an electric current that aids in helping keep the inside of the cell negative. Potentials that originate in the action of the sodium pump are quickly abolished by anything that disrupts metabolism (and the formation of ATP); the resting potential, however, which is dependent on concentration gradients of ions across the membrane, is not destroyed unless the gradients are lost by leakage of ions across the cell membrane.

Two separate mechanisms involved in ionic movements, therefore, occur in the cell membrane: one is concerned with permeability changes accompanying the action potential, during which ions move in response to concentration gradients and electrical potential; the second is concerned with the movements of ions against those gradients.

Accompanying phenomena. The action potential, the most easily detectable aspect of the nerve impulse, is accompanied by other phenomena. A minute quantity of heat, sufficient to raise the temperature of a nerve by only a few millionths of a degree centigrade, is liberated at about the time of the action potential. This initial heat arises from a brief heat pulse, which is liberated at about the time of the rising phase of the action potential, and much of it is reabsorbed as the action potential subsides. The small initial heat, therefore, is the net result of two much larger positive and negative phases of heat production. According to the local-circuit theory for the spread of the action potential, heat should be given off from the local-circuit currents. If, however, the currents arise, as the ionic theory suggests, from the movement of ions in response to concentration gradients, movement would be accompanied by a cooling that would match the heat given off. A better explanation for the positive and negative parts of the initial heat may be that they arise from changes in the cell membrane during an action potential.

Liberation of heat

The initial heat production is followed by a slower and more extensive heat production, known as the recovery

heat, which probably arises from the energy released during metabolic reactions that help to restore the ionic concentration gradients to the resting state. The recovery heat is accompanied by an increase in the rate of oxygen consumption by the nerve cell. Immense technical difficulties are involved in attempts to detect and follow the minute and rapid temperature changes that accompany the nerve impulse.

The action potential is accompanied by a transient change in the birefringence of the nerve; *i.e.*, its ability to affect the passage of light. The effect is extremely small, and the measured light changes in intensity only slightly. The change occurs in or near the cell membrane, and experiments with the voltage-clamp technique (see below *The voltage-clamp technique*) have provided evidence that the phenomenon is associated with the voltage differences across the membrane rather than with membrane currents. Complicated changes that cause scattering of light also accompany the nerve impulse, some are transient, others long-lived; a tetanus, for example, produces a relatively easy to detect and prolonged change in the optical properties of a nerve fibre. Another optical effect involves certain fluorescent dyes, which adsorb onto the fibres and increase in fluorescence when the nerve is stimulated. Optical methods eventually may be refined so as to provide a more convenient method of monitoring impulses in nerves than electrical methods.

TRANSMISSION OF THE NERVE IMPULSE

The conduction of an impulse along a nerve fibre is an electrical process; at the end of a fibre, however, the impulse is transmitted to a neighbouring cell by a chemical mechanism. According to the chemical theory of nerve transmission, the arrival of an impulse at the end of a nerve fibre causes the release of a chemical compound that excites the neighbouring cell. Although several chemical transmitter compounds exist, only a few have been identified; *e.g.*, acetylcholine, gamma-aminobutyric acid. The mechanism of nerve-impulse transmission has been studied particularly at the neuromuscular junction, the point at which a nerve affects a muscle fibre, and at the synapse between motor nerve cells in the spinal cord and other nerve cells.

General features. The impulse in a nerve fibre leading to a synapse (presynaptic fibre) initiates a response in a fibre leading away from the synapse (postsynaptic fibre) after a delay of a fraction of a millisecond. The delay arises because of the electrical discontinuity in transmission; *i.e.*, the intervention of a nonelectrical mechanism. In addition, impulses are transmitted in one direction only across a synapse. The action potential initiated in the postsynaptic cell is preceded by an electrical change called an endplate potential (EPP) at the neuromuscular junction and an excitatory postsynaptic potential (EPSP) at the synapse between nerve cells. Since the full time course of the preliminary change may be masked by the action potential, which occurs when depolarization of the membrane reaches the threshold for excitation, it is studied either by activating only a few synapses on a nerve-cell body or after certain drugs have been applied to a neuromuscular junction. Curare, for example, which diminishes the size of the endplate potential at a vertebrate neuromuscular junction without changing its time course, acts by making the junction insensitive to the chemical transmitter (acetylcholine), thereby preventing the endplate potential from attaining the threshold required to initiate an action potential in the muscle fibre.

Summation and facilitation

Both the EPP (endplate potential) and the EPSP (excitatory postsynaptic potential) differ from the action potential in being capable of summation; *i.e.*, the arrival of a second impulse at a presynaptic nerve terminal shortly after the first is added to it, so that the size of the synaptic potential increases with the number of impulses. One practical consequence of summation is that a graded response is achieved in the relatively inexcitable crustacean muscle fibre by summation of endplate potentials. Successive impulses in the presynaptic fibre occasionally yield successively larger endplate or postsynaptic potentials, a phenomenon known as facilitation. Facilitation arises in part from the release of larger quantities of transmitter substance by successive impulses and, together with summation, plays an important role in the central nervous system. An action potential occurs in the postsynaptic fibre when the transmitter substance is present in quantities sufficient to cause adequate depolarization of the nerve-cell membrane amounting often to about ten millivolts.

In the central nervous system, many thousands of presynaptic fibres may make functional contact with the cell body and dendrites of a single neuron (postsynaptic cell), thereby influencing its response. Incoming impulses must be synchronized in order to excite a postsynaptic fibre to its threshold before it is stimulated to give an action potential. The additive properties of synaptic potentials permit the elaborate integrative action that occurs in the nervous system.

The active process. *Role of the transmitter substance.* The endplate potential (EPP), like the excitatory postsynaptic potential (EPSP), is relatively long lasting; developing in frog muscle in about one millisecond and lasting for about 20 milliseconds, it spreads decrementally (*i.e.*, the size of the potential decreases as it travels from the point of origin) for a short distance. The initial stages represent only the active process of the EPP. The active process is an increase in the ability of the postsynaptic membrane to conduct electricity and is caused by the chemical transmitter. In contrast to the action potential, during which the cell contents become positively charged, however, the active process involves a decrease in the resting potential of the postsynaptic fibre membrane (*i.e.*, the membrane is depolarized) to a potential that approaches zero but does not become positive. The potential arises because the membrane of the postsynaptic fibre undergoes a nonselective and simultaneous increase in permeability to both sodium and potassium ions in the presence of the transmitter.

Fate of the transmitter

The junction between excitable cells is commonly the site of enzymes (organic catalysts) that function in the rapid destruction of the chemical transmitter, thereby ensuring its brief action; postsynaptic potentials are prolonged and enhanced, therefore, if the enzyme is poisoned. This information has medical applications; conditions in which the quantity of transmitter released is too low, for example, may be alleviated by drugs that reduce the activity of the enzyme. Drugs such as prostigmine, eserine, and Tensilon act by interfering with the activity of the enzyme, called cholinesterase, that breaks down acetylcholine at the neuromuscular junction; certain nerve gases act in a similar way. Other substances act differently at the neuromuscular junction; *e.g.*, botulinum toxin prevents the release of acetylcholine, and hemicholinium interferes with its synthesis.

Miniature potentials. One muscle endplate releases a few million acetylcholine molecules for each impulse. The resting synaptic region of a postsynaptic fibre shows miniature potentials, which have properties similar to postsynaptic potentials resulting from stimulation of presynaptic fibres. The miniature potentials, which occur randomly, are not the effects of single transmitter molecules; rather, they occur as a result of leakage of packets of transmitter molecules from the resting presynaptic nerve fibre. Each packet contains thousands of transmitter molecules. The packets probably represent the contents of the small sacs, or vesicles, known to occur at the ends of presynaptic nerve fibres. The frequency of miniature potentials increases when the presynaptic fibre is depolarized. The concentration of ions such as calcium and magnesium at the junction influences the release of packets of transmitter substance from depolarized nerve terminals. Calcium is necessary for their release; magnesium has an opposing action.

Inhibitory postsynaptic potentials. The stimulation of certain presynaptic fibres causes an increase in the threshold of excitation of postsynaptic cells rather than their excitation; synapses and neuromuscular junctions at which this occurs are said to be inhibitory. Changes in membrane potentials, known as inhibitory postsynaptic po-

tentials (IPSP's), which follow stimulation of a presynaptic fibre, hyperpolarize (*i.e.*, increase the resting potential of) the postsynaptic cell membrane, thus making more difficult the stimulation of an action potential. During an IPSP, the postsynaptic membrane often becomes selectively permeable to negatively charged particles (anions), and chloride ions enter the cell, their negative charges causing the increase in membrane potential.

The transmitter substance at inhibitory neuromuscular junctions is probably, in crustaceans, a compound called gamma-amino-butyric acid. In other animals, it has not been identified with certainty. It is likely that acetylcholine is an excitatory transmitter at certain synapses in vertebrates and mollusks and an inhibitory transmitter at other molluscan synapses. Glutamic acid may be the excitatory agent at some crustacean neuromuscular junctions; acetylcholine is important at many parasympathetic nerve endings, noradrenaline or adrenaline at sympathetic ones.

Electrical transmission. At certain synapses, where there is electrical, rather than chemical, transmission of the impulse, no synaptic delay occurs, and vesicles may not be present in presynaptic fibres; in addition, presynaptic and postsynaptic cell membranes generally are close to each other or fused. Electrical transmission of impulses occurs in the giant nerve fibres of earthworms, and, unlike most synapses, the junctions transmit impulses in either direction. The ciliary ganglion of the chick also permits action currents to spread in both directions. An excitatory synapse in the abdominal nerve cord of the crayfish depends upon electrical transmission but operates only in one direction; the synaptic contact allows currents to flow from the presynaptic to the postsynaptic fibre but not in the reverse direction. An example of electrical transmission in an inhibitory synapse is the Mauthner neuron of the goldfish, in which action currents, derived from adjacent nerve cells, are channelled into a specialized region, the axon cap, of the Mauthner cell and leave the cell in distant, nonsynaptic regions. The entering current hyperpolarizes the synaptic zone, raising the local threshold to excitation.

STUDY OF THE NERVE IMPULSE

The voltage-clamp technique. Applied by the Nobel Prize winners A.L. Hodgkin and A.F. Huxley, the technique known as the voltage clamp is used to measure currents flowing across a nerve-fibre membrane relative to the voltage across the membrane. First used on the giant nerve fibres of squid, the technique involves the insertion of two long wire electrodes along the fibre, so that a uniform potential may be maintained across a substantial length of the fibre membrane. One electrode is used to measure the voltage inside the nerve fibre, the second to introduce electrical current that can flow across the fibre membrane. When the membrane potential is changed to a new steady value by the flow of current across the fibre, the current crossing a specific membrane region is measured with two appropriately placed electrodes outside the fibre. This procedure was used to establish the time course and strength of currents flowing through the nerve membrane at known voltages and to characterize the ionic flow across the membrane.

Phases of membrane current

If a nerve fibre is suddenly depolarized (*i.e.*, resting potential decreased) by more than 20 millivolts, three phases of current may be measured. The first phase is a very brief surge of outward-flowing current, the result of the discharge of the stored charge in the membrane; this is called the displacement current. A second and transient phase follows, in which current flows inward through the cell membrane if the depolarization does not exceed 110 millivolts; otherwise, the transient current flows outward. After this transient current, a third phase of persistent outward current flow develops. The transient current in the second phase, which, over a large range of depolarization, flows in a direction contrary to the voltage gradient, is caused by sodium ions moving into the cell. The persistent current evoked by depolarization in phase three arises principally from the movement of potassium ions outward from the fibre. The phenomena that occur during voltage-clamp experiments with squid nerves also occur in other nerve fibres.

Applications

A set of equations for membrane conductances and ionic equilibrium potentials has been developed to describe the electrical phenomena in nerves. Although derived from observations made under the restricted conditions of the voltage clamp, these equations account for both the form and the velocity of the normal action potential in the squid nerve. In addition, they anticipate the positive phase following the action potential; they predict a well-defined threshold depolarization for excitation and the all-or-none response; and they account for the movements of both sodium and potassium ions that accompany each nerve impulse. The rate constants for changes in the ability of membranes to allow the passage of ions are temperature dependent, increasing threefold for a 10° increase in temperature; this property accounts for the marked temperature dependence of the action potential, which is of shorter duration as temperatures increase. Other phenomena that can be described from the equations relating to the voltage clamp include refractory period, anode break excitation, and subthreshold responses. The equations do not account for afterpotentials, apart from the positive phase seen in squid nerve fibres, or for phenomena associated with the sodium pump.

The voltage-clamp technique is valuable for providing information about many factors that affect nerve conduction. Calcium and magnesium ions, for example, affect the electrical properties of nerve fibres. Removal of calcium ions from the environment of a nerve fibre often leads to spontaneous activity and the voltage clamp indicates that the effects of calcium-ion deficiency are equivalent to a reduction of the resting potential. Magnesium ions have a calcium-like action but are less effective. One hypothesis of the effect of calcium ions is that they adhere to the outer surface of the cell, producing local potential gradients in the membrane, but they may have a more important role in nerve conduction. Calcium ions enter excited fibres faster than resting ones, and action potentials have been reported in axons placed in a sodium-free but calcium-rich environment. Voltage-clamp techniques show that local anesthetics, which reversibly block the conduction of nerve impulses, reduce the ability of the nerve membrane to conduct ionic currents. Of particular interest is tetrodotoxin, a powerful poison extracted from the puffer fish; in minute concentrations, this poison blocks nerve conduction by lowering the conductance of the membrane for sodium currents. Tetrodotoxin, reversible in its action on isolated nerve fibres, has no effect when applied inside a fibre. The separate characters of sodium and potassium current suggests that there may be zones in which each current is conducted.

Artificial and simplified membranes. The successful description of the nerve impulse in terms of the ionic theory has led to attempts to explain the chemical basis for the action potential in terms of structural models. One possibility involves a class of complex lipid molecules, called phospholipids, that are found in membranes. It has been postulated that these molecules, which contain electric charges, might change their positions in the membrane as a result of the electrical changes occurring during an action potential; in this way, they could provide a route through the membrane for sodium ions during depolarization (resting potential decrease) that later changes to a path allowing the entry of potassium ions. Thin membranes, approximately two phospholipid molecules in width, have been synthesized in attempts to construct membranes similar in structure to natural ones. Artificial membranes, or bilayers, have high electrical resistance and do not alone give action potentials; however, their resistance can be lowered by the addition of various small molecules. Of interest is a specific type of antibiotic, of which valinomycin is an example, that increases the permeability of both bilayers and cell membranes to potassium rather than sodium ions. Other compounds also cause bilayers and cell membranes to undergo phenomena reminiscent of those found in nerve fibres.

Possible role of phospholipids

Another method used to study the nerve impulse involves removal of the cytoplasm from squid giant nerve fibres; the cytoplasm is squeezed out of the fibre as toothpaste might be squeezed from a tube, and the fibre, including the membrane, is then filled with an artificial solution. If a potassium salt solution is used, the nerve fibres can conduct many impulses over several hours (Figure 4D); the bulk of the cytoplasm, therefore, apparently has little significance in the conduction of the nerve impulse, although the nature of the salt solution within the fibre determines whether or not the fibre is excitable. The action and resting potentials are abolished if identical solutions are used both inside and outside the fibre.

BIBLIOGRAPHY. E.D. ADRIAN, *The Basis of Sensation: The Action of the Sense Organs* (1928), *The Mechanism of Nervous Action* (1932), *The Physical Background of Perception* (1947), three monographs by a great authority, valuable for those interested in the role of the nerve impulse; J. BERNSTEIN, *Elektrobiologie* (1912), a classic of historical importance (in German); K.S. COLE, *Membranes, Ions, and Impulses: A Chapter of Classical Biophysics* (1968), an important monograph, individualistic in style, mathematical in approach; J.C. ECCLES, *The Physiology of Nerve Cells* (1957), *The Physiology of Synapses* (1964), authoritative monographs; J. ERLANGER and H.S. GASSER, *Electrical Signs of Nervous Activity* (1937), and A.V. HILL, *Chemical Wave Transmission in Nerve* (1932), two works of considerable historic interest; A.L. HODGKIN, *The Conduction of the Nervous Impulse* (1964), a brief but lucid exposition by a master of the field; H. HYDEN, "The Neuron," in J. BRACHET and A.E. MIRSKY (eds.), *The Cell*, vol. 4, pp. 215–323 (1960), a textbook account of structure and chemistry; G. KATO, *The Microphysiology of Nerve*, new ed. (1934), an account of useful experiments by delicate techniques, now mainly of historical interest; B. KATZ, *Electric Excitation of Nerve* (1939), *Nerve, Muscle, and Synapse* (1966), *The Release of Neural Transmitter Substances* (1969), brief and lucid monographs by a foremost authority; K. LUCAS, *The Conduction of the Nervous Impulse*, rev. by E.D. ADRIAN (1917), a book of historic interest; I. TASAKI, *Nervous Transmission* (1953), a useful monograph by an outstanding experimentalist, "Conduction of the Nerve Impulse," ch. 3 in *Handbook of Physiology*, sec. 1, *Neurophysiology*, vol. 1. (1959), a valuable textbook account; J.W. WOODBURY, "The Cell Membrane: Ionic and Potential Gradients and Active Transport," and "Action Potential: Properties of Excitable Membranes," in T.C. RUCH *et al.*, *Neurophysiology*, 2nd ed., ch. 1–2 (1965), a useful textbook.

(T.I.S.)

Nerves and Nervous Systems

All living organisms are able to react to external and internal changes in their environments. Some nonliving things can also react to changes; mercury in a thermometer expands as the temperature rises, and a photographic emulsion becomes chemically different when exposed to light. But the reactions of living forms differ from these relatively mechanical responses in that they tend to preserve the integrity, well-being, or status quo of the living organism. An animal, for example, tends to avoid excessive heat, which is harmful to it. A plant bends toward the light, which is necessary for its photosynthetic processes. In such manner do the reactions of living organisms increase the chances of survival of both the individual and the species. Dealt with in this article are coordinative activities based on the responsiveness of cells generally and on the reactions of nerve cells in particular. The integrative system of plants is dealt with generally, by way of comparison; details of plant coordination are covered in the article STEREOTYPED RESPONSES. For the entire complement of man's nervous coordination see the article NERVOUS SYSTEM, HUMAN.

This article is divided into the following sections:

I. General features
II. Nervous coordination
  Intracellular coordination
  Organelle systems
  Nervous systems
    The neuron, or nerve cell
    The transmission of the nerve impulse
III. Invertebrate nervous systems
  The earliest nervous systems
  Diffuse nervous systems
    Nerve net
    Simple radial systems
  Centralized nervous systems
    Complex radial systems
    Simple bilateral systems
    Moderately cephalized systems
    Complexly compartmentalized systems
IV. The vertebrate nervous system
  The central nervous system
    The brain and its components
    The spinal cord
    The brain coverings, cavities, and cerebrospinal fluid
    Neuroglia: non-nervous tissue
  The peripheral nervous system
    Cranial nerves
    Spinal nerves
    Sensory and motor fibres
    The autonomic nervous system
  Embryonic development of the vertebrate nervous system
    Development of the spinal cord and its nerves
    Development of the brain and its nerves
  Evolution of the vertebrate nervous system
    The primitive condition
    Encephalization
    Dominance of the cerebrum
V. Biodynamics of the vertebrate nervous system
  Controls exerted on the nervous system
    Intrinsic control
    Extrinsic influences
  Interaction between nervous and endocrine systems
    The neurosecretory cell and the concept of neurosecretion
    Neuroendocrine interactions

I. General features

The simplest type of living response is the direct one-to-one stimulus-response reaction. The change in the environment is the stimulus; the reaction of the organism to it is the response. In the case of single-celled organisms the response is the result of a property of the cell fluid (cytoplasm) called irritability. In protistan organisms, such as algae, protozoans, and fungi, simple responses in which the organism moves toward or away from the stimulus are termed taxes (singular: taxis). In larger and more complicated organisms, those in which response involves the synchronization and integration of events in different parts of the body, a control mechanism, or "controller," is located between the stimulus and the response. In multicellular organisms, the controller consists of two basic mechanisms by which integration is achieved—chemical regulation and nervous regulation.

Integrative mechanisms

In chemical regulation, substances called hormones are produced by definite groups of cells and are either diffused or carried by the blood to other areas of the body where they act on so-called target cells, influencing metabolism or inducing synthesis of other substances. The changes resulting from hormonal action are expressed in the organism as influences on, or alterations in, form, growth, reproduction, and behaviour patterns.

Plants respond to a variety of external stimuli by utilizing hormones as controllers in a stimulus–response system. Directional responses of movement are known as tropisms and are called positive when the movement is toward the stimulus and negative when it is away from the stimulus. When a seed germinates, the growing stem turns upward toward the light; the roots turn downward away from the light. Thus, the stem shows positive phototropism and negative geotropism, while the roots show negative phototropism and positive geotropism. In this example, light and gravity are the stimuli, and directional growth, which results in movement, is the response. The controllers are certain hormones synthesized by cells in the tips of the plant stems. These hormones, known as auxins, diffuse through the tissues beneath the tip and concentrate toward the shade side, causing elongation of these cells and, thus, a bending of the tip toward the light. The end result is that the growing part of the plant is maintained in an optimal condition with respect to light.

In animals, in addition to chemical regulation via the endocrine system, there is another integrative system called the nervous system. A nervous system can be de-

fined as an organized group of cells specialized for the conduction of an impulse—an excited state—from a sensory receptor through a nerve network to an effector, the site at which the response occurs. The controller in this type of stimulus-response system is a specialized cell called a neuron. In contrast to the endocrine system, the actions of the nervous system are generally extremely rapid, short acting, and precise, involving point-to-point transmission between cells.

Conduction of impulses

Organisms that possess a nervous system are, therefore, capable of much more complex behaviour than are organisms that lack one. The nervous system, being specialized for the conduction of impulses, allows rapid responses to stimuli from the environment. Many of the responses mediated by the nervous system are directed toward preserving the status quo, or homeostasis, of the animal. Stimuli that tend to displace or disrupt some part of the organism call forth a response that results in reduction of the adverse effects and a return to a more normal condition. Organisms with a nervous system are also capable of a second group of functions that initiate a variety of behaviour patterns. Animals, thereby, may go through periods of exploratory or appetitive behaviour, nest building, and migration. Although these activities are beneficial to the survival of the species, they are not always performed by the individual in response to a particular need or stimulus. Finally, learned behaviour can be superimposed on both the homeostatic and initiating functions of the nervous system.

II. Nervous coordination

INTRACELLULAR COORDINATION

All living cells have the property of irritability, or responsiveness to environmental stimuli, which can affect the cell in different ways, producing, for example, electrical, chemical, or mechanical changes. These changes are expressed as a response, which may be the release of secretory products by gland cells, the contraction of muscle cells, the bending of a plant-stem cell, and the beating of whiplike "hairs," or cilia, by ciliated cells.

The responsiveness of a single cell can be illustrated by the behaviour of the relatively simple protozoan amoeba, a single-celled organism.

Responsiveness of amoebas

Unlike some other protozoans, an amoeba lacks highly developed structures that function in the reception of stimuli and the production or conduction of the response. The amoeba behaves as though it had a nervous system, however, because the general responsiveness of its cytoplasm serves the functions of a nervous system. An excitation produced by a stimulus is conducted to other parts of the cell and evokes a response by the animal. An amoeba will move to a region of a certain level of light. It will be attracted by chemicals given off by foods and exhibit a feeding response. It will also withdraw from a region with noxious chemicals and show an avoidance reaction upon contacting other objects.

ORGANELLE SYSTEMS

In more complex protozoans, however, the specialized cellular structures, or organelles, that serve these functions include receptors such as stiff sensory bristles in ciliates and the light-sensitive eyespots of flagellates. Effectors that enable a protozoan to respond to stimuli include cilia (slender hairlike projections from the cell surface), flagella (greatly elongated whiplike cilia), and other organelles associated with drawing in food or with locomotion. Protozoa also have subcellular cytoplasmic filaments that, like muscle tissue, are contractile. The vigorous contraction of the stalked protozoan *Vorticella*, for example, is the result of contraction of a threadlike structure (myoneme) in the stalk.

Although Protozoa clearly have specialized receptors and effectors, it is not certain that there are special conducting systems between the two. In a ciliate such as *Paramecium*, the beating of the cilia—which propels it along—is not random, but coordinated. Beating of the cilia begins at one end of the animal and moves in regularly spaced waves to the other end, suggesting that coordinating influences are conducted longitudinally. A system of fibrils connecting the bodies in which the cilia are rooted could provide conducting paths for the waves, but coordination of the cilia could also take place without such a system. Each cilium could respond to a stimulus carried over the cell surface from an adjacent cilium, in which case, coordination would be the result of a chain reaction from cilium to cilium.

The best evidence that formed structures are responsible for coordination comes from another ciliate, *Euplotes*, which has a specialized band of ciliary rows (membranelles) and widely separated tufts of cilia (cirri). By means of the coordinated action of these structures, *Euplotes* is capable of several complicated movements in addition to swimming (*e.g.*, sharp turns, moving backward, spinning). The five cirri at the rear are connected to the anterior end in an area known as the motorium. The fibres of the motorium apparently provide coordination between the cirri and the membranelles. The membranelles, cirri, and motorium constitute a neuromotor system.

NERVOUS SYSTEMS

As has already been indicated, the basic pattern of nervous coordination in animals is the organization of the stimulus–response system into receptor, adjustor (or controller), and effector units. External stimuli are received by the receptor cells, which, in most cases, are neurons. (In a few instances, a receptor is a nonnervous sensory epithelial cell, such as a hair cell of the inner ear or a taste cell, which stimulates adjacent neurons.) The stimulus is modified, or transduced, into an electrical impulse in the receptor cell. This incoming excitation, or afferent impulse, then passes along an extension, or axon, of the receptor to an adjustor, or association neuron. All neurons are capable of conducting an impulse, which is a brief change in the electrical charge on the cell membrane. Such an impulse can be transmitted, without loss in strength, many times along an axon until the message, or input, reaches another neuron, which in turn is excited. The adjustor neuron selects, interprets, or modifies the input from the receptor and sends off an outgoing, or efferent, impulse to an efferent, or motor, neuron. The efferent neuron, in turn, makes contact with an effector such as a muscle or gland, which produces a response.

In the simplest arrangement, the receptor–adjustor–effector units form a functional group known as the reflex arc. Sensory cells carry afferent impulses to a central adjustor or internuncial neuron, which makes contact with a motor neuron. The motor neuron carries efferent impulses to the effector, which produces the response. Three types of neurons are involved in this reflex arc; a two-neuron arc, in which the receptor makes contact directly with the motor neuron, is possible but probably not of widespread occurrence. Simple reflexes are prompt, short-lived, automatic, and involve only a part of the body. Examples of simple reflexes are the contraction of a muscle in response to stretch, blinking of the eye when the cornea is touched, and salivation at the sight of food. Reflexes of this type are usually involved in preserving homeostasis.

Adjustor neurons of complex nervous systems

The differences between simple and complex nervous systems lie not in the basic units but in their arrangement. In higher nervous systems, there are more adjustor neurons concentrated in the central nervous system (brain and spinal cord) that mediate the impulses between afferent and efferent neurons. Sensory impulses from particular receptors travel over specific pathways to the central nervous system. Within the central nervous system, though, the impulse can travel over multiple pathways formed by numerous adjustor neurons. Theoretically, the impulse can be distributed to any of the efferent motor neurons and produce a response in any of the effectors. It is also possible for many kinds of stimuli to produce the same response. Man, for example, walks in response to a variety of stimuli. The connections of the adjustor neurons in the central nervous system make possible all the special pathways between the receptive areas and the final common pathways.

As a result of the integrative action of the adjustors, the behaviour of the organism is more than the simple sum of its reflexes; it is a unitary, integrated whole that exhibits coordination between many individual reflexes. Reflexes can occur in a complicated series or sequence producing elaborate behaviour patterns. Behaviour in such cases is characterized not by inherited, stereotyped responses but by flexibility and adaptibility to circumstances. Many automatic, unconditioned reflexes can be modified by or adapted to new stimuli. The remarkable experiments of the Russian physiologist Pavlov, for example, showed that if an animal salivates at the sight of food while another stimulus, such as the sound of a bell, occurs simultaneously, the sound alone can induce salivation after several trials. This response, known as a conditioned reflex, is a form of learning. The animal's behaviour is no longer limited by fixed, inherited reflex arcs but can be modified by experience and exposure to an unlimited number of stimuli. The most evolved nervous systems are capable of even higher associative functions: of other forms of learning, of thinking, and of memory. The complex manipulation of the signals necessary for these functions depends to a great extent on the number and intricacy of the arrangement of the adjustors.

The neuron, or nerve cell. In multicellular organisms, the diffuse general irritability of single cells does not permit adequate coordination between widely separated regions of the body. Instead, specialized cells have evolved to channel and focus that irritability. These cells —known as neurons—are specialized for the reception of stimuli, conduction of excitation, and transmission of the signal to other cells.

A nerve cell consists of a cell body and cell processes (Figure 1). There is a great variety of neuron types, depending on size, shape, number of processes, and function. The cell body is composed of a more or less centralized nucleus and the surrounding cytoplasm. The nerve cell processes are of two types: short, multiple processes are dendrites and a single, long process is an axon. Neurons with one process are called unipolar and are especially common in invertebrates. If the process originating from the cell body gives rise to two branches running in opposite directions, the cell is said to be pseudo-unipolar. Bipolar cells have two separate processes. Some vertebrate sensory cells are bipolar, such as neurons in the retina of the eye and in the inner ear. Multipolar cells have more than two processes—a single axon and two or more dendrites. Multipolar neurons, which are the predominant type in the vertebrate central nervous system, can be further subdivided into two types: neurons with a large cell body and an extremely long axon—up to three feet—and neurons with a small cell body and a short axon.

Functionally, nerve cells can be classified as sensory, associative (internuncial), or motor. A sensory neuron is a nerve cell specialized for the reception of stimuli. A nerve cell that carries impulses to the effector organs is a motor nerve cell, or motoneuron. The internuncial nerve cell, or interneuron, is an adjustor or integrator; it receives the excitations arising from a receptor cell but distributes, inhibits, combines, and modifies this information before passing it on to a motor neuron.

Functional classification of neurons

From *Biological Sciences Curriculum Study, Biological Science: An Inquiry into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich

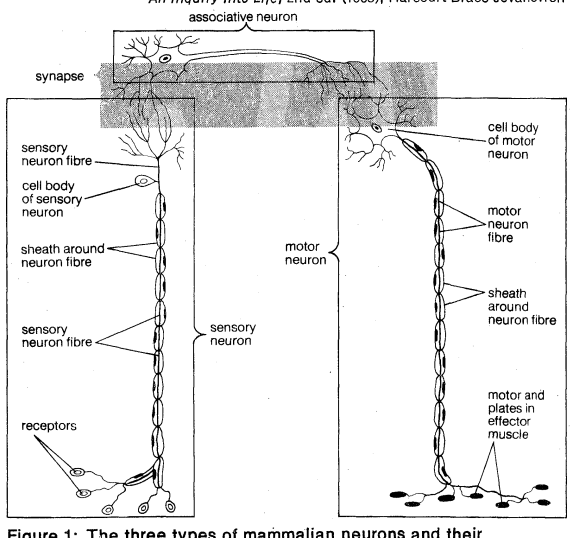


Figure 1: The three types of mammalian neurons and their relationships in a reflex arc (see text).

The cell body. Despite the great varieties of neurons, there are certain similarities in structure among the different types. The nucleus is a large, round body centrally situated in the cytoplasm and enclosed in a nuclear envelope. The nucleus contains a smaller body known as the nucleolus and dispersed material called chromatin. In addition to the usual cell components, such as mitochondria and the Golgi apparatus, the cytoplasm of the nerve cell contains special organelles, such as the Nissl bodies and neurofibrils, all of which are evident in nerve cells when they are stained with the appropriate chemicals and viewed under an ordinary optical microscope. Viewed with the greater magnifying power of the electron microscope, however, the Nissl bodies are a delicate channel-like system—the endoplasmic reticulum—which pervades the cell with flattened membranous cisterns and attached particles called ribosomes. Two types of delicate threads in the cytoplasm extend into the nerve processes—tiny neurofilaments and larger tubules known as neurotubules, or microtubules.

Large neurons often contain large bodies of golden-brown pigment. These structures, known as lipofuscin pigment granules, seem to be a by-product of nerve-cell metabolism; they become more abundant as the cell ages. Brown or black melanin granules are also sometimes present, as are lysosomes (granular bodies containing certain enzymes) and varying amounts of nutrients such as lipids and glycogen.

The nerve processes. Each neuron gives rise to a single axon, which arises from a conical elevation known as the axon hillock. The hillock and the axon cytoplasm, or axoplasm, can be distinguished from other regions of the cytoplasm by the lack of Nissl bodies. The axon contains mitochondria, neurofilaments, microtubules, and elongated channels of smooth endoplasmic reticulum. Along its course, the axon may give off collateral branches, which divide into even smaller branches. The main branch forms an extensive terminal brush, the telodendrion, which forms either a connection (synapse) with another neuron or with an effector such as a muscle fibre or a gland cell.

The nerve cell gives rise to one or, more commonly, several dendrites. The dendrites, usually short and confined to the vicinity of the cell body, branch repeatedly and sometimes form elaborate "trees," which increase enormously the receptive surface of the neuron. The farthest (distal) branches contain tiny spines, thorns, or knoblike gemmules upon which axons of other neurons converge. In internal structure, dendrites represent extensions of the neuronal cytoplasm. Unlike axons, they contain Nissl bodies.

The dendrites receive nerve impulses from the axons of other neurons terminating on them. In this way a single neuron may receive messages from a great many neurons. It has been estimated that there may be 200,000 axon terminals on the elaborate dendritic tree of a single Purkinje cell, a flask-shaped cell in the cerebellum of the brain. Thus, the dendritic tree plays a major role in integrating information received from different sources. The impulses travelling over the axons to the dendrites excite or inhibit electrical activity in localized regions of the dendrite membrane. The graded changes in electrical potential in the dendrite raise or lower the ability of the neuron to generate a nerve impulse. The impulse arises in the initial segment of the axon, is conducted to the axon terminal, and is then transmitted to other neurons or to effector cells.

The dendrite tree

The nerve cell body comprises only a small portion of the total cell mass, the major portion being the processes. The cell body maintains the cytoplasm of the processes by means of axoplasmic flow, which proceeds at a rate of

about one millimetre (0.04 inch) per day. Some structures within the axoplasm, however, such as vesicles, move at a greater rate.

The nerve fibre. The nerve fibre consists of the axon and covering or sheath cells (Figure 2). In the peripheral nerves, which extend into appendages, the sheath is

Thomas L. Lentz

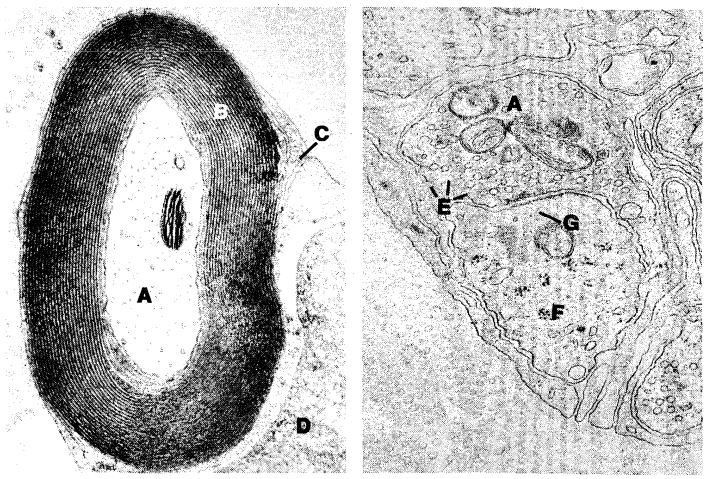


Figure 2: (Left) Electron micrograph of myelin sheath around nerve axon showing (A) nerve axon, (B) myelin sheath, (C) Schwann cell membrane, and (D) Schwann cell nucleus. (Right) Typical synapse in which (A) the axon terminal, containing many (E) synaptic vesicles, is separated from (F) the dendrite by (G) a synaptic cleft. (Photographs magnified about 28,000×.)

formed by so-called Schwann cells; in nerves of the central nervous system, the covering is formed by cells called oligodendroglia. In addition, some axons are enclosed by a fatty (lipoidal) covering, the myelin sheath, formed by Schwann cells or oligodendroglia. Nerves with a myelin sheath are called myelinated nerves; those lacking a sheath are unmyelinated. Myelin is responsible for the white colour of the nerves, brain, and spinal cord.

Interruptions of myelin at intervals along a nerve fibre are known as the nodes of Ranvier. These nodes represent points of discontinuity between individual Schwann cells arranged along the nerve fibre. In these regions, the axon is covered only by fingerlike extensions of the Schwann cells. Oblique lines running across the nerve fibre within Schwann cell segments, known as incisures of Schmidt-Lanterman, are formed by separations of myelin layers.

Each nerve fibre—myelinated or unmyelinated—is enclosed in a delicate connective-tissue covering called the endoneurium. Bundles of nerve fibres are surrounded by a more extensive layer of connective tissue, the perineurium. Finally, the entire peripheral nerve trunk is encapsulated by another connective sheath, the epineurium.

In nerve conduction, the myelin sheath acts as an insulator that prevents the impulse from jumping to adjacent axons. The nerve impulse is a wave of excitation, called the action potential, that is formed by an electrical change, or depolarization, of the membrane of the axon and passes down the nerve fibre. In unmyelinated nerves, conduction is a smooth and continuous process. In myelinated nerves, however, the impulse travels by means of jumps (saltatory conduction) between the nodes of Ranvier, a more rapid process than the continuous conduction in unmyelinated nerves.

Speed of conduction in the nerve fibre depends on the size of the axon and the thickness of the myelin sheath. Nerve fibres can be grouped into three main classes: A, B, and C fibres. The A fibres are myelinated, one to 20 microns (one $\mu = 0.001$ millimetre) in diameter, and are the fastest, conducting impulses at a rate of five to 120 metres (16 to 400 feet) per second; B fibres have a thinner myelin sheath, a diameter up to three microns, and conduct impulses at a rate of three to 15 metres (ten to 50 feet) per second; C fibres are the smallest, often less than one micron across, are unmyelinated, and conduct impulses slowly, at a rate of 0.6 to two metres (two to seven feet) per second.

Degeneration and regeneration of neurons. A nerve cell body lost because of injury or disease cannot be replaced. If, however, a nerve fibre is injured, cut, or crushed, the nerve cell body undergoes a series of changes preparatory to replacing the fibre. Beyond the point of injury, the axon and its myelin sheath die after a succession of changes called Wallerian degeneration. The neurilemma, however, persists as a hollow tube. The cell body and fibre closest (proximal) to the injury undergo changes called retrograde degeneration. The axon and myelin degenerate up to the first node of Ranvier, while in the cell body, the Nissl substance disperses. If the nerve fibre is injured very near the cell body, the neuron may die; if the cell body survives, however, regeneration of the peripheral fibre begins. (Although axons in the central nervous system attempt to regenerate, they seem unable to make connections across a wound gap.)

Tiny sprouts appear at the proximal end of the nerve fibre. The dilated ends of these sprouts are known as growth cones or end bulbs. Advancing sprouts, which enter the wound along with Schwann cells and connective-tissue cells called fibroblasts, then invade the persisting neurilemmal tube of the distal stump and grow down this pathway. Growth occurs at a rate of two to four millimetres (0.08 to 0.16 inch) per day; thus, a long nerve may require months to regenerate completely. Once it reaches its termination, it connects with the end organ, and normal function is restored. During this process, the axon increases in diameter, and the myelin sheath is reformed.

Rate of nerve regeneration

The transmission of the nerve impulse. In simple reflex arcs, as well as in complicated multineuron pathways, impulses must be transmitted from nerve cell to nerve cell over minute gaps called synapses. A synapse is formed by the termination of an axon on the dendrites (axodendritic synapse), the cell body (axosomatic synapse), or the axon (axoaxonic synapse) of an adjacent nerve cell, or postsynaptic cell. Some neurons have only a few synapses, but others, like the Purkinje cell, may have hundreds of thousands. Each synapse represents a connection with another neuron that has information to convey. Each neuron represents a centre in which all this information is integrated and sent along as a nerve impulse to still other neurons.

Near the termination of the axon, the myelin sheath, if present, ends, and the axon ends in a dilated bulb. The bulb usually contains a few mitochondria and many small sacs called synaptic vesicles (Figure 2). At the synapse the axon ending is separated from the postsynaptic neuron by a microscopic space called the synaptic cleft. Each neuron, therefore, is a discrete cellular unit anatomically separate from all other neurons.

Transmission of the nerve impulse can be either chemical or electrical, but chemical synapses are by far the more numerous. In chemical transmission, a specific chemical is synthesized by the nerve cell, stored at the axon terminal, and released at the appropriate time, affecting the postsynaptic cell. These chemical neurotransmitters are stored in the synaptic vesicles. Neurons in which the transmitter is the chemical acetylcholine are said to be cholinergic; neurons utilizing norepinephrine as the transmitter are known as adrenergic.

Upon arrival of an impulse at the nerve ending, the neurotransmitter is released from the synaptic vesicles. The transmitter diffuses across the synaptic cleft and produces a change in permeability of the postsynaptic membrane. The change results in an electrical change in the postsynaptic cell. The effect on the postsynaptic neuron may be excitatory or inhibitory. After the neurotransmitter is liberated, it is destroyed by enzymes so that its effect is localized and of short duration.

In electrical synapses, the arriving nerve impulse passes directly from the presynaptic to the postsynaptic element, whose membranes are very closely apposed, with only the tiniest gap separating them. This close apposition probably accounts for the low resistance between cells at these points, which allows the impulse to travel directly from cell to cell. In both chemical and electrical synapses, transmission is polarized; that is, it can pass in only one

direction across the synapse. (See the article NERVE IMPULSE for the biochemistry and biophysics of nervous conduction.)

III. Invertebrate nervous systems

THE EARLIEST NERVOUS SYSTEMS

Interest in the evolutionary origin of the nervous system has arisen primarily from the need to reduce the extremely complex nervous system of man and other higher animals to its simplest characteristics. Knowledge of the nature of the earliest nervous system might reveal the basic underlying mechanisms of nervous function in all nervous systems.

Among several theories about the manner in which the nerve cell arose, one of the earliest proposed that all the components of the nervous arc—the receptor, conductor, and effector—were originally contained in a single cell type. This cell, called a neuromuscular cell, then gave rise to separate nerve and muscle cells. This theory, however, has not been widely accepted because true neuromuscular cells have not been identified with certainty in any existing animal.

A more widely accepted view of the origin of the nervous system suggests that effectors evolved first. Sponges, for example, are considered to represent the state in evolution when muscle, an effector, is present in the absence of nerve cells. The next stage was the appearance of receptors, or sensory cells, which terminated on muscle cells. This stage constitutes a receptor–effector system. Finally, another nerve cell—the adjustor, or central organ—appeared between the receptor and effector, forming the receptor–adjustor–effector system that is found in higher invertebrates and vertebrates.

The origin of nerve cells

Nerve cells are thought to have arisen from the epidermis, or body covering, because they occur in this layer in some of the simple invertebrates and because nervous tissue develops from the outer layer (ectoderm) in the embryo. The cell that gave rise to neurons could have been excitable and conductile to the extent that many non-nervous cells and protozoans are. Being epithelial, this pre-nervous cell may also have been capable of secretory activity. Being excitable, the cell could respond to environmental stimuli. By the appearance of long cytoplasmic processes, the conductive properties of the cell could have become better developed. Finally, if the secretory capabilities of the cell were directed to the production of active substances that, when released, affected other cells, such as muscle, a means of transmission would be available. This primitive nerve cell, then, could have fulfilled the criteria necessary in any nerve cell —excitability, conductivity, and transmissibility.

The nerve cell probably appeared first in some primitive multicellular animal that was ancestral to more complicated invertebrates. In view of the functions performed by a nervous system, it is apparent that an organism possessing such a system would have a considerable advantage over those lacking one. The organism would be capable of more complex behaviour, expressed perhaps in better abilities to feed or to detect and escape from its enemies.

DIFFUSE NERVOUS SYSTEMS

Among the invertebrates in general, the least complex nervous systems are found in forms that are small, relatively simple in structure or uncomplicated in behaviour. Of the two general types of nervous systems—diffuse and centralized—the diffuse system is the original and most primitive type. It is found in the coelenterates (*e.g.*, hydroids, jellyfishes, sea anemones, corals) and in the ctenophores, or comb jellies. The even more primitive sponges have scattered cells with neural characteristics, but it is not certain whether these cells constitute a true nervous system. In the diffuse systems, nerve cells are distributed throughout the organism, usually beneath the outer epidermal layer. Large concentrations of nerve cells (*e.g.*, a brain) are not found in these animals, but there may be small local concentrations called ganglia.

Nerve net. The common hydroid *Hydra* has a nervous system that consists entirely of a nerve net—a meshlike system of individual and separate nerve cells and fibres dispersed over the organism but more concentrated near the "mouth" (Figure 3A). This widespread arrangement of fibres allows diffuse conduction of nervous excitation over a large number of available pathways, so that, if a hydra is split longitudinally but left connected by a small piece of body wall, stimulation of one side will result in contraction of both sides. In most higher animals, by contrast, interruption of specific nerves blocks responses to stimuli, because there are no alternate conducting pathways.

Several cell types comprise the nervous system of *Hydra*. Most of the nerve cells are found in the outer epithelial layer. Ganglion cells, situated at the base of the epidermis, are spindle shaped and have long processes. Sensory cells are elongated and oriented perpendicular to the surface of the animal; the outward extensions such as cilia act as sensory receptors, while inward extensions join the network at the base of the epithelial cells. Some of the nerve cells of *Hydra* are capable of secretion and are known as neurosecretory cells. Neurosensory cells are deeply situated ganglion cells with cilia reaching the body surface.

Simple radial systems. The nervous system of coelenterates, which corresponds to the radially symmetrical bodies of these animals, is distributed over the surface as a nerve net. In sea anemones the neurons are more highly concentrated near the mouth. Although local concentration of neurons reaches its highest degree in the jellyfish, the concentration is not sufficient to be considered a central nervous system. The jellyfish of the class Hydrozoa have two circular thickenings of the nerve net near the margin of the bell. These thickenings constitute the nerve rings. From the nerve rings, other thickenings of the net extend inward toward the mouth. In jellyfish of the class Scyphozoa there are no marginal rings, but ganglion cells are concentrated into usually eight marginal ganglia. The marginal ganglia occur near marginal bodies, or rhopalia, which are complex sensory structures that may include sensory or olfactory pits; sensory tentacles; a statocyst, or equilibrium organ; and photoreceptor organs, or ocelli.

Nerve nets

The marginal bodies initiate the rhythmic pulsation by which the jellyfish swims. Rapid conduction over the nerve net excites all the musculature of the bell nearly simultaneously.

CENTRALIZED NERVOUS SYSTEMS

One of the most prominent trends in the evolution of the nervous system is the centralization of nervous elements. In echinoderms (*e.g.*, starfishes, sea cucumbers, sea lilies) the nervous system is sufficiently concentrated to be called a central nervous system, but there is no true brain. In the platyhelminthes, or flatworms, and most higher invertebrates, however, there is an anterior collection of nerve cells, or a brain, and a longitudinal nerve cord. In the simpler types, the central nervous system is not distinctly separated from the peripheral nervous system, whereas in more advanced types, the nervous system is increasingly concentrated in the cephalic, or head, end of the animal.

Complex radial systems. A complex radial nervous system is found in the echinoderms, which, like the coelenterates, have a radial type of body construction. In fact, the nervous system of echinoderms shares some primitive characteristics with that of coelenterates, but also shows some advanced features. The major components of the nervous system are a nerve ring around the mouth, radial nerves extending into the arms, and a nerve network, or plexus, beneath the body covering. The radial nerves and nerve ring constitute the central nervous system of echinoderms. These are actually just thickened regions of the subepidermal plexus that contain both nerve cells and fibres. Specialized sensory organs are not common in these animals, although many starfishes have eyespots and sensory tentacles at the tip of each arm. The junction of each radial nerve with the circumoral nerve ring acts as a centre for the coordinated movement of the arms.

Simple bilateral systems. The flatworms were the first invertebrates to exhibit bilateral symmetry and also the first to develop a central nervous system with a brain (Figure 3B). The nervous system of a free-living flatworm such as *Planaria* consists of a brain, longitudinal nerve cords, and peripheral nerve plexuses. Located in the anterior portion of the animal, the brain is composed of two cephalic ganglia joined by a broad connection called a commissure. Longitudinal nerve cords, usually three to five pairs, extend posteriorly from the brain; they are connected by transverse commissures, and smaller, lateral nerves extend from the cords. The lateral nerves give rise to the peripheral nerve plexuses. The submuscular nerve plexus, consisting of sensory cells, ganglion cells, and their processes, is situated in the loose tissue (mesenchyme) below the subepidermal musculature. A much finer subepidermal plexus is located at the bases of the epithelial cells above the muscular layer.

Sensory receptors in *Planaria*

Planaria are richly supplied with sensory receptors. Single sensory cells in the nerve plexuses are widely scattered over the animal. Sensory organs also are present and include ciliated pits and grooves, auricles, the frontal organ, statocyst, and eyes. The ciliated pits and grooves contain chemical receptors (chemoreceptors), which permit the animal to detect food. The statocyst is responsible for balancing and such reactions as rising to the surface of the water or sinking. The eyes, or ocelli, may occur as a pair situated anteriorly or may be scattered abundantly over the head region depending on the species. Short optic nerves connect the eyes with the brain.

Moderately cephalized systems. Basic similarities in the nervous systems of the annelid worms, mollusks, and arthropods include an anteriorly situated brain, connectives running from the brain around the esophagus and joining paired longitudinal cords, and ventral nerve cords with ganglia along their length. The trend toward greater centralization and cephalization of nervous functions is continued within these groups, reaching its peak in the higher mollusks and arthropods.

Annelids. The brain of most annelids, terrestrial earthworms and the leeches, is relatively simple in structure. The earthworm brain is a bilobed mass lying above the pharynx in the third body segment (Figure 3C). Sensory nerves leave the brain and run forward into the prostomium (extreme anterior end) and first segment. The brain of the active, predatory polychaetes is more complicated. In some, the brain can be divided into a forebrain, midbrain, and hindbrain; a single pair of circumesophageal or circumpharyngeal connectives leave the brain, surround the anterior gut, and connect with the ventral nerve cord.

The most primitive annelids have a pair of ventral nerve cords joined by transverse connectives; the more advanced forms have the cords fused to form a single cord. A ganglionic swelling of the cord is found in each body segment, with the most anterior ganglion, the subpharyngeal ganglion, being the most prominent. Two to five pairs of lateral nerves leave each ganglion to innervate the body wall of that segment. Separate ganglia supply the leglike parapodia of polychaetes. A subepidermal nerve plexus occurs over the whole body. Another plexus, called the enteric, stomodaeal, or sympathetic system, is found in the wall of the gut.

Giant axons, usually few in number, are found running the length of the cord. They may belong to one cell or be composed of many neurons. These axons are capable of vary rapid conduction of impulses to the segmental muscles; their main function is to permit the worm to contract very rapidly as a defense against predators.

The usual slow crawling movements of worms are mediated by a series of reflex arcs. During crawling, the contraction of muscles in one segment stimulates stretch receptors in the muscle. Impulses are carried over sensory nerves to the cord, causing motor neurons to send impulses to the longitudinal muscles, which then contract. The longitudinal pull activates stretch receptors in the following segment, and a wave of contraction moves along the worm.

Some of the annelids, especially the active polychaetes, are richly supplied with sensory organs. The most highly developed eyes consist of a retina, a vitreous body, a lens, and an outer cornea. Optic nerves leave the eye and enter the brain. Spiral organs are light sensitive regions on the surfaces of the antennae. Nuchal organs are ciliated sensory pits, or slits, located on the head; they contain chemoreceptors necessary for the detection of food. Other ciliated sense organs include the dorsal and lateral organs, which are found on the trunk segments. Statocysts, necessary for orientation of the worm in burrowing, are located in the head.

Simple mollusks. The nervous systems of the lower mollusks (snails, slugs, and bivalves such as clams and mussels) conform to the basic annelid plan but are modified to conform with the unusual anatomy of these animals. In snails, a pair of cerebral ganglia constitutes the brain, which overlies the esophagus. Nerves leave the brain anteriorly to supply the eyes, tentacles, and a pair of buccal ganglia. These last ganglia, also called the stomatogastric head ganglia, innervate the pharynx, salivary glands, and a plexus on the esophagus and stomach. Other nerve cords—the pedal cords—leave the cerebral ganglia ventrally and terminate in a pair of pedal ganglia, which innervate the foot muscles. Another pair of nerve cords—the visceral cords—leave the brain and run posteriorly to the visceral ganglia. The pleural ganglion, supplying the mantle, or fleshy lining of the shell, and the parietal ganglion, innervating the lateral body wall and mantle, are located along the visceral nerves. Intestinal ganglia connected with the pleural ganglia innervate the gills, osphradium (a chemical sense organ), and mantle. Because of the torsion that occurred during the evolutionary history of the snails, some components of the nervous system are twisted and located in asymmetrical positions. In addition, there has been some migration forward and fusion of ganglia. Sense organs of snails include eyes, tentacles, statocysts, and osphradia.

Modified nervous system of snails

In the bivalves, a cerebropleural ganglion is situated on either side of the esophagus. An upper pair of nerve cords leaves these ganglia and runs posteriorly to the visceroparietal (visceral) ganglia. The visceral ganglia supply the mantle, adductor muscles (which close the shell), and internal organs. A second pair of nerve cords runs ventrally to the pedal ganglia. Most of the sense organs are found at the edge of the mantle. In the scallop, for example, the eyes are set in a row. They are well developed and consist of a cornea, lens, and a retina, in which the photoreceptor cells are directed away from the source of light (an arrangement much like that in the vertebrate retina).

Complexly compartmentalized systems. *Complex mollusks.* The highest degree of development of the invertebrate nervous system is attained by the cephalopods (squids, cuttlefishes, and octopuses) among the mollusks and by the insects and spiders among the arthropods. Although the basic plan of these nervous systems is similar to that of the annelids, there are several advances. First, there is a high degree of cephalization, with nervous functions concentrated in the head region of the animal. In addition, ganglia are fused and further forward, and nerve cells, less abundant in the peripheral nervous system, are situated in the brain or ganglia so that the nerve cords consist only of nerve fibres. Finally, control and coordination of specific functions, such as locomotion and feeding, are compartmentalized in particular parts of the nervous system.

The complex nervous system of the cephalopods is correlated with the active movement and predatory habits of these animals. Most of the ganglia typical of mollusks are concentrated or fused in a brain that encircles the esophagus. Nerves extend from the brain to ganglia at the base of the arms or tentacles and from the ganglia the length of the arms. A pair of large pallial nerves connects the brain with a pair of stellate ganglia on the inner surface of the mantle. The stomatogastric ganglia supply nerves to the digestive tract.

A great variety of body functions are centralized in the brain and compartmentalized to specific brain regions. These activities may be local, simple, and uncoordinated with other regions or may be extensive, complex, and

coordinated, involving large groups of muscles. The highest centres of the cephalopod brain are the associative areas, concerned with discrimination between objects, learning, and memory.

Giant-fibre system of the squid

The giant-fibre system—also seen in earthworms and insects—is very well developed in the squid. The diameter of giant fibres is many times greater than the diameter of most other nerve fibres. Giant neurons in the brain send fibres to the retractor muscles of the head and the funnel or to the stellate ganglion. Fibres from the stellate ganglion fuse to form giant fibres that innervate the mantle. Because of their large size, these fibres are capable of rapid conduction that, in turn, permits extremely rapid movement.

The eyes of cephalopods are especially well developed and bear close structural resemblance to the vertebrate eye. The eye fits into a socket of cartilaginous plates separate from the cartilages that protect the brain, and external muscles permit its movement. A transparent cornea covers the surface of the eye and can be focussed for both near and far objects. There is a pupil formed by an iris diaphragm, which can regulate the amount of light reaching the retina. The retina contains light-sensitive cells that differ from those in the vertebrate eye in being directed toward the source of light. The axons of the photoreceptors, or rod cells, form the optic nerves, which terminate in the extremely large optic lobes of the brain.

Arthropods. The other complex compartmentalized nervous system is found in arthropods. The arthropodan brain consists of three main regions: the protocerebrum, deutocerebrum, and tritocerebrum (Figure 3D). The anterior protocerebrum, which receives the nerves of the eyes and other organs forward, contains centres or neuropils such as the optic centres and bodies known as corpora pedunculata. The neuropils function as integrative systems for the anterior sense organs, especially the eyes, and in control of movement; they also are the centres for the initiation of complex behaviour. The deutocerebrum contains the association centres for the first antennae. The posterior tritocerebrum contains association neuropils for the second antennae (of crustaceans) and gives rise to nerves that innervate the mouthparts and the anterior digestive canal. The latter comprise the stomatogastric system, which regulates the intake of food and the movement of the gut necessary for digestion. This system bears a resemblance to the vertebrate autonomic system (see below).

The ventral nerve cord, connected to the brain by the circumesophageal connectives, is basically composed of a double row of ganglia connected longitudinally by connectives and transversely by commissures. Different groups of arthropods exhibit different degrees of fusion of the ganglia. In insects, the first ganglion, the subesophageal, is formed by fusion of three pairs of ganglia; it sends nerves to the mouthparts and to the salivary glands. The segmental ganglia in the thorax and abdomen provide nerves to the appendages, dorsal muscles, sense organs, and heart. Insects have three pairs of thoracic ganglia and up to ten abdominal ganglia.

Giant-fibre systems—found in most arthropods with an elongated abdomen, such as shrimps, lobsters, and scorpions—take part in reflexes in which there is a tail flick or a jumping movement serving as a rapid means of escape.

The most common sensory receptors in arthropods are the cuticular hairs, many of which are mechanoreceptors, sensitive to touch, vibration, water currents, or sound waves; some hairs are chemoreceptors, which detect odours or chemicals in the water. Hairs situated near the joints are stimulated by body movements and thus provide a sense of the position of the joint or appendage during locomotion or flight. Many sensory cells and organs are concentrated in the antennae, and a statocyst is found at the base of each antenna.

The most highly developed organ for sound reception is the tympanal organ of grasshoppers, crickets, and cicadas. It consists of modified parts of the cuticle (the tympanum), to which are attached mechanoreceptor sensory neurons. Sound waves that produce vibrations in the tympanum are detected by the nerve cells.

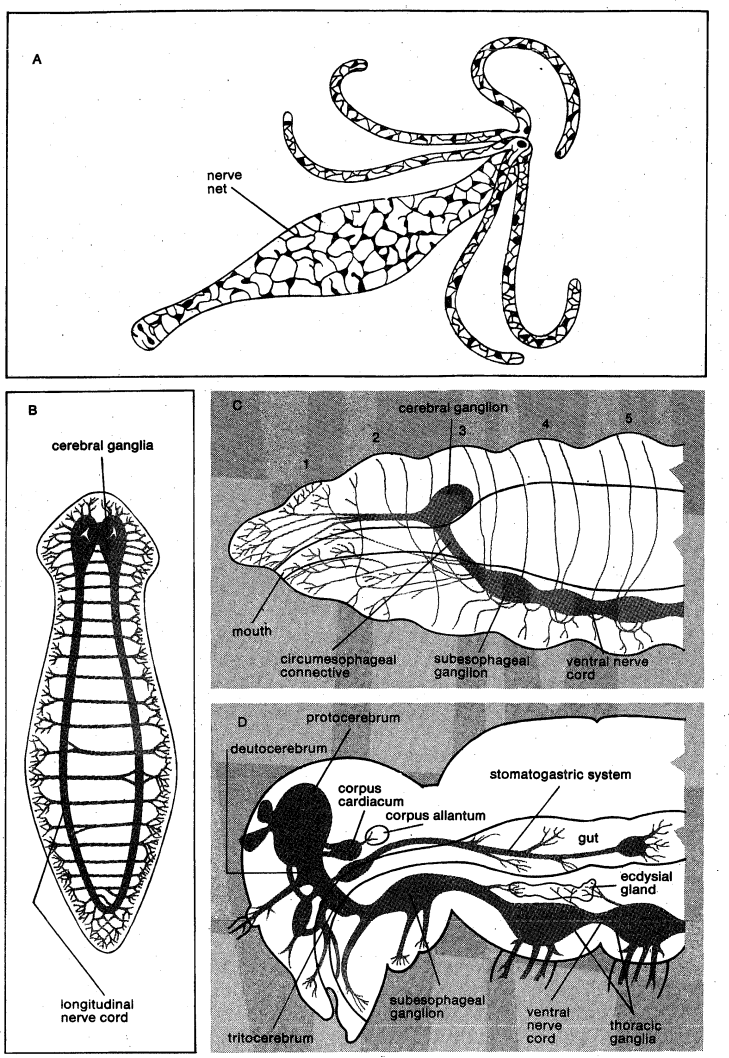


Figure 3: *Invertebrate nervous systems.*
(A) The nerve net of *Hydra*. (B) Central nervous system of a flatworm, composed of a brain (cerebral ganglia) and longitudinal nerve cords; (C) annelid, moderately cephalized with a brain, ventral nerve cord, and ventral ganglia; and (D) arthropod, highly cephalized and compartmentalized.

Spiders have several pairs of simple eyes with cup-shaped retinas. Crustaceans and insects, however, have a pair of well-developed compound eyes, each consisting of a large number of visual units called ommatidia. Each ommatidium contains six to eight sensory receptors arranged under a cornea and refractile cone and is surrounded by pigment cells, which adjust the intensity of light. Each ommatidium can act as a separate eye, capable of responding to its own visual field. Such an arrangement seems particularly well suited for detecting movement across a wide visual field.

Neurosecretory cells in crustaceans and arthropods

Neurosecretory cells, which have been identified in all the major invertebrate groups, reach their highest degree of development in the crustaceans and arthropods. The principal system of insects consists of neurosecretory cells in the protocerebrum of the brain. The axons of these cells form nerves that innervate structures called corpora cardiaca, situated just posterior to the brain. A corpus cardiacum is an organ in which neurosecretory products are stored in the nerve terminals for later release into the vascular system.

The neurosecretory cells of the protocerebrum manufacture a hormone called ecdysiotropin, or brain hormone. Sensory nerve impulses reaching the brain regulate the release of the hormone from the nerve endings into the blood within the corpora cardiaca. The hormone then stimulates a non-neural endocrine gland, the ecdysial gland, located in the thorax. The ecdysial glands in turn release a hormone, ecdysone, which initiates molting dur-

ing larval development and also stimulates differentiation into adult tissues. Another hormone, however, the juvenile hormone, keeps tissue in a juvenile or larval form. This hormone is released by the corpora allata, another pair of non-neural endocrine glands, located behind the corpora cardiaca. The successive life stages in insect development are, therefore, determined by the varying levels of ecdysone and juvenile hormone (see HORMONE; ENDOCRINE SYSTEMS).

IV. The vertebrate nervous system

The nervous system of vertebrates has two main divisions: the central nervous system, consisting of the brain and spinal cord, and the peripheral nervous system, which in man includes 12 pairs of cranial nerves and 31 pairs of spinal nerves, and the autonomic, or involuntary, nervous system. Because the arrangement of nerve cells and their fibres is so complex, it is helpful to understand some basic terms used in describing the vertebrate nervous system.

Anatomical structures such as the nervous system are described according to their position. In four-legged animals, the upper (back) surface is called dorsal and the lower (belly) surface ventral. The terms anterior, cranial, cephalic, or rostral refer to the head end of the body; posterior, or caudal, to the tail end. In man, since he stands erect, the situation is more complicated: dorsal becomes equivalent to posterior, and ventral is the same as anterior; cranial is often called superior, and caudal inferior. Objects near the middle plane of the body are medial and those farther away are lateral. Proximal refers to structures nearest the central bulk of a structure and distal to ones away from it. In referring to another structure, if it is located on the same side of the body, it is known as ipsilateral; if it is on the opposite side, it is contralateral.

Nerve cell bodies, or neurons, are often gathered into localized masses. In the peripheral nervous system these accumulations are called ganglia; in the central nervous system they are called nuclei. Portions of the central nervous system in which neurons predominate are called gray matter; areas in which myelinated nerves dominate are called white matter. Efferent, or motor, nerve fibres carry impulses away from the central nervous system; afferent, or sensory, fibres carry impulses toward the central nervous system. Visceral fibres innervate the viscera such as the heart and intestine, and somatic fibres innervate the body-wall structures such as skin and muscle. In the central nervous system, the nerve fibres are organized in bundles called tracts, or fasciculi. Ascending tracts carry impulses along the spinal cord toward the brain, and descending tracts carry them from the brain or higher regions in the spinal cord to lower regions. The tracts are often named according to their origin and termination; for example, the corticospinal tract consists of fibres running from the cerebral cortex in the brain to the spinal cord.

The account given here applies generally to the nervous system of mammals and particularly to that of the primates, including apes, monkeys, and men.

THE CENTRAL NERVOUS SYSTEM

The brain and its components. The brain is the greatly enlarged anterior portion of the central nervous system. It is enclosed within the cranial cavity of the skull and enveloped by membranes known as the meninges. Within the brain itself are cavities called ventricles, containing cerebrospinal fluid. The brain is divided into two major regions, the telencephalon (cerebral hemispheres) and the brainstem. The cerebral hemispheres include the cerebral cortex and basal ganglia. The brainstem is further subdivided into the diencephalon (thalamus and hypothalamus), mesencephalon (midbrain), metencephalon (pons and cerebellum), and the myelencephalon (medulla oblongata); the cerebellum lies behind the medulla and pons.

Major regions of the brain

Cerebrum. The two cerebral hemispheres make up the largest portion of the brain. Each hemisphere consists of an outer pallium, or cerebral mantle, and the deeply situated basal ganglia. The pallium is composed of an outer layer of gray matter, the cerebral cortex, consisting of the cell bodies of neurons, and an internal mass of white matter comprised of nerve fibres.

The surface of the hemispheres is wrinkled and convoluted by shallow furrows, or sulci; the ridges between the sulci are convolutions, or gyri. Deeper furrows are known as fissures. A deep longitudinal fissure incompletely separates the two hemispheres into right and left halves. At the base of the fissure, the hemispheres are joined by the corpus callosum, a broad band of nerve fibres. The two major grooves on the lateral surface of the brain are the lateral fissure (fissure of Sylvius), which starts at the base of the brain and extends upward and backward on the lateral surface, and the central sulcus (sulcus of Rolando), which runs from the middle of the dorsal border of the hemisphere downward almost to the lateral fissure.

The hemispheres are divided into lobes that have specific functions. The frontal lobe extends from the front of the brain to the central sulcus. The parietal lobe is the region behind the central sulcus and above the lateral sulcus. The temporal lobe lies below the lateral sulcus. The occipital lobe lies at the back of the brain, posterior to the temporal and parietal lobes. A fifth lobe, the insula, lies buried in the lateral sulcus. The olfactory lobe, or rhinencephalon, forms the lower (inferior) surface of the cerebrum.

The cells in the cerebral cortex are arranged indistinctly in six layers, from outside inward: molecular (I), external granular (II), external pyramidal (III), internal granular (IV), internal pyramidal (V), and fusiform (VI).

Each half of the cerebral cortex can also be divided into several regions depending on the distribution of afferent fibres or the origin of efferent fibres that connect with the deeper nerve centres. On this basis, the major functional regions of the cortex are the primary motor area, primary somesthetic area, primary visual area, primary auditory area, and the association areas.

The primary motor area is located on the anterior wall of the central sulcus and adjacent portions of the precentral gyrus of the frontal lobe. This region of the cortex controls the skeletal muscles on the opposite side of the body. The impulses travel over the corticospinal paths to the motor neurons of the spinal cord. The corticospinal tracts are formed by the axons of the Betz cells, giant pyramidal cells in layer V. The different parts of the body are represented in specific regions of the motor cortex, so that stimulation of a particular region of the cortex results in a particular movement.

General sensations from the skin, muscles, joints, and tendons reach the cortex by way of the thalamus. The fibres project to the somesthetic area located in the postcentral gyrus of the parietal lobe. The sensory cortex contains a pattern of localization of parts of the body similar to that in the motor cortex. Lesions of this area interfere with sensory perception but do not abolish it because some important sensations (*e.g.*, pain) enter consciousness at the level of the thalamus.

The primary visual area is located in the calcarine fissure of the occipital cortex. This region receives fibres carrying visual impulses from the lateral geniculate body; damage produces defects in vision. The primary auditory area is in the transverse temporal gyrus, which lies buried in the floor of the lateral cerebral fissure and receives auditory impulses from the medial geniculate body; injury can cause mild deafness.

Association areas of the cortex

Association areas, connected with the various motor and sensory areas, occupy a large portion of the cortex in higher vertebrates and are especially dominant among higher primates such as man. Association areas situated near the primary sensory areas are necessary to formulate sensory stimuli into images and to comprehend their meaning. Objects come to be associated with a constellation of past experiences, or memories. When sensory impulses excite these memories, the object is recognized. Before performing complex voluntary movements, there must first be a mental formulation of the plan, an idea.

The idea is transferred by association fibres to the motor area so that the movement can be executed. Complex sensory and motor associative mechanisms also underlie the comprehension of language.

The large remaining part of the frontal lobe, in front of the motor areas, is the prefrontal cortex, or frontal association area. Much of the prefrontal cortex is connected to the thalamus, which in turn has connections with autonomic centres in the hypothalamus. The frontal lobes also receive association fibres from other regions of the cortex. This region of the brain is responsible for certain aspects of higher intellectual functions and emotional behaviour in man.

Located on the inferior surface of the brain, the rhinencephalon is concerned primarily with the sensation of smell and the reception and integration of olfactory impulses (the rhinencephalon also includes the limbic system described below). The olfactory portion of the rhinencephalon includes the olfactory nerves, bulbs, tracts, and parts of the pyriform lobe. These regions of the brain constitute the paleopallium, which, in lower vertebrates, is very prominent.

A complex of structures forms the limbic system, or visceral portion of the brain. The limbic system consists of the limbic lobe, parts of the temporal and frontal cortex adjacent to the lateral sulcus, some thalamic and hypothalamic nuclei, and parts of the basal ganglia. The limbic lobe is formed by gyri surrounding the anterior end of the brainstem and corpus callosum. Parts of the limbic lobe, the hippocampus, and dentate gyrus constitute another old—in terms of evolution—part of the cerebral cortex, the archipallium. There are complicated interconnections between structures of the limbic system and with other parts of the brain; one major connection is the fornix, running from the hippocampus to the hypothalamus. The limbic system, through its connections with the cortex and hypothalamus, is thought to play a major role in the emotional responses to either hypothalamic activity or to higher thought processes and appears to be involved in reactions of fear, rage, or aggression and in sexual behaviour.

Deep within the cerebral hemispheres are masses of gray matter forming the basal ganglia, which regulate movement. These include the caudate nucleus, putamen, globus pallidus, and amygdaloid nuclear complex. Only the last named nucleus is part of the limbic system. The other nuclei form the corpus striatum, which constitutes the extrapyramidal system (distinguished from the pyramidal system originating in the motor areas of the cerebral cortex). Fibres from the corpus striatum run to the thalamus and subthalamus.

Diencephalon. The diencephalon is the most anterior end of the brainstem. Although small, it is a complex and important part of the nervous system, being composed of the thalamus, geniculate bodies, hypothalamus, subthalamus, and epithalamus.

The thalamus: an integration centre

The thalamus, an ovoid mass of gray matter lying on either side of the third ventricle, is composed of several groups of nuclei connected with other parts of the brain. The subcortical nuclei receive fibres from the major sensory systems and connect only with subcortical structures such as the hypothalamus and basal ganglia. Another group of nuclei, the cortical relay nuclei, receive fibres from the sensory tracts and also from the cerebellum, reticular system, and basal ganglia. Fibres from these nuclei project to the cerebral cortex. The association nuclei do not receive fibres from the sensory systems but are connected with other nuclei in the diencephalon and project to the association areas of the cerebral cortex. The thalamus is a major sensory integration centre, receiving all sensory impulses except the olfactory ones. It also seems to play a major role in the regulation of the state of consciousness, alertness, and attention and is responsible for some of the emotional experiences that accompany sensory perception.

The geniculate bodies are two pairs of small swellings on the lateral surfaces of the thalamus. The optic tracts, collections of nerve fibres from the retina of the eyes, transmit visual impulses to the lateral geniculate bodies, which then relay them to the visual cortex. The medial geniculate bodies receive auditory impulses via parts of the midbrain and relay them to the temporal lobe of the cerebral cortex.

The hypothalamus lies below the thalamus and forms the floor of the third ventricle. On the ventral surface of the brain are the crossing of the optic tracts (the optic chiasma), a slight eminence called the tuber cinereum, the infundibulum (a stalk connecting the tuber cinereum with the neural lobe, or pars nervosa, of the pituitary), and a pair of small swellings called the mammillary bodies. Although small, the hypothalamus has extensive and complex fibre connections. It receives fibres from olfactory nuclei, thalamus, hippocampus, amygdala, frontal lobe, and midbrain. From the hypothalamus, tracts proceed to the thalamus, tegmentum of the midbrain, autonomic centres and reticular formation of the brainstem, and neural lobe of the pituitary. The hypothalamus has important functions related to temperature regulation, water balance, and glandular secretion. Some of these functions are controlled through the connections with the pituitary; others are mediated through control of autonomic activity. If the external temperature drops, for example, the posterior hypothalamus initiates activities necessary to prevent heat loss (*e.g.*, constriction of blood vessels) and, under more extreme conditions, initiates muscular and metabolic activities that actually increase heat production (*e.g.*, shivering, oxidation of body resources). If the external temperature rises, activities that increase heat loss (*e.g.*, sweating and dilation of blood vessels) are initiated. Another regulatory activity of the hypothalamus is control of food intake by interaction of a feeding centre, which produces a desire to eat, and a satiety centre, which decreases the appetite. Finally, through its connections with the autonomic nervous system, the hypothalamus is a principal centre of emotional expression.

The subthalamus lies between midbrain and the thalamus. It receives connections from the basal ganglia. The epithalamus, which forms the roof of the diencephalon, has a curious small outgrowth, the pineal body (epiphysis), connected to it. The functions of the pineal body are obscure, but it is thought to be an endocrine gland that influences sexual function. It may be responsible for the sexual cycles associated with seasonal light changes such as occur in birds and small mammals. In lower vertebrates, it contains light-sensitive cells.

Possible functions of the pineal body

Midbrain. The midbrain, or mesencephalon, is the shortest portion of the brainstem. It is divided into three parts: the tectum, or quadrigeminal plate; the crura cerebri; and the tegmentum. The crura cerebri and tegmentum together comprise the cerebral peduncles, two large projections that contain fibres passing between the cerebral cortex and lower levels of the central nervous system. The two parts of the peduncles—crura cerebri and tegmentum—are separated by a black coloured nucleus known as the substantia nigra. A red nucleus is found in the raised part of the tegmentum and, giving rise to the rubrospinal tract, is part of the reticular formation of the brainstem. Fibres of the superior cerebellar peduncles enter the tegmentum. The cavity of the midbrain, the cerebral aqueduct, is surrounded by the central gray substance formed of the nuclei of two nerves that participate in eyeball movement, the trochlear and oculomotor nerves, and a portion of the trigeminal nucleus.

The quadrigeminal plate contains four rounded elevations called the corpora quadrigemina. The two higher projections are the superior colliculi, which receive fibres from the optic nerves and serve as reflex centres for eye movements and blinking. The lower pair—the inferior colliculi—are auditory reflex centres and receive fibres from the lateral lemnisci, bands of nerve fibres on either side of the peduncles. Cells in the nuclei of the tectum send fibres to motor neurons in the brainstem by way of the tectobulbar tracts and to the spinal cord by way of the tectospinal tracts.

Pons. The pons is an oblong-shaped mass located just ahead of the medulla. The middle cerebellar peduncles leave the lateral aspect of the pons, and the superior

cerebellar peduncles enter the cerebellum from the dorsal aspect. The abducens nerve—another nerve that influences the eyeball movement—leaves the ventral surface of the pons. The trigeminal nerve arises from the lateral surface; the facial and acoustic nerves emerge from the angle formed by the pons, medulla, and cerebellum.

The pons is divided into a dorsal portion known as the tegmentum and a ventral, or basilar, portion. The dorsal portion of the pons contains the forward extension of the reticular formation from the medulla. It also contains the motor nuclei of the facial, abducens, and trigeminal nerves, the main sensory and spinal nuclei of the trigeminal nerve, and the vestibular and cochlear nuclei of the acoustic nerve. The trapezoid body and the lateral lemniscus, which goes to the midbrain, are formed by fibres from the cochlear nuclei.

Fibres leaving the four vestibular nuclei extend into the cerebellum. Other fibres form a tract called the medial longitudinal fasciculus, through which branches descend through the medulla to the spinal cord; others ascend and end in the nuclei of the muscles of the eyeball. Still other fibres from the nuclei give rise to the two vestibulospinal tracts. The vestibular nuclei are thus important relay centres for coordinating movements of the eyes, neck, and body and for maintaining equilibrium, posture, and muscle tone.

The basilar portion of the pons consists of transverse and longitudinal fibres between which are the pontine, or pontile, nuclei. The longitudinal tracts are the corticospinal, corticobulbar, and corticopontine tracts. The transverse tracts, formed by the axons of nerve cells in the pontine nuclei, cross to the opposite side and form the middle cerebellar peduncle. The cells of the pontine nuclei receive terminations from the corticopontine tract and thus act as a relay station between the cerebral cortex and cerebellum.

Cerebellum. The cerebellum is divided into two large lateral portions, the cerebellar hemispheres, and a narrow middle portion called, because of its "tangled worm" appearance, the vermis. Three lobes can be distinguished on each hemisphere. The flocculonodular lobe in the posterior lower region consists of paired bumps, called flocculi, and the nodulus, or lower part of the vermis. The anterior lobe comprises that part of the cerebellum lying anterior to the primary sulcus. The posterior lobe, located between the other two lobes, constitutes the main part of the cerebellum. The flocculonodular node is part of the vestibular system involved in equilibrium. In terms of evolution, it is the oldest part of the cerebellum and is called the archicerebellum. The anterior lobe receives most of the proprioceptive impulses (originating in muscles and tendons) from the spinocerebellar pathway and regulates muscle tone. It is also an old part of the cerebellum and is called the paleocerebellum. The posterior lobe is the most recent development of the cerebellum—the neocerebellum—and is most prominent in primates and man. This portion of the cerebellum is involved in the coordination of fine, skilled movements.

"Old" and "new" parts of the cerebellum

The cerebellum, like the cerebrum, consists of a superficial mantle of gray matter and an inner mass of white matter. The surface of the cerebellum has a large number of sulci and fissures between narrow ridges called folia, or gyri. The cortex of gray matter has an outer molecular layer with many nerve cell processes, a single row of very large neurons (Purkinje cells), and an inner granular layer with many small nerve cells. Deep within the white matter are four pairs of cerebellar nuclei—the fastigial, globose, emboliform, and dentate—which act as relay stations between the cerebellar cortex and other parts of the brain.

The cerebellum is attached to the brainstem by the cerebellar peduncles, which are composed of large numbers of nerve fibres entering and leaving the cerebellum. These fibres connect the cerebellum with the cerebral cortex, brainstem, cranial nerve nuclei, and the spinal cord. The superior cerebellar peduncle, or brachium conjunctivum, contains most of the fibres leaving the cerebellum. These fibres arise from the dentate and other cerebellar nuclei and extend to the red nucleus, thalamus, and reticular formation. The middle cerebellar peduncle, or brachium pontis, consists entirely of fibres from the nuclei embedded in the pons. The inferior cerebellar peduncle, or restiform body, contains one efferent tract, which runs from the fastigial nucleus to the vestibular nuclei. It also contains several groups of afferent fibres that arise from the vestibular nerve and nuclei, inferior olivary nuclei, gracile and cuneate nuclei, and the posterior spinocerebellar tract from the spinal cord.

Conscious perception does not occur in the cerebellum nor can the coordinative functions of the cerebellum be voluntarily modified. In order to control or modify muscular activity, the cerebellum must receive information about position and movement. It is informed of the commands being issued by the cerebral cortex over the cortico-ponto-cerebellar pathway, of the status of the muscles and joints of the body over the spinocerebellar tracts, and of other sensory events from the skin, eyes, and ears.

Within the cerebellar cortex, the sensory signals are evaluated and appropriate corrections are made for any mistakes or inaccuracies of muscle activity. Nerve impulses are dispatched from the Purkinje cells to the cerebellar nuclei, from which there are several routes over which impulses reach the voluntary motor system to modify muscular movements.

Medulla oblongata. The medulla, the lower part of the brainstem, extends from the pons to join the spinal cord at the opening (foramen magnum) at the base of the skull. On its ventral surface, the corticospinal tracts can be seen as two longitudinal ridges on either side of the anterior median fissure. Such cranial nerves as the hypoglossal, glossopharyngeal, vagus, and accessory emerge from the lateral surface of the medulla. The dorsal, anterior surface of the medulla is formed by the thin roof of the fourth ventricle, alongside of which are the prominent inferior cerebellar peduncles.

Internally, many of the long fibre tracts of the spinal cord extend into the medulla. They may pass through directly or change their positions; their complete course and function are described along with the spinal cord (below). In the lower region of the medulla, the ventral pyramidal tracts form a prominent crossing (decussation) of the fibres to opposite sides. Dorsally, the fasciculi gracilis and cuneatus from the spinal cord terminate in the nuclei gracilis and cuneatus. The fibres of the cells in these nuclei arch downward as the internal arcuate fibres and cross the midline as the decussation of the medial lemniscus. The medial lemniscus travels up the brainstem and terminates in the thalamus.

In the upper half of the medulla, the inferior olivary nucleus in the ventrolateral region is a convoluted band of gray matter receiving fibres from the cerebral cortex and other higher structures and from the spino-olivary tract. From the olive (a prominence on each side of the medulla nearest the pons) fibres join the spinocerebellar fibres in the inferior cerebellar peduncle and proceed to the cerebellum. On the dorsal part of the medulla are the origins of three cranial nerves: the nucleus of the hypoglossal nerve, supplying the muscles of the tongue; the dorsal motor nucleus of the vagus nerve, supplying the viscera (heart, lungs, intestines); and the vestibular nuclei, receiving sensory fibres from the vestibular division of the acoustic nerve. Laterally, the solitary tract and nucleus convey taste impulses from several nerves. The spinal trigeminal tract and nucleus, which extends from higher levels into the lower portion of the medulla, convey sensory impulses from the face. The nucleus ambiguous supplies fibres via several nerves that innervate the muscles of the throat and voice box.

Around the central canal of the medulla is a diffuse network of cells and fibres called the reticular formation. It contains a respiratory centre, which controls the rhythmic respiratory movements. Other cells in the reticular formation control blood pressure, heart activity, and intestinal movements.

The reticular formation

The spinal cord. The spinal cord with its coverings lies loosely within the vertebral canal of the backbone. It extends from the foramen magnum at the base of the skull to the lower border of the first verterbra of the

lumbar section. The spinal cord is roughly cylindrical in shape, though somewhat flattened. The swollen regions of the cord that innervate the appendages are known as the cervical and lumbar enlargements. Below the lumbar enlargement, the cord tapers in a cone called the conus medullaris. From the conus, a slender non-nervous filament, the filum terminale, runs downward and attaches to the tailbone, or coccyx. The spinal nerves are formed by fusion of dorsal and ventral roots, which originate from the spinal cord. The spinal nerves pass through the openings between the vertebrae. Since the vertebral column is longer than the spinal cord, the spinal nerves must run downward at an angle to the cord to reach their corresponding intervertebral openings. In the lumbar and sacral regions, the roots of the spinal nerves extend past the end of the spinal cord in a bundle of nerve fibres called the cauda equina because of its fancied resemblance to a horse's tail.

There are a number of longitudinal grooves, or furrows, on the surface of the cord (Figure 4). A deep anterior, or ventral, median fissure and a shallow posterior, or dorsal, median sulcus divide the cord into equal right and left halves except in the midpoint, where the two halves are joined by tracts that cross. The posterolateral and anterolateral sulci are located more laterally. Dorsal roots of the spinal nerves enter the cord at the posterolateral sulcus and the ventral roots leave at the anterolateral sulcus. In the cervical and upper thoracic portions of the cord, the posterior intermediate sulcus runs between the median and posterolateral sulci.

Spinal columns. The spinal cord consists of centrally placed gray matter surrounded by white matter. The gray matter, which contains nerve cell bodies, unmyelinated nerve fibres, dendrites, and terminal branches, roughly has the form of a butterfly or the letter H. The crossbar is the gray commissure, in the centre of which runs the central canal of the spinal cord. The gray substance extending dorsally (posteriorly) from the commissure is called the dorsal, or posterior, column, or horn; that extending ventrally (anteriorly) is the ventral, or anterior, column, or horn. Between the dorsal and ventral horns is a slight lateral projection of gray matter known as the lateral horn. The cells of the gray matter can be divided into two major columns, motor and receptor. The motor neurons are situated in the ventral and lateral gray columns and give rise to the ventral roots of the spinal nerves. Large motor neurons in the ventral horn supply the voluntary skeletal muscle, and smaller neurons give rise to efferent fibres that innervate the sensory endings in muscles (neuromuscular spindles). The motor cells of the lateral column give rise to the preganglionic fibres, which leave the cord via the ventral root and terminate in the sympathetic ganglia of the thoracic, lumbar, and sacral portions of the autonomic system. The receptor part of the spinal cord is in the dorsal columns, where many fibres of the dorsal root (sensory) cells terminate. These cells—the central, internuncial, commissural, or association neurons—are confined to the central nervous system. Most of their fibres divide in the shape of a T as they enter the white matter and form ascending and descending branches. When the fibres remain on the same side of the spinal cord, the cells are known as association, or ipsilateral, column cells. If the axons cross to the other side of the spinal cord, the cells are called contralateral, or commissural, cells. The most prominent group of cells in the dorsal column is the nucleus dorsalis, or Clarke's column, which sends fibres into the dorsal spinocerebellar tract. The substantia gelatinosa is the chief associative centre for incoming impulses.

The funiculi of the white matter

The white matter is divided into three main columns known as funiculi. The dorsal funiculus lies between the dorsal median septum and the dorsal lateral sulcus. The lateral funiculus is situated between the dorsal lateral sulcus and ventral lateral sulcus. The region between the ventral lateral sulcus and ventral median fissure is the ventral funiculus. The white matter in the two halves of the spinal cord is joined by the ventral white commissure, which lies directly in front of the ventral gray commissure.

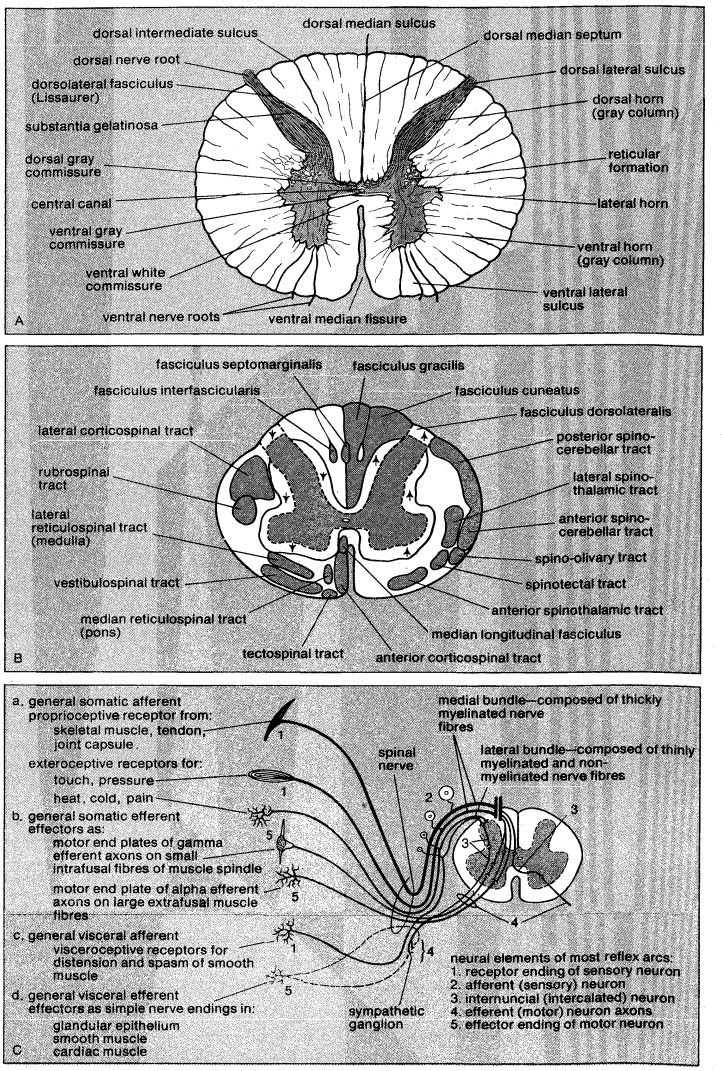


Figure 4: (A) Transverse section of the mammalian spinal cord in the midthoracic region. (B) (Left) Descending and (right) ascending pathways of the spinal cord. (C) Functional components of a spinal nerve. Receptor and effector endings and reflex connections in the spinal cord are shown.

From (A) C.M. Goss (ed.), *Gray's Anatomy of the Human Body*, 28th ed. (1966), Lea and Febiger; (B,C) R.C. Truex and M. Carpenter in A. Elwyn (ed.), *Human Neuroanatomy*, 6th ed. (© 1969), The Williams and Wilkins Co., Baltimore

Spinal tracts. Within the dorsal, lateral, and ventral columns, the nerve fibres are organized in specific bundles called tracts, or fasciculi. The fasciculi connect different levels of the spinal cord or connect the spinal cord with the brain. The tracts are called descending or ascending depending on the direction in which the nerve fibres run within them. The descending tracts, formed by the fibres of nerve cells located in the brain, end near motor neurons in the ventral horns of the spinal cord and so are mainly associated with motor activities. The ascending tracts, which originate from neurons in the receptor region of the spinal cord gray matter, carry to the brain impulses related to sensation. Other (intersegmental) tracts carry fibres involved in certain reflex activity.

The corticospinal, or pyramidal, tracts are the largest and most important of the descending tracts. They consist of nerve fibres of pyramidal cells in the cerebral motor cortex. At the level of the medulla, some of the fibres cross to the opposite side and form the lateral corticospinal tract. The uncrossed fibres continue as the ventral corticospinal tract. The fibres of the pyramidal tracts terminate on motor cells in the ventral horn and carry the impulses for voluntary motor and skilled activities.

The rubrospinal tract in the lateral funiculus arises from the contralateral red nucleus of the midbrain. The fibres end on neurons in the ventral horn and carry impulses that control muscle tone in certain muscle groups.

The two vestibulospinal tracts (medial and lateral) in the ventral funiculus originate from the vestibular nuclei at the junction of the medulla and pons. Impulses are carried to this area from the internal ear over the vestibular nerve. The fibres of these tracts terminate on lower motor neurons in the ventral horns. These tracts are involved in balance reflexes that maintain equilibrium.

The reticulospinal tracts (lateral and medial) descend in the ventral funiculus. They arise in the reticular substance of the medulla and pons. Some of the fibres end on the preganglionic neurons of the lateral gray columns and carry impulses related to several autonomic activities such as temperature and respiratory regulation. Other fibres act on certain motor neurons to influence muscle tone.

The tectospinal tracts descend from neurons of the contralateral superior colliculus—an optic relay centre—and mediate postural reflexes, especially of the head and upper extremity, in response to visual stimuli. Fibres of the olivospinal tract arise from the inferior olive of the medulla and help coordinate movements of the head and upper extremity.

Descending autonomic pathways

The pathways just described eventually terminate on ventral horn cells that innervate the voluntary striated muscles. The spinal cord also contains descending autonomic pathways for the innervation of visceral structures such as smooth muscle, heart muscle, and glandular tissue. Many of these fibres originate in the hypothalamus. The pathways are diffuse and located mainly in the ventral and ventrolateral portions of the white matter. The fibres terminate on preganglionic cells in the lateral gray column.

The largest ascending tracts to the cerebral cortex are found in the dorsal white columns. These pathways are the fasciculus gracilis and the fasciculus cuneatus, which are formed by the ascending branches of the central processes of the dorsal root ganglion cells. The fasciculus gracilis contains long ascending root fibres from the sacral, lumbar, and lower thoracic ganglia. The fasciculus cuneatus contains fibres from the upper thoracic and cervical ganglia. All these fibres terminate on neurons in the nucleus gracilis and nucleus cuneatus in the medulla. The axons of these latter neurons then cross the midline, travel upward in the medial lemniscus, and end in the ventral dorsolateral nucleus of the thalamus. From the thalamus, neurons in turn send axons to the sensory areas of the cerebral cortex. The fibres in the dorsal columns carry impulses from tactile receptors necessary for the ability to detect the size, shape, and texture of objects by feeling them. In addition to tactile impulses, these tracts carry the proprioceptive impulses that convey information concerning the movements of limbs and the position of different parts of the body.

Axons of neurons in the dorsal horn cross in the anterior white commissure and turn upward in the contralateral ventral spinothalamic tract. The fibres terminate on cells of the ventral dorsolateral nucleus of the thalamus, from which the impulses are relayed to the cerebral cortex. The ventral spinothalamic tract carries the impulses of simple or light touch.

The central processes of pain and temperature receptors enter the zone of Lissauer (dorsolateral fasciculus) and ascend or descend one or two segments before ending on cells in the substantia gelatinosa and dorsal gray matter. The axons of these cells cross in the anterior white commissure and ascend in the opposite lateral funiculus as the lateral spinothalamic tract to the ventral dorsolateral nucleus of the thalamus, from which fibres extend to the cortex.

The uncrossed posterior spinocerebellar tract arises from the cells of Clarke's column (dorsal nucleus) of the dorsal horn. Its fibres enter the cerebellum by way of the inferior cerebellar peduncle and terminate in the cortex of the cerebellum. This tract relays impulses from stretch, touch, and pressure receptors, information utilized in the coordination of posture and movement of individual limb muscles. Other axons from the dorsal horn cross in the ventral white commissure and join uncrossed fibres in the ventral spinocerebellar tract. These fibres, which enter the cerebellum in the superior cerebellar peduncle, also convey information concerning movement and posture.

Besides these major ascending pathways, several other ascending fibre systems are found in the spinal cord. These include the spino-olivary tract to the inferior olivary nucleus, spinoreticular fibres to the brainstem reticular formation, the spinotectal tract to the superior colliculus, a spinovestibular system to the lateral vestibular nucleus, spinopontine fibres to the pontine nuclei, and a spinocortical tract to the cerebral cortex.

Intersegmental tracts

Intersegmental tracts connect different levels of the spinal cord. In the majority of reflex arcs, there is at least one central, or internuncial, neuron between afferent and efferent neurons. The internuncial neuron may then send its axons to motor cells of the same segment or to higher and lower segments. Commissural cells send axons to the white matter of the opposite side of the spinal cord. All these nerve fibres, which may be ascending or descending, crossed or uncrossed, begin and end in the spinal cord and connect its various levels. They constitute the proprius bundles (fasciculi proprii), spinospinal, or fundamental columns of the spinal cord, and are found in all the funiculi of the white matter immediately outside the gray matter.

In addition to the proprius bundles, there are some systems associated with the major conduction pathways. The dorsal root neurons whose ascending branches form the fasciculus gracilis and fasciculus cuneatus tracts also have descending branches, which are involved in intersegmental reflexes. In the cervical and thoracic cord, these descending fibres form a small bundle, known both as the fasciculus interfascicularis and as the comma tract of Schultze, located in the dorsal funiculus. In the lumbar region, the fibres descend in the septomarginal fasciculus, situated in the middle of the dorsal median septum.

The brain coverings, cavities, and cerebrospinal fluid. *Meninges.* The spinal cord and brain are tightly covered by three connective tissue envelopes, the meninges: the outer layer is the dura mater, the middle layer the arachnoid, and the inner the pia mater.

The dura mater is tough, dense, and firm. The dura mater of the brain has two layers: a thin inner layer and a thicker outer layer that serves as the covering (periosteum) for the inner surface of the cranial bones. In the spinal cord the vertebrae have their own coverings, and, while the dura is separated from the vertebral periosteum by the epidural space, it is connected to the spinal cord on each side by a series of ligaments. The nerve roots emerging from the cord penetrate and become invested by the dura.

The arachnoid is a thin, delicate membrane devoid of blood vessels. From its inner surface, numerous strands connect to the pia mater across the subarachnoid space. The dura and arachnoid pass over the sulci and fissures on the surface of the brain and spinal cord without dipping into them. The pia mater, however, is a thin membrane that closely adheres to the surface of the brain and spinal cord, following each crevice. The pia contains a large number of blood vessels.

Ventricles. The central nervous system begins its development as a hollow tube. The cavity persists in the adult as the central canal of the spinal cord and the ventricles of the brain. There are two large lateral ventricles in the cerebral hemispheres. The third ventricle lies in the thalamic region and is connected to the lateral ventricles through openings called the interventricular foramina (foramina of Monro). The cerebral aqueduct (aqueduct of Sylvius) connects the third ventricle with the fourth ventricle in the pons and medulla. The central canal of the spinal cord extends from the fourth ventricle.

Cerebrospinal fluids. All these interconnected spaces —the ventricles, central canal, and subarachnoid spaces —are lined with a cellular layer called the ependyma and filled with cerebrospinal fluid, a clear, colorless, noncellular fluid that bathes the interior of the brain and spinal cord and cushions against any jarring shocks. It might also provide a pathway for certain substances to reach the intercellular spaces of the brain and for others to be carried away.

Sources of cerebrospinal fluid

There is constant production, circulation, and resorption of cerebrospinal fluid. The principal sources are the choroid plexuses, membranes found in part of the wall of the lateral ventricles and in the roof of the third and fourth ventricles. They are formed by vascular pia mater pushing the ependyma into the ventricles. Fluid from the lateral ventricles passes through the foramina of Monro into the third ventricle. More fluid is added here and passes through the aqueduct of Sylvius to the fourth ventricle, where additional fluid is added. From the fourth ventricle, it reaches the subarachnoid space on the surface of the nervous system by passing through medial (foramen of Magendie) and lateral (foramina of Luschka) openings. At these regions, the subarachnoid space is very wide forming dilated cisterns. The largest is the cisterna cerebellomedullaris (cisterna magna) between the cerebellum and medulla. The apertures of the fourth ventricle open into this cavity. From the cisterns, the cerebrospinal fluid diffuses into the subarachnoid spaces around the brain and spinal cord, most of it absorbed by pia–arachnoid extensions (villi) or granulations (Pacchionian bodies) into venous sinuses of the dura.

Neuroglia: non-nervous tissue. Although the number of nerve cells in the central nervous system is enormous, there is an even greater number of non-nervous cells called glial, or neuroglial cells (comprising the neuroglia), which intermingle with the nerve cells in the central nervous system. There are three types of glial cells: astrocytes, oligodendrocytes, and microglia. Besides glial cells, the ependyma and the so-called satellite cells on the surface of neurons in peripheral ganglia belong to the non-nervous tissues of the central nervous system. The Schwann cells of peripheral nerves could be considered as peripheral neuroglia.

The astrocytes are the largest and most numerous of the glial elements, with long branching processes that invest the nerve cells and provide much of the packing material of the nervous system. They provide structural support to nerve cells, may help regulate the transfer of substances from blood capillaries to the nerve tissue, and may prevent the spread of impulses into improper channels. The oligodendrocytes are smaller than the astrocytes and have fewer processes. They are aligned in columns along the nerve fibres and are believed to form the myelin sheaths in the central nervous system. The satellite cells on the surface of some neurons appear to be a type of oligodendrocyte.

The microglial cells are smaller than the other glial cells and are found between nerve cells and along blood vessels. They are considered to be the scavengers of the nervous system, moving about under conditions of inflammation or injury and removing particles and debris in the manner of phagocytes in the blood.

THE PERIPHERAL NERVOUS SYSTEM

Nerve fibres classified by function and innervation

Nerve fibres can be classified according to their function and the regions they innervate. Nerve fibres innervating the body wall structures are designated as somatic; those innervating the internal organs are called visceral. Sensory fibres, which carry impulses to the central nervous system, are called afferent; motor fibres, carrying impulses away from the brain or spinal cord, are efferent. General somatic afferent fibres carry exteroceptive sensory impulses (of pain, temperature, touch, pressure) from the skin and connective tissue and proprioceptive sensory impulses (stretch, position) from the muscles, tendons, and joints. General visceral afferent fibres carry sensory impulses from the visceral organs including the blood vessels. Special somatic afferent fibres carry impulses from the sense organs such as the eye and the ear. Special visceral afferent fibres convey the impulses for smell and taste. Somatic efferent fibres carry impulses to skeletal muscle. General visceral efferent fibres innervate smooth muscle, the heart, and glands. Special visceral efferent fibres innervate the muscles of mastication, facial muscles, and muscles of the pharynx and larynx.

On the basis of anatomical origin nerves are classified as cranial, when their roots arise in the brain, or spinal, when their roots originate in the spinal cord.

Cranial nerves. The 12 pairs of cranial nerves, named from studies of man and other mammals (see NERVOUS SYSTEM, HUMAN), are organized similarly in all vertebrates with minor differences. These nerves are the olfactory (I), optic (II), oculomotor (III), trochlear (IV), trigeminal (V), abducens (VI), facial (VII), acoustic (VIII), glossopharyngeal (IX), vagus (X), accessory (XI), and hypoglossal (XII). All 12 nerves are found in reptiles, birds, and mammals. In fish and amphibians the first ten are present, and the eleventh is part of the tenth nerve. In addition, all vertebrates have an anterior nerve, the terminal, which supplies the olfactory organs. The terminal nerve is rudimentary in primates.

The cranial nerves can be grouped into several functional categories. Special somatic sensory fibres are carried by nerve I (smell), nerve II (vision), and the cochlear (hearing) and vestibular (equilibrium) divisions of nerve VIII. In fish and aquatic amphibians, sensations of water movement are carried in lateral line nerves. The special visceral sense of taste is conveyed by nerves VII, IX, and X. These sense of smell is often classified also as special visceral because it is associated with the visceral function of digestion. Some cranial nerves innervate the voluntary muscles derived from the embryonic somites. These nerves, called general somatic efferent, supply the muscles of the eyeball (nerves III, IV, and VI) and the muscles that move the tongue (nerve XII). These nerves also carry proprioceptive (locational) impulses from these muscles. Several cranial nerves innervate the branchiomeric (gill) arches of fish and the derivatives of the arches in higher vertebrates. The nerves supplying these structures have mixed motor and sensory components and include nerves V (first arch, the jaws), VII (second arch, the hyoid arch), IX (third arch), and X and XI (other arches). Sensations from the face (touch, pain, temperature, proprioception) are carried in general somatic afferent fibres of nerve V. General visceral afferent fibres in nerves VII, IX and X carry sensations from visceral organs, including the heart and intestines. Special visceral efferent (motor) fibres supply the muscles of the gill arches in fish and amphibians. In higher vertebrates they innervate the muscles of chewing (nerve V), facial expression (nerve VII), pharyngeal muscles of swallowing and laryngeal muscles of the vocal cords (nerves IX and X), and some muscles of the neck (nerve XI). Several of the cranial nerves have general visceral efferent (parasympathetic motor) fibres of the autonomic nervous system. Fibres in nerve III supply the muscles for focussing the eye and constricting the pupil. The lacrimal (tear) gland is controlled by nerve V and the salivary glands by nerves VII and IX. The thoracic viscera (heart, lungs, esophagus, glands, intestine) are supplied by the vagus (X) with a small contribution from nerve XI.

The primitive hypothetical vertebrate may have exhibited a segmental pattern of peripheral nerves associated with the brain similar to the arrangement in the spinal cord. The evolution of cranial nerves, however, did not involve a union of dorsal and ventral roots as in the spinal cord. Instead, one or another root dropped out. In the primitive situation, the dorsal root contained visceral motor fibres as well as sensory fibres, while the ventral roots were somatic motor in character. Thus, the cranial nerves of primates that are largely somatic motor (III, IV, VI, and XII) represent original ventral roots. The branchiomeric nerves with sensory and visceral motor components represent dorsal root nerves (V, VII, IX, X and XI).

Spinal nerves. A spinal nerve arises from each segment of the spinal cord. The number of spinal nerves varies from ten pairs, in some tailless amphibians, to over 500 pairs, in snakes. The nerves are named for the level of emergence from the spinal cord, as cervical (neck region), thoracic (chest region), lumbar and sacral (lower back region), and coccygeal (tailbone).

Each spinal nerve is formed by two roots, a dorsal sensory and a ventral motor root, which arise from the spinal cord. In the most primitive vertebrates, the roots remain as distinct nerves. In all other vertebrates, the

dorsal and ventral roots fuse to form the spinal nerve, which passes through an opening (foramen) in the vertebra. The sensory ganglion, a cluster of cell bodies of the sensory neurons, is situated on the dorsal root.

The dorsal roots contain general somatic afferent fibres (sensory fibres from muscle, skin, and other somatic structures) and general visceral afferent fibres (from visceral structures like the intestine). The ventral roots contain somatic efferent fibres, the motor fibres innervating somatic structures, and the general visceral efferent fibres that are preganglionic axons for the autonomic system.

After a short distance the spinal nerve branches into four divisions called rami: dorsal ramus, ventral ramus, ramus communicans, and meningeal ramus. The dorsal ramus innervates the muscles and skin of the back. The large ventral ramus supplies the ventrolateral portion of the body wall and the extremities. The small meningeal branch re-enters the intervertebral foramen and supplies the meninges and vertebral column. The ramus communicans joins the sympathetic ganglion. The general visceral efferent fibres leave the spinal nerve through the white portion (myelinated nerves) of the ramus communicans. Postganglionic general visceral efferent fibres from the sympathetic ganglion join the spinal nerve over the gray ramus (unmyelinated fibres). The general visceral afferent fibres reach the spinal nerve via the white ramus.

When, during development, the embryonic segments migrate to their adult positions—reflecting roughly the changes that occurred during evolution—the spinal nerves migrate with them, resulting in the intermixing and crossing of adjacent nerve trunks in certain areas.

Role of vertebrate plexuses

In many vertebrates, collections of nerves (plexuses) emerge at areas of concentration along the spinal cord. The cervical plexus emerges in the neck region and provides nerves that supply many muscles in the neck and chest as well as the diaphragm. The brachial plexus supplies nerves to the upper extremities—shoulder, forelimb, and hands. The thoracic spinal nerves follow more regularly and do not collect in a plexus. They innervate the muscles of the back, ribs, and abdomen. The lumbosacral plexus innervates the hindlimbs and feet primarily, and the pudendal and coccygeal plexuses supply nerves to the anus and to the genital organs.

Sensory and motor fibres. The spinal nerves as well as the peripheral branches are composed of sensory and motor nerve fibres. The cell bodies of the sensory neurons are located in the sensory ganglia of the dorsal roots. Each sensory neuron has a single process that leaves the cell body and then branches into two. One branch enters the dorsal horn of the spinal cord, and the other runs to the periphery and ends in the sensory receptor organs. The cell bodies of the motor neurons are found in the ventral horn of the spinal cord. From there, axons run into the ventral root, join the spinal nerve, and finally innervate the skeletal muscles of the body.

Receptors can be grouped in three categories on the basis of the origin of the impulse: exteroceptors, proprioceptors, and interoceptors. The exteroceptors receive stimuli from the surface of the body; they carry the sensations of touch, pressure, skin pain and temperature, smell, sight, and sound. The proprioceptors receive sensations from the joints, muscles, and tendons. Exteroceptors and proprioceptors constitute somatic receptors. The interoceptors carry stimuli from the internal organs and are classified as viscероceptors; they carry the sensations associated with digestion, excretion, sex, deep pain, and circulation. (For a detailed account of the reception of sensation see SENSORY RECEPTION; specific receptors and their function are considered in MECHANORECEPTION; PHOTORECEPTION; and SOUND RECEPTION.)

The autonomic nervous system. Smooth muscle (digestive tract and internal organs), heart muscle, and glandular tissues of the body receive their innervation from the autonomic, visceral, or vegetative, nervous system. The reactions of the visceral organs are to a great extent involuntary and unconscious and are controlled by visceral efferent neurons. These same organs also have visceral afferent nerves, which initiate visceral reflexes. The cell bodies of these sensory nerves are located in the dorsal root ganglia. Descending impulses from the cerebral cortex and brainstem as well as impulses from somatic sensory nerves also influence autonomic activity.

Types of neurons of efferent pathways

The efferent pathways of the autonomic nervous system involve two different neurons. The first, or preganglionic, neuron is situated in the brain or spinal cord. The preganglionic cell sends its axon to an autonomic ganglion where it synapses with a postganglionic cell, which in turn innervates the visceral organs.

There are three major outflows of preganglionic fibres from the central nervous system to the autonomic ganglia. The cranial outflow consists of fibres within some of the cranial nerves. The thoracolumbar outflow leaves the spinal cord in the thoracic and upper two lumbar nerves. The sacral outflow passes through the second, third, and fourth sacral nerves. The autonomic nervous system consists of the thoracolumbar division, which comprises the sympathetic system, and the cranial and sacral divisions, which together form the craniosacral, or parasympathetic, system.

The sympathetic division. The preganglionic fibres of the sympathetic division originate from neurons in the intermediolateral gray column of the spinal cord, leave the cord in the ventral roots, and pass to the sympathetic trunks, where some terminate. The sympathetic trunks are chains of ganglia close along either side of the spinal cord from the base of the skull to the coccyx.

Other preganglionic fibres from the spinal cord pass through the sympathetic ganglia as splanchnic nerves and end on the prevertebral, or collateral, ganglia, which are situated near the visceral branches of the aorta.

The actions of the sympathetic nervous system on visceral structures include dilation of the pupil of the eye, acceleration of the heart rate, increased force of contraction of the heart, and dilation of the coronary arteries. The blood vessels of the viscera and skin are constricted, blood pressure rises, and blood supply to the muscles, lungs, heart, and brain is increased. Respiration increases, and the bronchioles of the lung become dilated. Gastrointestinal motility is decreased, and secretion of the salivary and gastrointestinal glands is inhibited. The secretion of sweat glands is increased. The urinary and rectal sphincters contract. Thus, the sympathetic nervous system prepares the body for intense muscular activity necessary in offense and defense (the fight-or-flight reaction). Pain, exposure to cold, or strong emotions, such as rage or fear, evoke sympathetic activity that mobilizes the body resources for violent reactions. At the same time, the peaceful activities of the visceral structures are suspended.

The effects of the sympathetic nervous system are augmented by the medulla of the adrenal gland, whose cells, when stimulated, release epinephrine and norepinephrine into the bloodstream, thereby increasing the effects of impulses in the postganglionic sympathetic fibres and producing generalized rather than localized responses.

The parasympathetic division. In contrast to the sympathetic system, the preganglionic fibres of the parasympathetic system are long and terminate on ganglia located within or very near the organs they innervate. The postganglionic fibres, however, are very short.

The actions of the parasympathetic nervous system are largely antagonistic to those of the sympathetic. The parasympathetic neurons constrict the pupil of the eye, decrease the heart rate, stimulate the flow of saliva and gastric juice, and stimulate the peristaltic movements of the gastrointestinal tract. They relax the sphincters of the bladder and rectum and contract the muscles in the wall of the bladder allowing the body to rid itself of wastes. Thus, the parasympathetic system is concerned with restoring and conserving bodily energy and resting vital organs. Digestion and absorption of nutrients is promoted and the heart given longer periods of rest.

EMBRYONIC DEVELOPMENT OF THE VERTEBRATE NERVOUS SYSTEM

Development of the spinal cord and its nerves. The nervous system develops from the outer cell layer (ecto-

derm) of the early embryo (Figure 5). The ectoderm along the dorsal side of the embryo near a structure called the notochord becomes thickened to form the neural plate, which is broader over the anterior, or cranial, end. The neural plate soon folds in (invaginates) to form a shallow neural groove whose flanks are then called neural folds. The neural groove deepens while the neural folds approach each other over the groove and fuse to form the neural tube. Fusion occurs first over the future cervical region and then continues both headward and tailward. When fusion is completed, the central nervous system is represented by a closed tube with an expanded cranial portion, the future brain, and a narrow caudal portion, which becomes the spinal cord.

Formation of the neural tube

From L. Hyman, *Comparative Vertebrate Anatomy*, copyright © (1942); University of Chicago Press

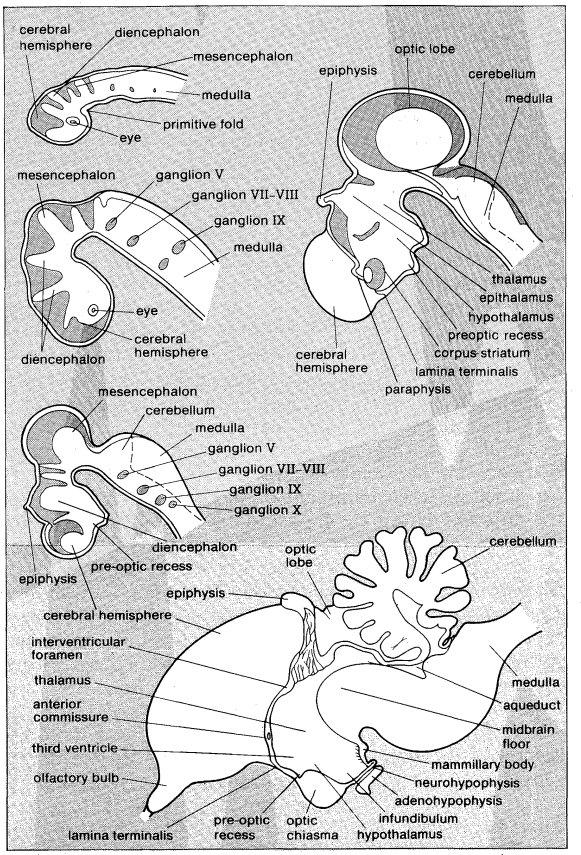


Figure 5: Stages in the early development of the chick brain and cranial nerves.

The lateral margins of the neural plate contain cells that will give rise to the neural crest. When the neural folds fuse, these cells come to lie between the ectoderm and the neural tube and later migrate to the sides of the neural tube. The neural crest gives rise to the sensory cells of the peripheral nervous system, peripheral cells of the autonomic nervous system, cells of the adrenal medulla, Schwann cells of peripheral nerves, pigment cells, and the pia mater and arachnoid membranes.

Undifferentiated stem cells, or germinal cells of the neuroepithelium, give rise to the neurons, neuroglia, and ependyma of the central nervous system. These dividing cells differentiate into medulloblasts and spongioblasts. The spongioblasts develop into ependymal cells and astrocytes. The medulloblasts give rise to astrocytes, oligodendroglia, and neuroblasts. The neuroblasts develop processes and become the neurons of the gray matter. The microglia are of middle layer (mesodermal) origin, perhaps developing from loosely-packed mesenchymal cells accompanying the blood vessels that grow into the central nervous system.

When the neural tube has closed, the spinal cord has thick lateral walls and thin roof and floor plates. A longitudinal groove, the sulcus limitans, running laterally on each side of the neural tube, divides it into ventral (motor) and dorsal (sensory) areas. Part of the lateral wall, called the mantle layer, represents the gray matter of the spinal cord. It thickens ventrally to form the basal plates, which will become the ventral horns of the adult spinal cord. Dorsal thickenings, or alar plates, will become the dorsal horns. The central cavity is reduced in size by fusion of its dorsal portion to form the posterior median septum. The ventral part of the cavity persists as the central canal of the spinal cord. The anterior median fissure is formed by an invagination in the floor plate. As cells differentiate in the mantle layer, they send their fibres into the marginal layer, which will become the white matter. At this point the neural tube has the basic configuration of the adult spinal cord.

The developing neuroblasts in the basal lamina of the developing cord send out processes that reach the developing muscle cells within the embryonic muscle blocks called myotomes. The fibres innervating each myotome are grouped in a bundle forming the ventral root. The neural crest cells give rise to the neuroblasts, which form the sensory ganglia. The proximal processes of these cells grow toward the spinal cord and form the dorsal root while the distal processes unite with the ventral root to form the spinal nerve. Neural-crest cells migrate along the outgrowing nerve processes and form the neurilemmal sheaths. Schwann cells form the myelin sheath present on some nerves. The sympathetic ganglia develop from neural-crest cells and from cells migrating from the spinal cord along the ventral roots. Some of the cells migrate further to form the prevertebral ganglia and the terminal autonomic ganglia in the walls of the viscera.

Development of the brain and its nerves. By the time the neural tube closes, three primary brain vesicles are apparent, the farthest forward being the prosencephalon, or forebrain, from which neural components of the eyes (optic vesicles) extend laterally. The intermediate dilation is the mesencephalon, or midbrain. The caudal dilation is the rhombencephalon, or hindbrain. The mesencephalon persists in the adult as the midbrain. The prosencephalon, however, gives rise forward to the telencephalon, or endbrain, and to a more caudal portion, the diencephalon. The telencephalon develops into the cerebral hemispheres, while the diencephalon carries the optic vesicles and includes the thalamus and hypothalamus. The rhombencephalon develops into the metencephalon and myelencephalon. The metencephalon becomes the pons and cerebellum, while the myelencephalon forms the medulla.

The three primary brain vesicles

The midbrain is separated from the hindbrain by a bend called the cephalic flexure. Farther back a cervical flexure appears at the junction of the brain and the spinal cord, and later a pontine flexure appears in a direction opposite to the cephalic and cervical flexures. Only the cephalic flexure persists in the adult, with the result that the cerebral hemispheres and diencephalon are folded downward at an angle to the rest of the brain axis.

The nerves develop in different ways. The olfactory nerve is formed by the axons of olfactory cells derived from neuroblasts in the nasal epithelium. The optic nerve arises from the stalk of the optic vesicle. The other nerves are formed according to their functional components. Fibres with a motor function arise from several nuclei appearing within the basal plate; sensory elements are derived from ganglia formed from neural-crest cells and in some cases from special ectodermal thickenings called placodes.

Although developing muscle cells of the embryo are capable of contracting spontaneously (myogenic activity) before they are innervated, as soon as a nerve fibre reaches the muscle cell, contraction occurs only in response to nerve impulses (neurogenic activity). At that time reflexes can take place in which the muscles respond to sensory stimulation. As the neurons differentiate and assume their functional capacities, tract systems are laid down so that the influences of higher centres are superimposed on the simple spinal reflexes. The tracts begin to function at the time the nerves become myelinated. At birth, reflex mechanisms controlling vital activities such

as respiration, suckling, and swallowing are well established. Control of movements by the cerebral cortex is not completed until the pyramidal tracts become myelinated after birth. After birth, cell division in the central nervous system stops and nerve cells that are lost thereafter cannot be replaced.

EVOLUTION OF THE VERTEBRATE NERVOUS SYSTEM

The primitive condition. The vertebrates constitute an advanced subdivision of the phylum Chordata. All chordates at some time in their life history have a rodlike bar called the notochord running the length of the body. Lower chordates (acorn worms, tunicates, and amphioxus), which lack a vertebral column, illustrate the most primitive features of the chordate nervous system. In these animals, the nerve cord is a rather uniform-appearing dorsally placed tube with a hollow cavity, which corresponds roughly to the spinal cord of the vertebrates, suggesting that the spinal cord is the most primitive component of the central nervous system (Figure 6).

The notochord

In amphioxus and in the lowest vertebrates such as lampreys, the sensory fibres and motor fibres leave the cord in dorsal and ventral roots to supply the adjacent body segments (myotomes). The dorsal and ventral roots remain separate nerves, and arise at alternate positions along the cord. In lower fishes, there is still alternation of dorsal and ventral roots, but the roots unite in a single spinal nerve. In higher vertebrates the two roots unite in a single spinal nerve and leave the cord at the same level, one above the other. Each spinal nerve supplies a single myotome. When appendages (fins, wings, arms, and legs) develop from several myotomes, the nerves continue to supply their original segments and branches of the spinal nerves become interwoven to form plexuses.

From A. Romer, *The Vertebrate Body*, 3rd ed. (1962); W.B. Saunders Company, Philadelphia

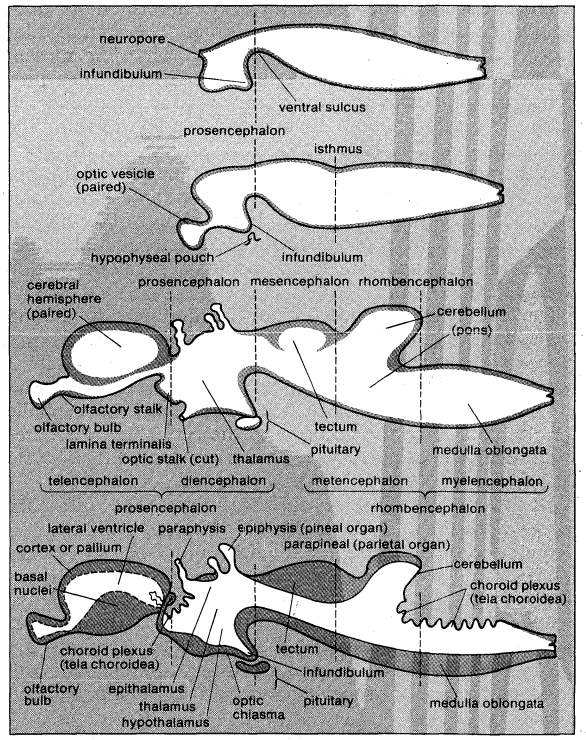


Figure 6: Development of principal vertebrate brain divisions.

The brain of vertebrates developed by the accumulation of nerve cells at the cephalic end of the nerve cord. At first, this diffuse collection of nerve cells regulated the reflex activity of spinal motor neurons. These cells are comparable to the reticular formation occupying the brainstem of higher vertebrates. The brainstem, thus, is the oldest portion of the brain in terms of evolution.

Encephalization. Early in the evolution of vertebrates, a special sensory system became associated with each major part of the brain: the olfactory organs with the forebrain, the eye with the midbrain, and ear and related organs with the hindbrain. Each of the three sections, furthermore, developed dorsal outgrowths of gray matter forming, respectively, the cerebrum, midbrain roof, or tectum, and cerebellum. With these developments, the three-part brainstem was then transformed into a brain of five regions: telencephalon, diencephalon, mesencephalon, metencephalon, and myelencephalon. The addition of these nerve centres to the primitive brainstem allowed greater coordination and association between the sensory and motor fibres.

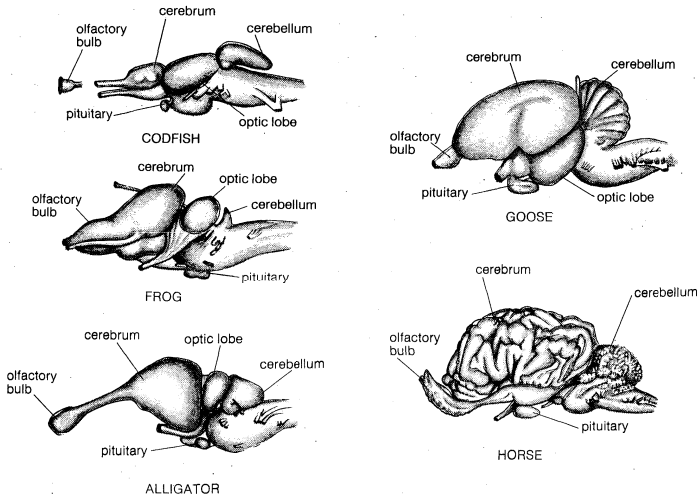


Figure 7: Representative series of vertebrate brains illustrating the evolutionary development of the brain.

From *Biological Sciences Curriculum Study Green Version High School Biology*, 2nd ed. (1968); Rand McNally & Co.

In tracing the development of the parts of the brain in the different vertebrate classes, some general features are apparent (Figure 7). There is a correlation between the size of a particular part of the brain and its importance in the life of the animal in question. Some neural structures (*e.g.*, olfactory bulb) have considerable size and importance in lower forms but are less conspicuous in higher forms. Progressing from the lowest to highest forms, there is a gradual cephalic shift of function from the lower brainstem to the higher cerebral cortex.

The hindbrain is comparable to an enlarged, anterior section of the spinal cord. In the gray matter are dorsal sensory and ventral motor columns similar to those present in the cord. The longitudinal continuity of these columns is preserved in the lower vertebrates, but in higher vertebrates the columns break up into discrete nuclei that serve some of the cranial nerves. The hindbrain exerts control over the spinal motor neurons through the reticular formation. Fish and tailed amphibians, in addition, have a pair of giant cells called the cells of Mauthner, which exert some control over the local spinal-cord reflexes responsible for the rhythmic swimming undulations of these animals.

The hindbrain is the area of reception of one of the main sensory systems, the acoustico-lateralis system, which consists of the ear (hearing and equilibrium) and the lateral-line organs (vibration and pressure). The latter, situated in rows along the head and body, are retained in fish but disappear in the land vertebrates.

Origin of cerebellum

The cerebellum originated as a specialized part of the acoustico-lateralis area. The oldest part of the cerebellum —the archicerebellum—is concerned with equilibrium and connected with the inner ear and the lateral-line system. The anterior lobe of the cerebellum represents the paleocerebellum, an area that regulates equilibrium and muscle tone; it constitutes the main mass of the cerebellum in fish, reptiles, and birds. In mammals, the development of the cerebral cortex and its connections with the cerebellum are correlated with the appearance of the large cerebellar hemispheres. This new part of the cerebellum, or neocerebellum, coordinates skilled movements initiated at cortical levels. In mammals a great mass of fibres connects the brainstem to the cerebellum; this re-

gion forms the pons, which, together with the cerebellum, constitutes the metencephalon. The caudal part of the hindbrain remains as the medulla (myelencephalon).

The midbrain (mesencephalon) and diencephalon comprise the anterior portion of the brainstem. Sensory and motor nuclei for cranial nerves extend from the hindbrain to the midbrain. The roof of the midbrain, or tectum, developed as the primary visual centre. The optic lobes, especially prominent in fish and birds, are a part of this area. In fish and amphibians, the tectum is the major centre of the nervous system and wields the greatest influence on body activity. While this area is still significant in reptiles and birds, it is supplanted in importance by the cerebral hemispheres. In mammals, the midbrain is greatly reduced in importance. Instead, most of the optic sensations are relayed to the cerebral cortex. With development of the cerebral cortex, the thalamus becomes less significant as an association area and more important as a relay centre for sensory impulses. Centres for visceral sensations and visceral motor responses become established in the hypothalamus.

Dominance of the cerebrum. Ascending the vertebrate scale, the cerebral hemispheres become more and more important as association centres (Figure 8). The cerebral hemispheres begin their development as paired

From A. Romer, *The Vertebrate Body*, 3rd ed. (1962); W.B. Saunders Company, Philadelphia

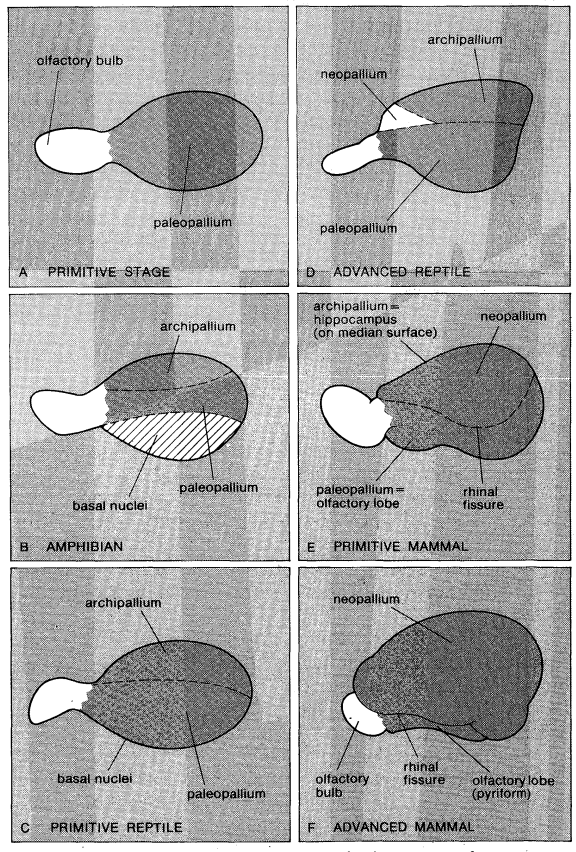


Figure 8: Differentiation of the cerebrum in the course of evolution.

The cerebral hemispheres as olfactory lobes

outgrowths of the forebrain and serve as centres of olfactory reception. In the lowest vertebrates, the forebrain is divided into the olfactory bulb—where the olfactory nerve fibres end—and the cerebral hemisphere. The hemisphere at this time, referred to as the paleopallium, is merely an olfactory lobe serving as an association area for olfactory impulses. The olfactory lobes are very prominent in animals such as amphibians, but in birds and primates in which sense of smell is less important, the lobes are reduced in extent. In amphibians, the hemispheres consist of three parts: the paleopallium (olfactory lobe), archipallium, and basal nuclei. These areas of gray matter are located deep to the surface. All three areas receive olfactory stimuli and discharge impulses to the brainstem. The archipallium is a correlation centre and forerunner of the mammalian hippocampus. The basal nuclei are equivalent to the corpus striatum and function as association areas with connections with the thalamus. In primitive reptiles, the basal nuclei have moved to the inner part of the hemisphere, whereas the other areas of gray matter have moved toward the surface. In advanced reptiles, a new association centre, the neopallium, appears between paleopallium and archipallium. In birds, there is nothing corresponding to the neopallium, but the basal ganglia (corpus striatum) are enormously expanded. In mammals, the neopallium becomes greatly enlarged to exceed all the other parts of the brain in size. This region assumes more and more of the higher types of neural activity in correlation, association, and learning. At first, the neopallium expands to envelop the other brain structures. The archipallium becomes folded into a small area on the median part of the hemispheres, where it remains as the hippocampus. The paleopallium (olfactory lobes) constitute a small ventral region on the hemisphere, the pyriform lobe. The corpus striatum (old basal nuclei) becomes a central part of the hemisphere. Further expansion of the neopallium in primates and man causes extensive folding and results in a very convoluted surface of the brain.

V. Biodynamics of the vertebrate nervous system

CONTROLS EXERTED ON THE NERVOUS SYSTEM

Intrinsic control. In the normal animal, a variety of influences constantly impinge on the nervous system and affect its activity. Stimuli from many external sources travelling over sensory fibres affect the activity of central neurons. Other external influences reaching the nervous system through the bloodstream may act directly on nerve cells. In addition, there are nervous activities not immediately referable to external sources but apparently arising within the nervous system itself. The number and diversity of influences on neural activity might be expected to elicit a variety of conflicting reflexes and behaviour patterns. What is remarkable about the normal nervous system, though, is that from moment to moment it sorts out various activities and arrives at a singular action.

Regulation in the nervous system is achieved by means of its integrative capabilities, which allow the nervous system to regulate its output in response to its input. Integration takes place at different levels within the nervous system. First, it can occur at the level of the neuron, which has several forms of activity that can be coupled, permitting the output of the neuron to be a complex function of its input. The input may either decrease the probability of output (inhibition) or increase it (excitation).

Secondly, integration occurs between organized groups of neurons. The central nervous system contains many specific fibre tracts and groups of neurons that are connected in different ways. One input system may lead to the integrative group of neurons, and one output line leads away. Several input lines may converge on one output, or several outputs can diverge from one input. Associative components, or interneurons, may be situated along the afferent or efferent pathways. The interconnections between these pathways allow for limitless processing of the nerve impulses. The systems can interact through excitatory or inhibitory tendencies or through alteration of frequency of impulses to transform the nervous signal. These influences are finally exerted on the motor neurons to the skeletal muscles and to the visceral structures; the motor neuron, thus, serves as the last neural link, or final common pathway.

Unification of behaviour

The highest orders of integration occur at the level of the organism and can be deduced from the study of behaviour. At any one time, the behaviour of an organism shows a singleness of purpose. Alternative courses of action are excluded in favour of unified behaviour. This function of focussing behaviour is carried out by the reticular activating system, a part of the reticular formation in the central core of the brainstem. The reticular activating system is responsible for arousal, wakefulness, and alertness of the cerebral cortex, permitting the cortex

to carry out its special functions of sensation, volition, and thought. It focusses action and unifies efforts by confining the direction of attention and galvanizing the specific system into action.

Extrinsic influences. Besides stimuli that reach the central nervous system over afferent nerves, many other extrinsic factors influence the nerve cells directly. These influences include nutrients, metabolites (chemical products of metabolism), hormones, and drugs. Most of these substances reach the nervous system through the bloodstream. First, though, they must cross the blood–brain barrier, which plays a major role in maintaining the environment necessary for proper functioning of neurons. Some substances, such as particulate materials, certain dyes, pigments, metals, large molecules, and bacteria, cannot enter the spaces between cells in the brain, although they can move between the cells in other tissues of the body. The actual barrier to these materials seems to be the brain capillaries, which are different physiologically from other capillaries.

The brain has an enormous oxygen requirement, necessary for the utilization of the sugar glucose, its major source of fuel. Therefore, any conditions that reduce the oxygen supply or lower the level of glucose in the blood affect the nervous system and may lead to seizures and coma.

Since the functioning of neurons depends in part on the movement of ions, it is not surprising that disturbances of ion levels will affect the nervous system.

The nutritional and metabolic state of the body also affects the nervous system. Deficiencies of the B vitamins have the most serious effect on the nervous system, causing widespread degeneration of peripheral nerves, damage to neurons in the brain, and degeneration of spinal cord tracts.

The physical state of the body, of course, plays a major role in the functioning of the nervous system. The well-being of the nervous system depends on the proper functioning of other organ systems including the circulatory, respiratory, digestive, and excretory systems. Disease, malfunction, or aging of any of these other systems may adversely affect the nervous system by failing to provide necessary oxygen and energy supplies, to maintain the proper environment for nervous activity, and to remove toxic wastes. Neurons are lost progressively as an organism ages, but because these cells do not divide after birth, there is no capacity for replacement.

The varied effects of drugs

Many drugs have an effect on the nervous system resulting in changes in behaviour, perception, thought, emotions, and other activities of the nervous system. These drugs—called neuropharmacological agents—may exert their influence by acting directly on nerve cells or by affecting the neurotransmitters. Since neurotransmitters are responsible for the exchange of information between nerve cells, any drugs that influence the synthesis or breakdown of these transmitters or that mimic their effects will have an influence on nervous activity. Drugs that have the same effects as acetylcholine are known as cholinergic drugs; they include agents that inhibit acetylcholinesterase, the enzyme that destroys acetylcholine. Other drugs, such as atropine and curare, block the effects of acetylcholine. Drugs that have the same effects as norepinephrine and epinephrine are called adrenergic drugs. Adrenergic blocking agents (*e.g.*, ergotamine) inhibit the effects of norepinephrine and epinephrine and, thus, oppose the effects of sympathetic stimulation.

A number of drugs seem to exert their major effects on the central nervous system. Some, such as alcohol, have a depressant effect on the nervous system through depression of the autonomic circulatory and respiratory centres in the brainstem. Barbiturates, like alcohol, depress the higher cerebral centres first and with increasing doses affect lower subcortical structures. Drugs such as opium and morphine appear to elevate the pain threshold and to alter the reaction to pain. These drugs are also addictive and have depressive effects on the central nervous system. Anesthetics cause central nervous depression ultimately leading to unconsciousness. Some of their effects seem to be exerted on the reticular activating system, which maintains wakefulness. Other drugs elevate mood and stimulate behaviour and are useful in the treatment of depressive states.

Tranquillizers are drugs that reduce anxiety and agitation without clouding the consciousness. These agents exert their action on subcortical structures including the reticular system, hypothalamus, and limbic system. Some drugs are capable of producing changes in perception, thinking, and emotion, causing hallucinations, illusions, and delusions. These drugs, called hallucinogens, include among them lysergic acid diethylamide (LSD) and related compounds.

INTERACTION BETWEEN NERVOUS AND ENDOCRINE SYSTEMS

The neurosecretory cell and the concept of neurosecretion. A special class of neurons called neurosecretory cells have been identified in all groups of higher animals. These cells engage in secretory activity to a degree above and beyond that observed in ordinary nerve cells. They have features characteristic of both neurons and endocrine cells. Like ordinary neurons, they respond to stimuli, they conduct action potentials, and they contain synaptic vesicles. Like endocrine cells, they produce hormonal substances (neurohormones) that act over a distance. These secretory products, however, are usually proteins or polypeptides, unlike the neurotransmitter substances produced by ordinary nerve cells.

Neurohormones

These neurohormones are synthesized in the cell body of the neurosecretory neuron and packaged in neurosecretory granules. The neurosecretory granules travel from the cell body down the axon and are stored in the nerve terminal. Neurosecretory cell terminals are situated near blood vessels; the combinations of neurosecretory cells with vascular channels are known as neurohemal organs. The corpus cardiacum of insects and the posterior lobe of the pituitary of mammals are examples of such storage depots. Upon appropriate stimulation, the neurosecretory cells release their secretory products into the vascular system. The neurohormones then travel in the blood and affect distant structures such as endocrine glands or other tissues.

Neuroendocrine interactions. The functional significance of the neurosecretory cell is that it acts as a mediator between the nervous and endocrine systems. The nervous system is involved in point-to-point transmission of signals with very short duration (fractions of a second). The endocrine system, on the other hand, is concerned with much slower and more diffuse transmission of messages. The significance of the neurosecretory cell is that it alone serves as a singular channel between the nervous and endocrine systems, receiving messages in neural language and transmitting this information in endocrine language.

In all cases afferent stimuli are channelled to the neurosecretory cells and influence the release of neurohormones. Certain neurosecretory hormones are actually factors that effect the release of other hormones by the anterior pituitary. Five of these factors cause the release of specific anterior pituitary hormones (follicle-stimulating hormone, luteinizing hormone, growth hormone, adrenocorticotropic hormone, and thyroid-stimulating hormone), and one inhibits the release of a hormone (luteotropic hormone, or prolactin). These hormones in turn act on their respective endocrine glands. The endocrine organs then release hormones that influence the target organs and tissues. Other neurosecretory hormones directly affect their target organs.

Neurosecretory cell bodies located in certain nuclei of the hypothalamus send axons to the neural lobe of the posterior pituitary and terminate near blood vessels. These neurosecretory cells elaborate two hormones, vasopressin, or antidiuretic hormone, and oxytocin. Vasopressin produces a generalized constriction of blood vessels, producing a rise in blood pressure. It also stimulates the kidney tubules to retain water and lose salt in the urine. Oxytocin causes contraction of the uterus and the flow of milk from cells in the mammary gland (see also the article ENDOCRINE SYSTEMS).

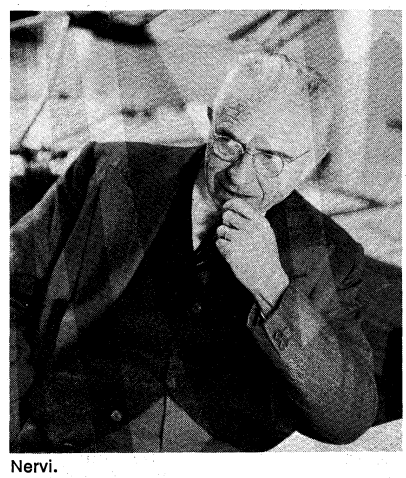
BIBLIOGRAPHY. I. ASIMOV, *The Human Brain: Its Capacities and Functions* (1963), a popular discussion of the structure and operation of the human brain and the nervous system in general; M.A.B. BRAZIER, *The Electrical Activity of the Nervous System*, 3rd ed. (1968), a college-level survey of the electrical activities of the nervous system; T.H. BULLOCK and G.A. HORRIDGE, *Structure and Function in the Nervous Systems of Invertebrates*, 2 vol. (1965), a monumental work, well-illustrated and containing comprehensive bibliographies, covering the available knowledge of the nervous systems of all the invertebrate groups; E.C. CROSBY, T. HUMPHREY, and E.W. LAUER, *Correlative Anatomy of the Nervous System* (1962), a textbook of neuroanatomy correlating morphology of the nervous system with neurophysiology, neuropharmacology, neurosurgery, and neurology; J.C. ECCLES, *The Physiology of Nerve Cells* (1957), a book written by a co-recipient of the 1963 Nobel Prize for Medicine, and concerned with the electrical properties of nerve cells and the transmission of impulses (available in paperback); H.C. ELLIOT, *Textbook of Neuroanatomy*, 2nd ed. (1969), a textbook for the college or medical student, containing a complete account of the structure and function of the human nervous system; D.H. FORD and J.P. SCHADE, *Atlas of the Human Brain* (1966), photographs of dissections of the brain; M. GABE, *Neurosecretion* (1966), a complete account of neurosecretion in invertebrates and vertebrates; E. GARDNER, *Fundamentals of Neurology*, 4th ed. (1963), a short textbook presenting the anatomical and physiological bases of neurology; C.J. HERRICK, *Brains of Rats and Men* (1926, reprinted 1963), a work still useful in presenting problems of human and animal behaviour to both scientist and layman; H. KUHLENBECK, *The Central Nervous System of Vertebrates* (1967–), a projected five-volume series to include basic concepts and a general survey of the nervous systems of both invertebrates and vertebrates; T.L. LENTZ, *Primitive Nervous Systems* (1968), an account of the neural organization of sponges, hydra, and planaria, with a hypothesis on the evolutionary origin of the nervous system; B.W. MCCASHLAND, *Animal Coordinating Mechanisms* (1968), contains the basic principles of nervous and hormonal coordination in animals, written for the beginning biology student; J.W. MILLEN and D.H.M. WOOLAM, *The Anatomy of the Cerebrospinal Fluid* (1962), a description of the anatomy and function of the ventricles of the brain, the choroid plexuses, and meninges; P.W. NATHAN, *The Nervous System* (1969), a complete account of the anatomy, physiology, and psychology of the nervous system of man and other animals, written for readers without an extensive background in biology; F.H. NETTER, *The CIBA Collection of Medical Illustrations*, vol. 1, *Nervous System* (1958), a magnificent collection of colour drawings of the central nervous system, spine, nerve tracts, autonomic nervous system, and diseases of the brain and spinal cord; G.H. PARKER, *The Elementary Nervous System* (1919), a classic book on the origin of the basic receptor–adjustor–effector system of neural function; A. PETERS, S.L. PALAY, and H. DE F. WEBSTER, *The Fine Structure of the Nervous System: The Cells and Their Processes* (1970), a description of nerve cells based on numerous electron micrographs; G.C. QUARTON, T. MELNECHUK, and F.O. SCHMITT (eds.), *The Neurosciences: A Study Program* (1967), an important, advanced text containing contributions by many distinguished scientists on a wide range of topics in neurobiology; C.S. SHERRINGTON, *The Integrative Action of the Nervous System* (1906, reprinted 1947), a classic on the physiology of reflex mechanisms (also available in paperback); R.C. TRUEX and M.B. CARPENTER, *Human Neuroanatomy*, 6th ed. (1969), a complete account of the structure and function of the human nervous system written for the medical student; J.R. WILSON, *The Mind* (1964), a well-illustrated introductory book from the Life Science Library, covering the anatomy of the nervous system, learning, intelligence, and mental illness; L.S. WOODBURNE, *The Neural Basis of Behavior* (1967), a useful review of the anatomy and physiology of the nervous system for the beginning student in biology, medicine, or psychology.

(T.L.L.)

Nervi, Pier Luigi

An Italian engineer-architect, Pier Luigi Nervi is the paradigm of the master builder of modern times. His work expresses a comprehensive understanding and knowledge of design, structure, and construction techniques with a profound understanding of social and economic processes. His audacious, imaginative structures typically are built of reinforced concrete, either cast in place or prefabricated on the site, in forms that span immense distances efficiently and with surpassing beauty. Although his initial interest in building was as a structural engineer and he never designed primarily for aesthetic effect, his works were soon recognized as having transcended engineering and achieved the status of architecture of the higher order.

Henri Cartier-Bresson—Magnum



Nervi.

Nervi was born in Sondrio, Italy, on June 21, 1891. He graduated from the University of Bologna in 1913 and soon afterward World War I began. Nervi served as a lieutenant in the Corps of Engineers of the Italian Army and, after the war, worked as an engineer in Bologna and Florence. In 1926–27 he designed his first significant work, a cinema in Naples, and followed it with the municipal stadium (Berta Stadium) in Florence, built in 1930–32.

Formation of Nervi and Bartoli firm

In 1932 Nervi and a cousin in Rome formed the contracting firm of Nervi and Bartoli, with which he would remain through the rest of his career. In 1935 the Italian Air Force held a competition for a series of hangars to be built throughout Italy. Nervi conceived them as concrete vaults, with huge spans, that could be constructed at low cost, and he won the commission. Between 1935 and 1941 he built hangars in Obertello, Orvieto, and Torre del Lago. All of these structures were destroyed during World War II.

Each of these early structures showed the growth of conceptual design that resulted from Nervi's ceaseless search for new solutions to structural problems. His creativity was not confined to the design of buildings; during World War II, he attempted to construct vessels for the Italian Navy made of concrete, but the project was not completed. After the war, he did succeed in building a 165-ton, motor-powered, concrete sailboat, with a hull 1.4 inches (3.6 centimetres) thick. It was named "Irene" for his wife. Subsequently he built a 38-foot (11.6-metre) ketch, the "Nennele," with a hull only a half inch (1 centimetre) thick. For both of these vessels, he employed ferrocemento, a material of his own invention, composed of dense concrete, heavily reinforced with evenly distributed steel mesh that gives it both lightness and strength.

This material was vital in Nervi's design for a complex he built for the Turin Exhibition in 1949–50. It is a prefabricated structure in the form of a corrugated cylindrical arch, spanning 309 feet (93 metres), based on modular components of glass and ferrocemento. Without the structural properties of this material, the entire conception of the complex would not have been feasible. In this, as in his other works, there is an intimate relationship between conception and execution.

A close relationship is evident between Nervi's work and his austere life. Both have been called humble but full of passion, economical of means, yet extremely warm. His solutions to building problems have always been direct, transmitting to the ground the stresses developed within the structures through the shortest path. His works have been relatively unaffected by the changes in taste that have accompanied the advent of new forms in architecture. As a professor at the University of Rome from

1947, Nervi taught that a designer can develop truthful solutions in three ways: by understanding the pure harmony of the laws of the physical world that regulate the equilibrium of forces and the resistance of materials; by honestly interpreting the essential factors of each problem; and by rejecting the limitations of the solutions of the past.

Design of UNESCO headquarters

In 1950, when UNESCO decided to build its new headquarters in Paris, Nervi was one of the architects selected to design it. The Hungarian-born U.S. architect Marcel Breuer, one of his collaborators, described Nervi's participation in the project as "a continuous search for a system: a system of geometric rhythm," and later he said of the man: "If there is a notion that arrogance and reckless irresponsibility are the very attributes of genius, a notion that to be a genius means not to be quite human, there is Nervi to disprove this notion."

In 1955, in association with a group of architects, Nervi helped design the first skyscraper in Italy, the Pirelli Building in Milan; it was the first office building to use a long-span structure—80 feet (25 metres). Although architects and engineers in the United States had long experience in the design and construction of skyscrapers, they had invariably designed them around frameworks consisting of series of smaller spans. For the Pirelli Building, Nervi used experimental models—as he often had—which he tested in the famous laboratory at Bergamo. His second skyscraper was built in Montreal, again in collaboration, and his third was Australia Square (1962–69; Sydney), a cylindrical tower of 50 stories. At the time, this was the highest concrete structure in the world. In 1957 and 1958–60, for the 1960 Rome Olympic Games, Nervi designed two sport palaces.

Work in the U.S.

His first building in the United States was commissioned by the Port of New York Authority: the George Washington Bridge Bus Terminal, in Manhattan, built in 1961–62. Subsequently he designed a precast, vaulted Field House for Dartmouth College in New Hampshire (1961–62) and, in collaboration with Pietro Belluschi, the Cathedral of San Francisco, four vertical-warped surfaces dramatically enclosing the vertical space of the main nave. In 1961 Harvard University appointed Nervi to the Charles Eliot Norton Chair of Poetry and awarded him, in 1963, an honorary degree; he later received the Gold Medal from the American Institute of Architects in recognition of his work.

After years of intense practice in Italy and abroad, Nervi was obliged to diminish his activities as a builder in the late 1960s. Assisted by two of his sons, Antonio, a structural engineer, and Mario, an architect, he began to confine his activities largely to designs. An increasing number of his projects now were done in association with foreign architects.

Although Nervi's primary concern has never been aesthetic, many of his works, nonetheless, reach the realm of poetry. Nervi's buildings achieve remarkable expressive force, as in the geometry of the slabs in the Gatti wool factory (1953), in Rome, and the mezzanine of the Palace of Labour, in Turin. Through his use of interpenetrating planes, of folded and bent plates, and of warping surfaces, Nervi has introduced a new three-dimensional vocabulary into architectural design. He reminded architects that "materials, statics, the technology of construction, economic efficiency and functional needs, are the vocabulary of the architectural speech."

Nervi's contribution has been compared to that of another builder whose work revolutionized architecture—Joseph Paxton, who built the Crystal Palace for the International Exhibition of 1851 in London. In both instances, highly rational and innovative structures resulted from a continuous process of devoted search and development, with an emphasis on modular construction, prefabrication, and extreme physical and visual lightness.

MAJOR WORKS

Berta Stadium, Florence (1930–32); Hangar for the Italian Air Force, Orvieto (1936); Exhibition Hall, Turin, Italy; (Salone B, 1948–49; Salone C, 1950); Kursaal Beach Casino roof, Ostia, Italy (1950); Sala delle feste delle terme, Chianciano, Italy (1952); UNESCO Headquarters, Paris—with Marcel Breuer and Bernard Zehruss (1953–58); Pirelli Building, Milan (1955–59); Palazzetto dello Sport, Rome (1957); Flaminio Stadium, Rome (1958–59); Palazzo del Lavoro (Palace of Labor), Turin, Italy; (1961); George Washington Bridge Bus Terminal, New York (1961-62); Place Victoria, Montreal (1962–65); Cathedral of San Francisco (1970); Coliseum, Richmond, Virginia (1971); Auditorium, Vatican City (1971).

BIBLIOGRAPHY. PIER LUIGI NERVI, *Costruire correttamente* (1955; Eng. trans., *Structures*, 1956), Nervi explains his work and construction techniques; *The Works of Pier Luigi Nervi* (1957); *Nuove Strutture* (1963; Eng. trans., *Buildings, Projects, Structures, 1953–63*, 1963; British title, *New Structures*, 1963); *Pier Luigi Nervi: Space and Structural Integrity* (1961), an exhibition catalog of the San Francisco Museum of Art—all contain a complete graphic presentation of Nervi's work, with quotations from him and comments by the author or editor; ADA LOUISE HUXTABLE, *Pier Luigi Nervi* (1960), describes Nervi's work from a positive critical approach, combined with personal notes; *Aesthetics and Technology in Building* (1965), based on Nervi's Charles Eliot Norton lectures given at Harvard University, 1961–62; AGNOLDOMENICO PICA, *Pier Luigi Nervi* (1969), a written (in Italian) and graphic exposition of Nervi's work up to 1968; and "Moretti and Nervi's Place Victoria," *Archit. Rec.*, 139:141–146 (1966); D. CANTY, "Nervi's Gilded Gateway," *Archit. Forum*, 126:68–71 (1967); R. BOYD, "Australia Square," *ibid.*, 130: 26–35 (1969), a description of an office building by Nervi in Sidney, Australia.

(E.F.C.)

Nervous System, Human

Man's nervous system, like those of other vertebrates, includes such neural structures as the brain, spinal cord, and nerves; these serve to receive stimuli (input), to process and store their effects, and to generate behaviour (output). Abnormal development or injury in the human nervous system can lead to difficulties that range from the grossest defects (*e.g.*, crippling, idiocy, and death) to the subtlest psychiatric symptoms (*e.g.*, transient irritability, suspiciousness, and forgetfulness).

The intimate and overriding role of neural tissues in the functions of all animals, including human beings, is reflected in many other articles, including NERVE IMPULSE; NERVES AND NERVOUS SYSTEMS; and NERVOUS SYSTEM DISEASES.

This article is divided into the following sections:

I. Fundamental considerations
- Origin and development
- Characteristic structures
- Nerve degeneration and regeneration

II. The central nervous system
- The brain
- The spinal cord

III. The peripheral nervous system
- Cranial nerves
- Spinal nerves

IV. The autonomic nervous system
- General organization and function
- Levels of structural organization
- Early development
- Comparative features
- Autonomic transmission
- General and integrative functions of the autonomic nervous system levels
- Pathophysiology

I. Fundamental considerations

ORIGIN AND DEVELOPMENT

The human nervous system appears to have emerged from millions of years of prevertebrate and lower vertebrate evolution. Neural elements that have survived the evolutionary process tend to be those that enhanced the survival of successions of organisms during this long period. Several aspects of evolutionary significance may be inferred from the developing human embryo.

The origin of both the human nervous system and part of the skin from the ectoderm (the outer layer of cells of the early, ball-shaped embryo) may be related to the eventual location of most sensory structures on or close to the body surface where they can receive stimulation from the external environment. The earliest nerve cell, or neuron, to evolve was probably a combined receptor (re-

ceiving environmental stimuli) and motor unit (producing muscle or gland responses). Such primordial types of neuron are still to be found in living sponges. In subsequent evolutionary stages, specialized receptor neurons and the effector (motor) nerve cells seem to have emerged, such sequences of cells being observable in living sea anemones. In more highly evolved stages, greater complexity derives from the interposition (intercalation) of one or more nerve cells between the receptor and effector neurons. These intercalated cells (interneurons, or internuncial neurons) are organized in circuits of various degrees of complexity as in the central nervous system (brain and spinal cord). In this theoretical reconstruction of evolution, the primordial nervous system, composed of neurons on the body surface, evolves into a central nervous system that interacts with the external environment through the peripheral nerves.

Neurons in primitive life forms

In early embryonic stages, the presumptive neural part of the ectoderm in man and other vertebrates detaches from the part destined to become skin, sinks beneath the surface, and becomes surrounded by another embryonic layer (mesoderm). During subsequent development of the embryo, many sensory neurons differentiate and grow until their receptor fibres terminate in or near such ectodermally derived structures as the outer skin (epidermis) and the mucous membranes of the eye, ear, and nose.

All known vertebrates, including the most primitive known fossils (the ancient ancestors of jawless, eel-like fishes), have complex nervous systems. Study of inner skull surfaces (cranial endocasts) of many fossil vertebrate remains indicates that in all of them the brains and cranial nerves are built on a basically similar plan, which includes a specialized brain (encephalization). The vertebrate nervous system is bilaterally symmetrical, most neural structures being repeated on the left and right sides of the body.

Differentiation

The human nervous system begins to differentiate structurally and to show clear signs of function early during prenatal life (see also EMBRYOLOGY, HUMAN). By 18 days after fertilization, the ectoderm thickens along the future dorsal (backside) midline of the body to form the neural plate, and, slightly later, the primordial eye, ear, and nose differentiate. The neural plate elongates, and its lateral edges rise to form the neural folds, which, in turn, extend and unite the midline to close off the neural tube from which the central nervous system develops. The space within the neural tube becomes the cavities (ventricles) of the brain and the central canal of the spinal cord. This developmental sequence produces the dorsally located, hollow central nervous system unique to vertebrates. Continuing to differentiate, the neural tube detaches from the skin ectoderm and sinks beneath the surface. During this process, an ectodermal column of cells on each side differentiates, becoming the neural crests that segregate into future parts (*e.g.*, dorsal-root ganglia) of the spinal nerves. The cephalic (head) end of the tube differentiates into primary brain vesicles called the forebrain, midbrain, and hindbrain. This is the morphological status of the nervous system at the end of the first month of pregnancy.

The cells of the central nervous system originate from the matrix (ependymal) layer of the neural tube—that is, the layer of cells that lines the cavity of the neural tube. The matrix cells differentiate and proliferate into specialized cells (neuroblasts), the progenitors of the neurons of the central nervous system, and other cells (glioblasts) from which associated non-neural cells, called glia, develop. With few exceptions, the neuroblasts, glioblasts, and their derived cells do not divide and multiply (as do most other cells) once they migrate from the matrix layer into the gray and white matter of the nervous system.

Prenatal brain growth

During prenatal life, the slender primordial brain of the neural-tube stage gradually differentiates into the globular-shaped brain by middle fetal life. During this interval the brain continues to grow differentially in size and to bend at three major flexures. The main outlines of the form of the brain are recognizable by the end of the third fetal month. Its typically adult size and shape, however, are not attained until early puberty.

The early development of the "contorted" brain from the neural-tube stage is the product of the formation of three flexures (midbrain, pons, and cervical); the differential enlargement of various regions, especially the cerebrum and cerebellum; the growth of the cerebral hemispheres over and at the sides of the midbrain and cerebellum; and the formation of convolutions ("valleys" called sulci and "hills" called gyri) in the cerebral cortex and of folia (leaflike gyri) in the cerebellar cortex. The forebrain differentiates into such structures (described in greater detail below) as the cerebral cortex, corpus callosum, lateral and third ventricles, basal ganglia, thalamus, and hypothalamus; the midbrain becomes the cerebral peduncles, superior and inferior colliculi, and cerebral aqueduct; and the hindbrain generates the cerebellum, pons, medulla, and fourth ventricle. Such deeper valleys as the central and calcarine sulci are discernible by the fifth fetal month, and every major gyrus and sulcus is normally present by the seventh. The minor sulci and gyri are not fully formed until after birth. The relative size and proportions of the brain and its parts are basically similar to those in the adult brain by two years after birth.

The nervous system exhibits observable activity within the second prenatal month. Seven weeks after fertilization, the head draws away from a stimulus applied to the upper lip. Many reflexes of the head, trunk, and extremities can be evoked from the embryo during the third month of pregnancy.

Postnatal growth of the brain

The postnatal growth of the human brain is rapid and massive, especially during the first two years. The typical brain weighs 350 grams (about 0.8 pound) in the full-term infant at birth, 1,000 grams (2.2 pounds) at the end of the first year, and about 1,300 grams (2.9 pounds) at puberty. This increase is mainly attributable to growth of pre-existing neurons, new glial cells, and myelination (development of fatty insulating sheaths on neuron fibres). This unique threefold growth of the brain during the first year may be interpreted as an adaptation that is essential to man's survival as a species with a large brain. Birth occurs at a developmental stage when the infant is not so helpless that it is unable to survive, yet is still small enough to pass through the bony opening of the maternal pelvis; if the brain were much larger (enough to support intelligent behaviour), normal delivery might not be possible.

The neurons and the non-neural glial cells of the central nervous system differentiate at various times during the development of the individual. With a few uncertain exceptions, all neurons have differentiated from the matrix layer of the neural tube before birth. This effectively means that a person is born with as many neurons as he will ever have (perhaps 10,000,000,000 in the brain cortex alone), losing the potential function of more and more of them from then on. According to one estimate, an average of approximately 50,000 neurons atrophy or deteriorate each day from ages 20 to 70. In a healthy person, in 50 years, less than 10 percent of the original neuronal complement is out of action. Indeed, by age 75, the weight of the brain typically is reduced from its maximum at maturity by about one-tenth; the flow of blood through the brain drops by about one-fifth; and the functional taste buds in the mouth are about two-thirds gone (see also AGING, HUMAN).

The early maturation of neurons makes the first years of life crucial to the level of intelligence that is exhibited at the adult stage. Impairment of normal development of neurons in childhood may result in irreversible abnormality. Thus, severe mental retardation in a genetic disorder called phenylketonuria (PKU) results from a specific enzyme deficiency that prevents sufferers from normally metabolizing the amino acid phenylalanine. Young children with PKU who are fed on a phenylalanine-deficient diet often are able to escape the consequences of this inborn defect. Because neural tissues incorporate amino acids, however, young children reared on diets that are too low in protein fail to develop normal brain structures. The intelligence of many hundreds of millions of the world's potentially normal children may continue to be

impaired as long as so many (about 60 percent) of them suffer protein-deficient diets.

Development of the spinal cord and spinal nerves

The spinal cord and spinal nerves develop from the neural tube and neural crest. Neuroblasts migrate from the matrix layer of the neural tube and differentiate into the neurons of the spinal cord. Some parts (cell bodies and dendrites; see below) of these neurons remain within the gray matter. Other parts, the axons (fibrils or axis cylinders), differentiate, grow, and terminate in a variety of centres within the spinal cord and brain. From some spinal neurons, axons, the basic elements of white matter, elongate, pass out of the spinal cord, and form spinal nerves before terminating in such structures as voluntary muscles (*e.g.*, those that move the limbs). Some neuroblasts migrate from the spinal cord to differentiate into neurons of the sympathetic nervous system (which involves the spinal nerves of the midspine from chest to lower back). Spongioblasts (*i.e.*, glioblasts) migrate from the matrix layer and differentiate into different kinds of glial cells (astroglia and oligodendroglia) found in both white and gray matter.

The neuroblasts within the neural crest differentiate into sensory relay assemblies of neurons (ganglionic cells) in the spinal nerves. Each of these ganglionic cells has a peripheral fibril that extends to a sensory receptor (*e.g.*, in the skin) and a central axon that terminates in the gray matter of the spinal cord. Other crest cells differentiate into such specialized structures as the neurilemma cells, which generate the myelin sheaths that insulate many nerve fibres. Mesodermal cells develop into connective tissue that supports the nerves and into blood vessels that nourish the nerves.

Mechanisms by which genetic influences regulate the formation of the complex patterns within the nervous system remain poorly understood. The axons of motor neurons (*e.g.*, those that activate muscles) grow out of the spinal cord through the primordial ventral roots of the spinal nerves that leave the cord, terminate in the somite or other mesodermal elements, and make connections (synapses) with primordial muscle cells (myoblasts). When the myoblast migrates to its final location, it pulls, or tows, its elongating axon along. The axons of other neurons and of neurilemma cells follow the path taken by the initial outgrowths of axons. In this way peripheral nerves, essentially bundles of axons (neuron fibrils), differentiate and develop. Innervation of all myoblasts tends to be ensured by a process in which myoblasts without neural supply seem to attract growing nerves while those already supplied tend to reject excessive synapses (hyperinnervation). In computer language, nerve development may be programmed by peripheral structures such as myoblasts.

CHARACTERISTIC STRUCTURES

The neuron is the basic structural and functional unit of the circuits and networks of the human nervous system. Each of the more than 15,000,000,000 neurons within the system has synaptic contacts with other neurons, and each acts as integrator, conductor, and transmitter of coded neural information (*e.g.*, as electrical impulses). Some neurons innervate muscle and glandular cells. Non-neural surrounding (interstitial) cells are in intimate contact with the entire neuron including its fibril processes (axons and dendrites). In a sense interstitial elements (called oligodendroglia, astroglia, neurilemma cells, satellite cells, and ependymal cells) are intercalated between the neurons and the blood vessels, extracellular fluids, and connective tissue cells. Except for a few small sites in the entire system, each neuron is effectively enveloped by these supporting, nourishing cells. The connective tissue in the nervous system is similar in structure and function to that in the rest of the body including microglia, or histiocytes, which serve as scavenger cells capable of ingesting bits of degenerating or waste matter in the nervous system. The extracellular fluids in the vicinity of the neurons and their processes contain electrolytes such as salt; thus the composition of these fluids has significant electrical and metabolic effects on the activity of the neuron as an information processor and signaller. The central nervous system is bathed in cerebrospinal fluid, both inside its cavities and outside in the subarachnoid space formed by membranes (meninges) that envelop the brain and spinal cord (see also CEREBROSPINAL FLUID).

Structure of neurons

Neurons. As seen with a microscope, a typical neuron includes a cell body (soma), thin threadlike processes called dendrites, and an axon. Any neuron has metabolic, or trophic, receptive, conductile, and transmissive functions associated with different parts of the cell.

The trophic structure is the cell body, containing the nucleus—the metabolic centre essential to the life of the neuron—as well as other cytoplasmic substructures within the cell. One or more dendritic processes receive and carry neural information into the cell body, and the axon conducts impulses away from the cell body to the terminal dendrites. The latter may chemically link and communicate (synapse) with the cell body of another neuron or with a gland cell or muscle cell at the motor end plate. Depending on their location in the body, receptive dendrites may receive stimuli directly from the environment (*e.g.*, pressure on the skin) or from other neurons through synaptic contacts; in most neurons the dendrites and remainder of the cell body function as the receptive segment of the cell. The conductile segment (typically the axon) is specialized for the transport of neural impulses from the receptive segment to the transmissive portion of the cell. The presynaptic transmissive segment (the short, terminal axonal portion) contains in vesicles chemical substances, called neurosecretions, which pass into a tiny gap (synaptic cleft) to exert a chemical influence on the next cell in the chain.

Nerve cells vary in shape and range in length from a fraction of an inch (*e.g.*, small neurons in the brain) to several feet in tall persons (*e.g.*, sensory neurons extending from the big toe into the central nervous system). In an elephant or whale, neurons can be appropriately longer. Not all nerve cells have the same microscopic appearance. Sensory receptor neurons, for example, may incorporate distinctively different structures. Thus there are non-neural onion-like capsules (pacinian corpuscles) capping some of the receptors in the skin, whereas, in the eye, there are retinal rods and cones, highly specialized nerve cells that are selectively responsive to different kinds of light (see also SENSORY RECEPTION, HUMAN; EYE AND VISION, HUMAN).

The entire neuron including all its processes is bounded by a thin plasma membrane that is selectively permeable to different chemical substances, barring some and permitting others to pass through. This process is crucial to the maintenance of the resting electrical potential of the neuron and to the neural impulses the cell generates.

Other elements of the cell body

In addition to the nucleus, the cell body contains such substructures as Nissl substance, Golgi apparatus, mitochondria, and neuroplasm (see also CELL AND CELL DIVISION). The Nissl substance appears as particles ranging in size from discrete clumps to dustlike bits that function (along with the nucleolus of the nucleus) in producing proteins required for the renewal of protoplasm. Neurons are such prodigious synthesizers of protein that they can replace one-third of their protein content each day. Neuronal protein may flow from the cell down the axon (axoplasmic flow) at the rate of three to 20 millimetres (0.1 to 0.8 inch) each day. The Golgi apparatus, visible only with special staining methods, is a fine network of cytoplasmic organelles associated with the secretory process. The mitochondria are the powerhouses of the cell, utilizing the energy-rich chemical adenosine triphosphate (ATP) for the neuron's metabolic activities (see also METABOLISM). The neuroplasm is the amorphous, flowing cellular protoplasm (cytoplasm); it normally contains a high concentration of potassium and other substances critical to the production of the nerve impulse and other metabolic activities of the neuron. The branching dendrites of the cell body contain their own complement of Nissl substance, mitochondria, and neuroplasm and range from a fraction of a millimetre to generally no more than a few millimetres in length.

The axon also shows mitochondria and neuroplasm but is without Nissl substance. It originates from the cell

body as the conical axon hillock and tapers to a small diameter in what is called the initial segment. The initial segment is considered to be the focal structure that acts as an all-or-none "gate" or "valve" for the neural impulse; at any rate, the neuron either generates all the electrical impulse it can, or it sends out none. The axon ends in terminal dendrites with knobs (boutons terminaux) at their tips. Each knob synapses to transmit excitatory or inhibitory influences to another cell.

Synapses

Synapses, or synaptic junctions, between neurons are specialized complexes where communication occurs between an axonal ending of one cell and a dendrite (axodendritic synapse), cell body (axosomatic synapse), or axon (axo-axonic synapse) of another. An axon and its branches may terminate in only a few or as many as thousands of synapses (a thousand may be an average). In turn, the cell body of one neuron may receive synaptic contacts from several hundred to more than 20,000 other neurons. The synapse consists of a presynaptic portion, an extremely narrow gap called the synaptic cleft, and a postsynaptic portion. The presynaptic knoblike portion shows vesicles that contain the precursor chemicals of the transmitter substance (*e.g.*, acetylcholine, norepinephrine) released at that synapse. The plasma membranes of the two adjacent cells are respectively called the presynaptic and the postsynaptic membrane at the synaptic cleft (about one millionth of an inch). Following the arrival of the all-or-none impulse, transmitter substances are released across the presynaptic membrane into the synaptic cleft to combine chemically with receptor sites in the postsynaptic membrane, evoking physiological response in the next cell in the chain.

Peripheral nerves. Peripheral nerves (*i.e.*, those serving parts in or near the body surface) are bundles of axons extending side by side, crudely analogous to coaxial cables. Each axon is comparable to one wire among the many in a cable, and the neurilemma cells and thin connective-tissue layer encapsulating each axon can be likened to the insulation around the wire. Nerve axons are bound together in subgroups (fascicles and groups of fascicles); all of these are enveloped by a covering of connective tissues to form the entire nerve. Many individual axons are covered by a myelin sheath, a complex of fatty materials and proteins that is the product of the layers of the plasma membrane and of the neurilemma (Schwann) cells. Other axons, which are surrounded by neurilemma cells but without myelin, are called unmyelinated fibres. The myelin sheath is a segmented, sausage-like layer interrupted at regular intervals by indentations called nodes of Ranvier. One neurilemma cell forms the myelin sheath for each internode (between two nodes of Ranvier). There is a direct relation between the internodal length and the speed of neural conduction: the longer the internodes, the greater the conduction speed of the impulse. Myelin improves the signalling efficiency of a nerve fibre; in effect, the impulse hops along the plump sheath from node to node in a process called saltatory conduction. The axons of unmyelinated fibres are encapsulated by neurilemma cells that do not produce myelin; these fine, small-diameter fibres have relatively slow speeds of conduction.

Other neurilemma cells—called satellite cells—encapsulate the cell bodies of sensory ganglia of the cranial and peripheral nerves and of the ganglia of the autonomic nervous system to be discussed later.

Cells of the central nervous system. In the central nervous system, neuroglial (glia) cells outnumber neurons five to ten times. Except for blood vessels and small phagocytic (scavenger) glial cells (microglia), connective tissue is sparse in the central nervous system. Extracellular fluid is present in the spaces (interstices) among the cells and nerve fibres. Oligodendroglia in the gray matter surround the cell bodies of neurons, whereas, in the white matter of the central nervous system, these glial cells produce the myelin sheaths of the axons. These glial cells are equivalent to the satellite, or neurilemmal, cells of the peripheral nervous system. Nodes of Ranvier and internodes are present in myelinated neurons of the central nervous system. Astroglia are cells that are intercalated between blood capillaries and neurons. Processes from these glial cells are in intimate contact with about 85 percent of the capillary surface; other processes are in close synaptic contact with cell bodies of neurons. Interrelations among neurons, blood vessels, glial cells, and extracellular fluid serve to maintain concentrations of chemical substances essential to the neuron's resting potential and all-or-none impulse. They also provide the basis for a kind of blood–brain barrier, a mechanism through which many substances (which readily pass from the blood into other tissues of the body) enter numerous parts of the central nervous system much more slowly and in lower concentration. Some antibiotics, for example, move so poorly from the bloodstream into the cerebrospinal fluid that they are not always a good choice for treating certain infections of the nervous system.

Physiological interrelations among neurons and surrounding tissues

NERVE DEGENERATION AND REGENERATION

An injured neuron exhibits metabolic changes that tend to repair the cell and restore its function. When a peripheral nerve is cut (transected), Nissl substance in each damaged neuron breaks up (a process called chromatolysis), the cell body swells, and the nucleus is displaced to the side of the cell body. In the nerve segment distal to the cut (*i.e.*, the part now disconnected from the central nervous system), the axon and the myelin sheath of each neuron fragment and are subsequently lost in favour of a continuous band of new neurilemma cells. Such so-called degenerative responses appear actually to be initial steps toward new growth, or regeneration. The cell body increases the rate at which it produces protein, which, in turn, flows down the axon to support the sprouting of several new branches at the reconstituting cut end. These growing tips elongate, invade the distal stump, and follow the bands of neurilemma cells, optimally at a rate of about four millimetres per day. Eventually the regenerating tips reach the sensory nerve receptors and motor endings.

Functionally effective regeneration, however, does not occur invariably. Each regenerating axon may send out many new branches that may become associated with different neurilemma bands. Each band, in turn, may act as a scaffold that guides a number of regenerating fibres. The eventually surviving axons are those that grow to terminate in the same kind of nerve endings that were theirs before the injury (*e.g.*, touch fibre to a touch ending in the skin). Those fibres that fail to form such appropriate functional connections eventually degenerate, while successful fibres remyelinate. Although peripheral nerves often can regenerate, transected axons in the central nervous system show regenerative changes but rarely form effective new synapses. In general, then, such injured neurons in the brain or spinal cord can be given up as functionally lost.

Sometimes an uninjured axon may sprout a branch from a node of Ranvier (collateral regeneration). This sprout invades and grows down a neurilemma band of a nearby injured nerve fibre. Thus a healthy neuron may incorporate the severed end of a damaged nerve cell and return that ending to full function. Successful collateral nerve regeneration occurs in both the peripheral and the central nervous systems. (C.R.N.)

II. The central nervous system

The human central nervous system consists of a brain, its connected spinal cord, and their associated membranes, fluids, and blood vessels. The central nervous systems of all vertebrates are similar in structure and function, but there is generally an increase in relative size and complexity as the phylogenetic scale is ascended. The comparative anatomy and physiology of nervous systems of all animals, as well as the nomenclature regarding anatomical structures of the nervous system and their position, are dealt with in the article NERVES AND NERVOUS SYSTEMS. Most aspects of the central nervous system of man are almost identical to those of other primates, dealt with rather extensively in the previously mentioned article.

The normal human brain, a unified, balanced, symmet-

rical, and harmonious organ of specialized and unique function, is endowed with considerably more potential than is realizable in the course of one person's lifetime. The continuity of personality throughout life hinges most directly upon the brain. An individual may lose almost any other part of his body—a kidney, his stomach, half his liver—or exchange his heart for another without losing his identity; his brain, however, seems to be where he "lives" as a legally and psychologically unique individual. If complete nerve regeneration were possible, and if parts of all severed nerves could be matched, there could still, in actuality, be no such procedure as a "brain transplant" —it would be a matter of a body transplant. A "well" brain would acquire another body in which to be housed.

The delicate brain and its spinal cord are provided with specially evolved, intricate protective and nutritive mechanisms of bone, muscle, ligament, and blood vessels, as well as supporting cushions of fluids.

Structure of the skull and spinal column

The human skull that encases the brain is remarkably structured in weight, contour, cavities, depressions, and openings for the exit and entry of nerves carrying an almost infinite variety of motor and sensory messages to and from various organs via other parts of the nervous system. The bony spinal column, extending caudally from the skull, normally comprises 30–33 individual bones (vertebrae) linked together by fine, multiple criss-crossing ligaments and intervertebral cartilage that permit great mobility (except at the lower end, where the column is relatively fixed as it joins the pelvis). The vertebrae are so structured in curves, configurations, and facets as to allow considerable flexibility of movement while still bearing the stress of erect posture and body weight. At the same time the spinal column protects the delicate nervous tissue of the spinal cord running most of its length. The column also has small channels (foramina) that allow for the entry and exit of peripheral nerves (see also SKELETAL SYSTEM, HUMAN).

The lowest part of the brain, the medulla oblongata, is continuous through the foramen magnum, the opening at the base of the skull, with the spinal cord. The brain and spinal cord, as previously mentioned, constitute the central nervous system; the nerves passing to and from the central nervous system form the peripheral nervous system. The human nervous system includes certain nerve centres and nerve fibres that, without conscious knowledge of the processes concerned, control the vital functions of the body, such as circulation of the blood and respiration (Figure 1). This system, called the autonomic nervous system, includes sympathetic and parasympathetic components. These systems will be described later. Through communicating branches the brain can influence organs that are supplied by the autonomic system —*e.g.*, the salivary glands and heart, both of which may be acted on by fear, hunger, and other emotions. Ordinarily, however, the internal organs function without awareness of the processes involved.

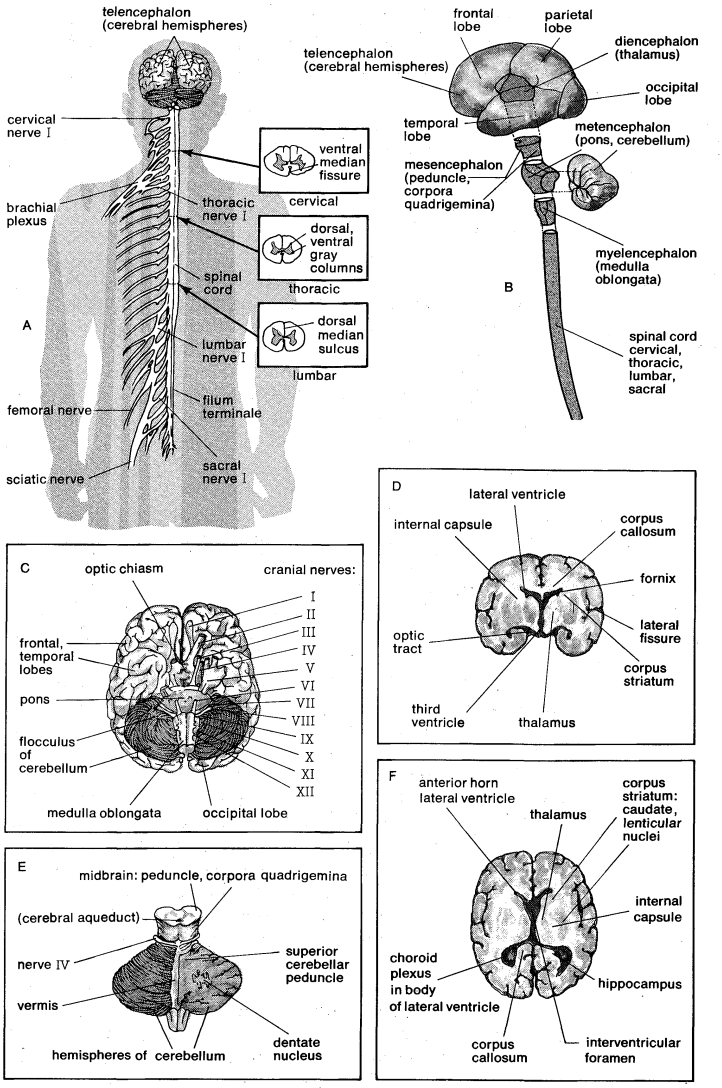


Figure 1: *Brain and spinal cord.*
(A) Posterior view; (B) major divisions. Various aspects of the brain; (C) base; (D) coronal section; (E) cerebellum and midbrain; (F) horizontal sections (right side at lower level).
From W.A. Dorland, *Dorland's Illustrated Medical Dictionary*, 23rd ed. (1957); W.B. Saunders Co.

THE BRAIN

General structure. The average male brain weighs about three pounds (1,400 grams). It is surrounded by three successive membranes called meninges and protected by the bony skull (neurocranium). The brain (encephalon) is divided into five parts: telencephalon (end brain), diencephalon (interbrain), mesencephalon (midbrain), metencephalon (afterbrain), and myelencephalon (medulla oblongata). The cerebrum comprises the telencephalon, diencephalon, and upper midbrain. The metencephalon is divided into the pons and the cerebellum. The pons is connected to the cerebellum by bridges of nerve fibres called cerebellar peduncles.

Each half of the bilateral telencephalon, called a cerebral hemisphere, consists of an outer gray mantle called the cerebral cortex, a deep mass of gray matter—the basal ganglia—and the white matter located between the two gray structures. The hemispheres are connected by bands of nerve fibres called commissures. The corpus callosum is the largest commissure. The cerebral cortex is characterized by ridges called convolutions or gyri, which are separated by grooves called sulci. The right and left cerebral hemispheres are separated by a midsagitally located, deep, longitudinal fissure.

Two major structures of the diencephalon are the thalamus and the hypothalamus. The brain has a series of cavities, called ventricles, that are filled with cerebrospinal fluid. All of these structures are described in more detail in this article.

Membranes of the brain. The membranes, or meninges, covering the brain are an outer, tough, fibrous layer, the dura mater; a thin, intermediate, weblike tissue, the arachnoid; and a soft, vascular inner covering, the pia mater.

The dura mater. The dura mater lines the cranial cavity. In it are meningeal arteries and veins that serve for the nutrition of the bone. If the dura mater is torn in an injury to the skull, an effusion of blood may occur between the dura mater and the bone, which may, by exerting pressure on the underlying brain, cause paralysis of the opposite side of the body. The inner surface of the membrane is smooth and moist. The dura mater also forms partitions (septa) between the hemispheres of the brain and cerebellum. These septa are folds of the dura mater and consist of two layers, blended where they touch but separated along the attached borders of the septa to form venous channels, or dural sinuses, by which blood and excess of cerebrospinal fluid is drained from the brain into the great veins of the neck, which carry the

Functions of the dura mater

fluid mixture back toward the heart (see also CEREBROSPINAL FLUID).

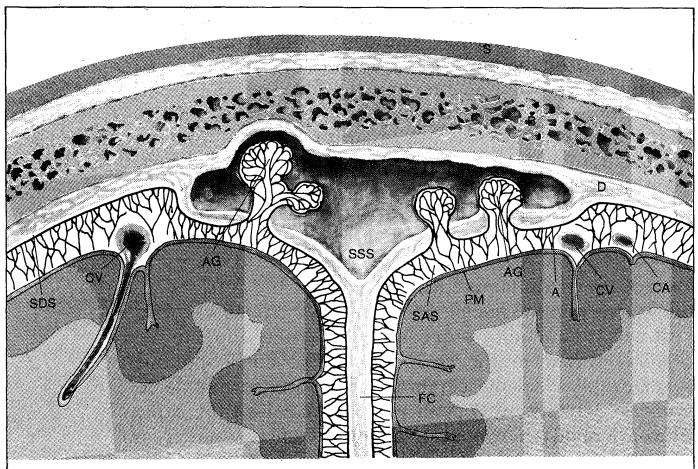
The secretion from the pituitary gland is also carried away into the general bloodstream by small venules, which open into the neighbouring cavernous and other venous sinuses. Absorption and passage of the cerebrospinal fluid is carried out, to a large extent, by small villous projections of the arachnoid membrane, which extend into the venous sinuses and spaces of the dura mater and are most numerous in the neighbourhood of the sagittal sinus. If this passage is prevented or retarded, the intracranial pressure of the cerebrospinal fluid rises, and one form of hydrocephalus results. After birth the arachnoid villi enlarge to form what are called pacchionian bodies, beginning at about 18 months. Between the dura mater and the subjacent arachnoid membrane is an interval called the subdural space. It contains a small quantity of fluid that serves to lubricate the smooth inner surface of the dura mater.

The arachnoid membrane. Beneath the dura mater is the arachnoid membrane, which, although thin, is not permeable to fluids. It is separated from the pia mater, beneath, by the subarachnoid space. This is traversed by a network of delicate fibrous bands. The meshes of this network are filled by the subarachnoid cerebrospinal fluid; the larger thin-walled cerebral arteries and veins covering the surface of the brain lie in the thin bands of fibrous tissue forming the net. In the lower part of the roof of the fourth ventricle (see below) are three openings in the arachnoid membrane and pia mater: a median, the foramen of Magendie; and two lateral, the foramina of Luschka. These form a communication between the cerebrospinal fluid in the ventricles of the brain and that contained in the subarachnoid space. Obliteration of these openings by meningitis produces an obstructive hydrocephalus, in which the accumulation of fluid is entirely intraventricular.

Openings in the arachnoid

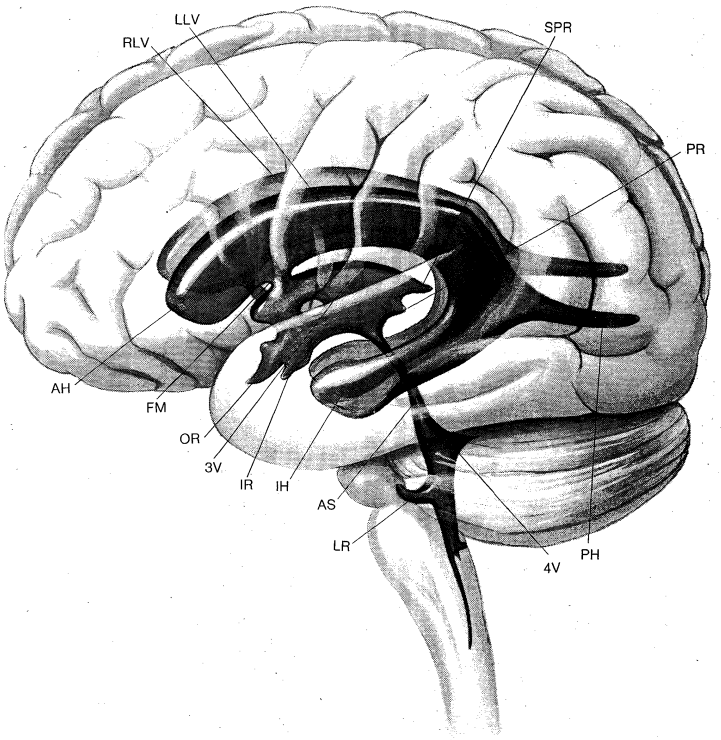
The pia mater. The pia mater is a delicate vascular membrane forming the immediate investment of the brain and dipping down into the fissures between the convolutions. It contains the smaller arterioles and venules that supply the subjacent brain. A large triangular fold of pia mater (velum interpositum) is included in the great transverse fissure lying between the corpus callosum and fornix above and the roof of the third ventricle and thalamus below. This pyramidal fold contains two great blood vessels, the cerebral veins of Galen, which drain the blood from the interior of the brain. Vascular fringes at the margin of the fold project into the lateral ventricles, and similar fringes project from under the surface of the fold into the third ventricle. These fringes are the choroid plexuses of the lateral and third ventricles, and a similar choroid plexus is found in the roof of the fourth ventricle. The fringes are covered by a secretory layer, called the choroidal epithelium, which elaborates and secretes some of the cerebrospinal fluid (see also CEREBROSPINAL FLUID).

After Frank H. Netter, M.D., *The CIBA Collection of Medical Illustrations;* Copyright © 1957, 1972 by CIBA Pharmaceutical Company, Division of CIBA-GEIGY Corporation; reproduced with permission. All rights reserved



A—arachnoid
AG—arachnoid granulation
B—bone
CA—cerebral artery
CV—cerebral vein
D—dura mater
FC—falx cerebri
PM—pia mater
S—skin
SAS—subarachnoid space
SDS—subdural space
SSS—superior sagittal sinus

Figure 2: Arachnoid and pia mater.



AH—anterior horn of lateral ventricle
AS—cerebral aqueduct of Sylvius
FM—interventricular foramen of Monro
IH—inferior horn of lateral ventricle
IR—infundibular recess
LR—lateral recess of fourth ventricle
LLV—left lateral ventricle
OR—optic recess
PH—posterior horn of lateral ventricle
PR—pineal recess
SPR—suprapineal recess
RLV—right lateral ventricle
3V—third ventricle
4V—fourth ventricle

Figure 3: Lateral view of cerebrum, cerebellum, and ventricles.

After Frank H. Netter, M.D., *The CIBA Collection of Medical Illustrations;* Copyright © 1957, 1972 by CIBA Pharmaceutical Company, Division of CIBA-GEIGY Corporation; reproduced with permission. All rights reserved

Ventricles. Ventricles are cavities containing fluid, situated in the substance of the brain, and lined by a thin membrane, the ependyma. There are four true ventricles: the right and left lateral ventricles, contained in the cerebral hemispheres; the third ventricle, situated in the diencephalon; and the fourth ventricle, located in the pons and medulla (hindbrain). Each lateral ventricle is connected with the third ventricle by a small opening, the interventricular foramen of Monro; and the third ventricle is joined to the fourth by a narrow channel; the cerebral aqueduct of Sylvius in the midbrain. The fourth ventricle communicates below with the central canal of the spinal cord and with the subarachnoid space by the foramina of Magendie and Luschka.

The cerebrospinal fluid contained in the ventricles and subarachnoid space acts as a mechanical support to the brain and spinal cord; it also performs functions similar to those of tissue fluid and lymph in other parts of the body (see TISSUES AND FLUIDS, ANIMAL). The brain floats in the cerebrospinal fluid. Blood is supplied to the brain by two paired arteries: the internal carotid and the vertebral arteries. These are in communication with each other at the base of the brain, where they form the so-called arterial circle of Willis. Venous return is arranged not alongside the arteries, as almost everywhere else in the body, but through a system of dural sinuses that drain blood to the jugular foramen and from there to the internal jugular veins.

Medulla oblongata. The medulla oblongata is situated in the lower and posterior part of the cranial cavity. It appears to be a direct continuation upward of the spinal cord but differs in its arrangement of fibres composing the nerve tracts and in disposition of the gray matter. It contains the vital cardiac, vasomotor, and respiratory

Vital centres within the medulla oblongata

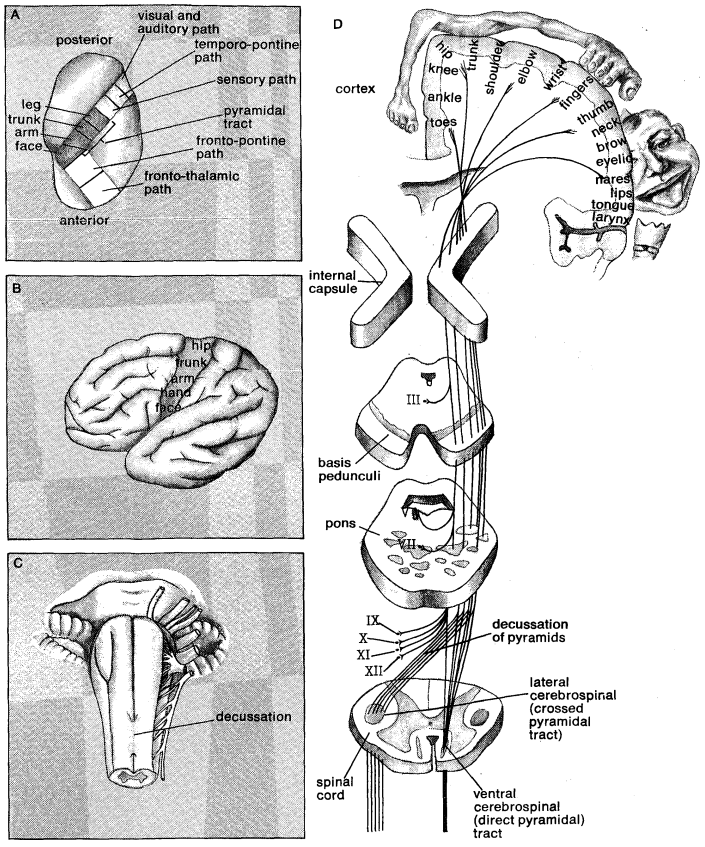


Figure 4: *Pyramidal system.*
(A) Horizontal section through internal capsule to show location of principal pathways; (B) lateral aspect of cerebral cortex to show topographic projection of motor centres (schematic); (C) ventral aspect of brainstem showing decussation of pyramids; (D) tracts of the pyramidal system.
After Frank H. Netter, M.D., *The CIBA Collection of Medical Illustrations;* Copyright © 1957, 1972 by CIBA Pharmaceutical Company, Division of CIBA-GEIGY Corporation; reproduced with permission. All rights reserved

centres. These are situated in the lower part of the medulla. Longitudinal bundles of nerve fibres connect the medulla oblongata with the pons, and two diverging bundles of fibres called the paired inferior cerebellar peduncles join it to the cerebellum.

The principal longitudinal tracts that connect the pons with the medulla are (1) the pyramidal tracts, (2) the medial longitudinal fasciculi, and (3) the medial lemnisci.

Pyramidal tracts. The pyramidal tracts consist of motor fibres, each of which originates from the motor areas of the cerebral cortex and descends through the internal capsule of the cerebrum, midbrain, and pons to the anterior part of the medulla. Here they form two parallel strands (pyramids), one on each side of a median vertical groove. In the lower part of the medulla oblongata, the greater number of the fibres of the pyramidal tract cross over to the opposite side of the spinal cord, where they form a bundle of descending fibres called the crossed pyramidal tract. The few remaining fibres are continued downward on the same side of the cord as the direct pyramidal tract. Eventually these fibres also cross to the opposite side. The crossing of the motor nerve fibres in the medulla oblongata is called the decussation of the pyramids, and, because, with few exceptions, all motor and sensory fibres cross to the opposite side, each cerebral hemisphere influences the muscles of and receives sensory impulses from the opposite side of the body.

Medial longitudinal fasciculi. The medial longitudinal fasciculi are paired tracts of nerve fibres that have their cells of origin in the vestibular nuclei of the brainstem. The fibres end in the motor nuclei of the muscles of the eyes and the descending ones in the nuclei of the neck muscles of the same and the opposite side. By means of these tracts, the position of the eyes and the head can be adjusted to each other and coordinated in such a way that the eyes remain fixed on a given object, regardless of the position of the head.

Medial lemnisci. Each medial lemniscus is a longitudinal tract of ascending sensory fibres lying close to the median plane between the pyramidal tract and the medial longitudinal fasciculus. The fibres of the right and left tracts cross the median plane, forming the sensory decussation. This lies in the medulla above the level of the motor, or pyramidal, decussation.

The sensory impulses coming from the spinal cord end in cell processing stations in the medulla, the gracile and cuneate nuclei. It is from these that fibres (internal arcuate fibres) cross to opposite sides and ascend in the medial lemniscus, which terminates in the thalamus. This lemniscus forms one link, or relay, in the main sensory tract to the cortex.

The olive

Immediately lateral to the pyramidal tract on the anterior aspect of the medulla is an oval swelling, the olive. The olive lies over a folded lamina of gray matter in the substance of the medulla; this is the nucleus of the inferior olive. It is connected with the opposite cerebellar hemisphere by fibres that cross the middle line and reach the cerebellum through the inferior cerebellar peduncle. The nuclei from which the motor fibres of the cranial nerves III through XII originate and those in which the sensory fibres terminate lie in the reticular formation of the medulla and pons. Like the spinal nerves, they are connected by nerve tracts with the opposite hemisphere (for a description of cranial and spinal nerves, see below *The peripheral nervous system*).

Pons. The pons is located between the midbrain and the medulla oblongata. Its posterior surface forms the floor of the upper half of the fourth ventricle. The posterior region of the pons, adjacent to the ventricular surface, is called the tegmentum, while the anterior region is the basilar portion. The tegmentum is part of a continuous structure of the same name extending rostrally (to the front) and caudally (to the rear) to form much of the midbrain and medulla. The basilar portion extends rostrally, as the crura cerebri of the midbrain, and caudally, as the pyramids of the medulla.

In common with the tegmentum of the midbrain and the medulla, the tegmentum of the pons contains nuclei of origin (motor) and of termination (sensory) of some cranial nerves, the tracts of ascending sensory pathways, sone descending motor pathways, and the reticular formation. The nuclei of the cranial nerves include those of the fifth, sixth, seventh, and eighth cranial nerves. The ascending pathways include the spinothalamic tract (pain and temperature from the body), trigeminothalamic tracts (pain, temperature, and touch from the head), medial lemniscus (touch from the body), the medial longitudinal fasciculus (vestibular influences to the nuclei innervating the eye muscles and correlating position of head in space to the direction of gaze; see below), and the lateral lemniscus (hearing). The descending pathways include the corticobulbar fibres, which influence the motor nuclei of the cranial nerves, and the fibres of the extrapyramidal system (see below). The reticular formation is an intricate neuronal network organized into complex ascending pathways, descending pathways, and local neuronal circuits.

The basilar portion of the pons includes: (1) the fibres of the pyramidal tracts, which course longitudinally from the crura of the midbrain, located rostrally, to the pyramids of the medulla; and (2) the fibres of the corticopontine tracts, pontine nuclei, and fibres of the pontocerebellar tracts. The latter group conveys information from the cerebral cortex to the cerebellum. The corticopontine tracts originate in the cerebral cortex and terminate to synapse (interconnect) with pontine nuclei on the same side of the body. The pontine nuclei have fibres that cross over in the basilar portion and course through the middle cerebellar peduncle before terminating in the cerebellum of the opposite side.

Cerebellum. The cerebellum consists of a medial part, known as the vermis, and two lateral hemispheres. Each hemisphere is connected with the brainstem by three

paired stemlike parts (peduncles): (1) the inferior cerebellar peduncle to the medulla; (2) the middle cerebellar peduncle; (3) the superior cerebellar peduncle. The last joins the cerebellum to the midbrain and conveys mostly efferent fibres, one bundle leaving the cerebellum and ascending to the contralateral red nucleus and to the contralateral thalamus, and another bundle going finally to the reticular formation of the medulla oblongata.

The cerebellum receives impulses from the equilibratory organ in the middle ear, from muscle spindles, from the exteroceptors of the skin, from the eye, and from the ear (acoustic organ), as well as from the cerebral cortex. These impulses reach the gray matter of the cortex of the cerebellum and are believed to pass from there to the deep cerebellar nuclei (dentate, emboliform, globose, and fastigial nuclei). From this nuclei a relay of fibres ascends to the opposite red nucleus, thalamus, and brainstem reticular formation (see below). The fastigial nuclei are integrated functionally with the vestibular system. The superficial surface of the cerebellum differs from that of the cerebral hemispheres. In place of convolutions, the vermis and hemispheres of the cerebellum are crossed by numerous transverse fissures, which mark off a series of folds, or folia. The general arrangement of these is seen in a median section through the vermis, which presents a branched appearance called the arbor vitae. The surface of the cerebellar hemispheres and central lobe is subdivided into lobes and lobules by deep fissures.

The cortex of the cerebellum consists of a superficial stratum, the molecular layer; an intermediate layer, which contains the cell bodies of branched Purkinje cells; and an inner deep stratum, the granular layer, which rests upon the central white matter. The latter is formed by medullated nerve fibres, which course to and from the gray matter. The Purkinje cells are large and have extensive connections. The cell bodies are pear-shaped and arranged in a single layer. From the outer end of each cell, processes (projections) branch out in the molecular layer.

Purkinje cells

These processes are the dendrites, and the branching takes place chiefly in a plane at right angles to the longitudinal axis of the folium in which it lies. The dendrites of the Purkinje cells are intersected at right angles by parallel fibres that run in the direction of the folium. These fibres are derived from axons of the granule cells in the inner stratum of the cortex, which pass outward and divide in the molecular layer in a T-shaped manner into right and left branches. The bodies of the Purkinje cells, moreover, are surrounded by a network of fibres that originate from basket cells in the molecular layer; and the Purkinje dendrites are also accompanied by delicate climbing afferent nerve fibres from the white matter.

Although areas of the cerebellar cortex cannot be mapped out by response of particular groups of muscles to electrical stimulation, it is possible on morphological grounds, by means of experimental work and by the tracing of tracts of nerve fibres entering the cerebellum, to locate areas of the cortex according to the fibres that they receive from the particular parts. The cerebellum, relatively large in man, weighs more than the entire spinal cord. It is the great modulator subserving the coordination of groups of muscles. The cerebellum smooths out the actions of these groups by delicately regulating and grading muscle tensions. Cerebellar reactions are unconscious; destruction of the cerebellum entails no loss of sensation or intelligence, although cerebellar disturbance may cause some proprioceptive misperception and poor muscular coordination.

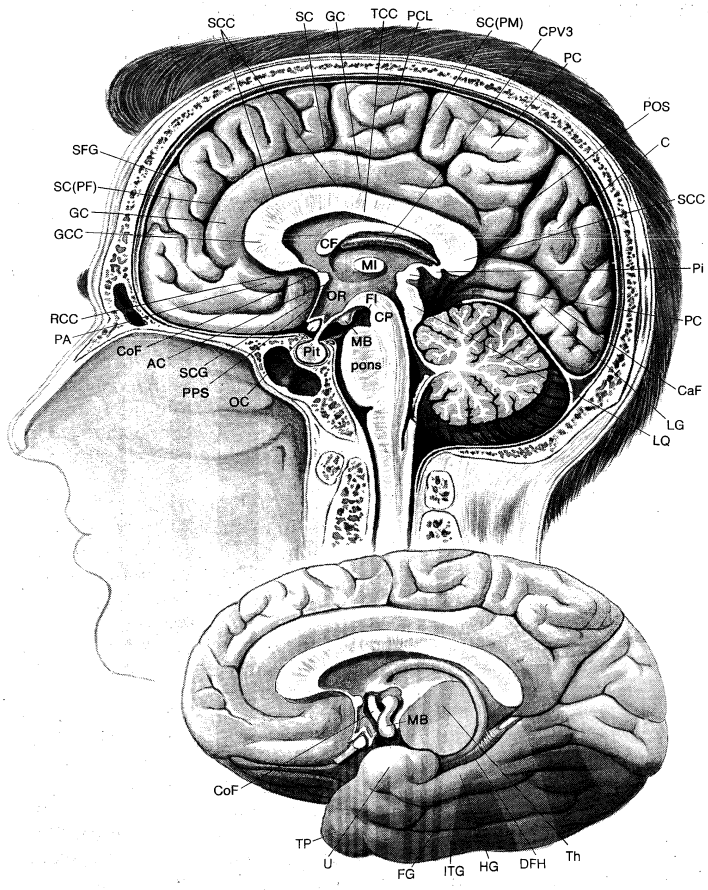
The midbrain. The midbrain (mesencephalon) connects the pons and cerebellum with the cerebrum. It is traversed by the cerebral aqueduct of Sylvius. The part that lies above the aqueduct, called the roof plate or tectum, is subdivided into four rounded swellings. These are the colliculi or corpora quadrigemina. The upper pair of these bodies receive nerve fibres from the retina, which reach them through the optic tracts and from the medial lemniscus. They are an important centre in lower mammals but are of less importance in primates. The lower pair serve as a cell station in the path of the auditory impulses, which commence in the cochlea and terminate in the cerebral cortex of the temporal lobe. The gray matter in the roof of the aqueduct receives an important tract of nerve fibres from the spinal cord, known as the spinotectal tract.

A similar bundle, the spinothalamic tract, also traverses the midbrain. This pathway conveys the more primitive sensations of pain, heat, and cold to the receptive centres in the thalamus. These have been described as protopathic sensations, to distinguish them from the finer and more recently evolved sensations of touch, which have been termed epicritic. The latter ascend in the posterior columns of the spinal cord to the gracile and cuneate nuclei. From these a relay of nerve fibres ascends through the medulla oblongata and, crossing in this to the opposite side of the brain, passes as the medial lemniscus through the pons and midbrain to the thalamus. From the thalamus another relay of fibres carries the sensory impulses to the cerebral cortex of the brain.

Spinotectal and spinothalamic tracts

The most ventral part of the midbrain forms the crura cerebri. These two diverging limbs are composed of nerve fibres coursing from the cerebral cortex to the spinal cord (pyramidal tract) and pons (corticopontine fibres). Behind, or dorsal to, the crura is a lamina of pigmented nerve cells (substantia nigra) that separates the pes pedunculi from the tegmentum. This is traversed by decussating fibres, the greater number being the cerebellar

After Frank H. Netter, M.D., *The CIBA Collection of Medical Illustrations;* Copyright © 1957, 1972 by CIBA Pharmaceutical Company, Division of CIBA-GEIGY Corporation; reproduced with permission. All rights reserved



APS—anterior parolfactory sulcus
C—cuneus
CaF—calcarine fissure
CF—body of fornix
CP—cerebral peduncle
CPV3—choroid plexus of 3rd ventricle
CoF—column of fornix
DFH—dentate fascia of hippocampus
FG—fusiform gyrus
FI—interpeduncular fossa
GC—gyrus cinguli
GCC—genu of corpus callosum
HG—hippocampal gyrus
ITG—inferior temporal gyrus
LG—lingual gyrus
LQ—lamina quadrigemina
MI—massa intermedia
MB—mammillary body
OC—optic chiasm
OR—optic recess
PA—parolfactory area
PC—precuneus
PC-L—paracentral lobe
Pi—pineal body
Pit—pituitary gland
PPS—posterior parolfactory sulcus
RCC—rostrum of corpus callosum
SC—sulcus cinguli
SC(PF)—sulcus cinguli (pars frontalis)
SC(PM)—sulcus cinguli (pars marginalis)
SCC—splenium of corpus callosum
SCG—subcallosal gyrus
SFG—superior frontal gyrus
TCC—trunk of corpus callosum
Th—thalamus
TP—temporal pole
U—uncus
PC—posterior commissure
POS—parieto-occipital fissure

Figure 5: Medial aspect of the brain.

fibres already mentioned as issuing from the dentate nucleus, travelling by the superior cerebellar peduncle to the midbrain and then crossing to the red nucleus of the opposite side and to the thalamus.

From the red nucleus a tract of nerve fibres descends to the spinal cord, where it forms connections with the motor cells in the gray matter of the anterior horn (see below *Spinal cord*). This is the rubrospinal tract. The red nucleus is also connected with the thalamus, the brainstem reticular formation, and the nucleus of the inferior olive. By means of these connections, it is believed that the cerebral hemispheres control the more reflex movements carried out by the spinal cord, such as balancing movements and the maintenance of posture.

In the tegmental part of the midbrain are also situated the longitudinal association tracts, known as the dorsal and medial longitudinal fasciculi, and the nuclei and roots of origin of the third and fourth cranial nerves and part of the fifth in the gray matter surrounding the cerebral aqueduct of Sylvius.

Cerebral hemispheres. Two of the most distinctive features of the human brain are the large size of the cerebral hemispheres and the high degree of specialization in the microscopic structure of the cerebral cortex. The cerebral hemispheres comprise the cerebral cortex, the lateral ventricles, white matter deep to the cerebral cortex, and the basal ganglia. The surface of each hemisphere is subdivided, for descriptive purposes, into lobes and lobules. Certain sulci and lines that are arbitrarily drawn between these are employed for demarcating the boundaries of these areas. The central sulcus, or sulcus of Rolando, is situated on the lateral surface and separates the frontal from the parietal lobe. The lateral sulcus, or sulcus of Sylvius, marks off the temporal lobe from the parietal and frontal lobes. On the median surface are the callosomarginal and parieto-occipital sulci, which limit the frontal, limbic, parietal, and occipital lobes.

The central, or Rolandic, sulcus marks the posterior limit of the important motor area of the cortex. Electrical stimulation of particular parts of this area produces definite movements of groups of muscles, the action of which is normally initiated and controlled by the part stimulated. Injury to the same part causes paralysis of the corresponding muscles.

The cortex of the occipital lobe, which surrounds the posterior part of the calcarine sulcus, is the visuosensory area for reception of visual impulses from the retina. The visuosensory area is surrounded by a zone that extends on to the outer aspect of the occipital lobe, termed the visuopsychic area.

The middle part of the first temporal gyrus and the adjacent gyri on the lower lip of the fissure of Sylvius are concerned with hearing. The auditory area of the brain receives sensory impulses by way of the auditory radiation from the medial geniculate body of the same side. These, the lower auditory centres, are connected with the opposite ear by means of the lateral lemniscus.

The front part of the hippocampal gyrus, with its hooklike end, the uncus, is the higher cortical centre for the sense of smell. This is the most primitive of the special senses. It is closely associated with the sense of taste and is more highly developed in many other vertebrate animals than in man. That part of the brain that is concerned with the sense of smell and with certain complex emotional responses is called the limbic lobe. There is also a large area behind the central sulcus that extends forward onto the motor area. This is the cutaneous sensory, or tactile, area.

The areas of cerebral cortex lying between the special and cutaneous sensory areas are believed to function as association centres between the different senses and as centres in which the memory of associated sensations is stored. The region in front of the motor cortex (premotor cortex) is connected by association fibres with all the various sensory and motor areas and, more especially, with that part of the adjacent motor area that is concerned with the movements of the eyes. It is therefore believed that this part of the cortex may control skilled movements that are dependent on impulses reaching the brain from the eyes and that require close attention and a knowledge or memory of past experiences—such movements, for instance, as those of the lips and tongue or of the hand, which have given man the powers of speech and of writing. In front of the premotor cortex is the prefrontal cortex, which is somehow involved with such qualities as disposition, outlook on life, drive, personality, planning for the future, and control of activities along ethical standards. If the upper and lower lips of the fissure of Sylvius are separated, a triangular area of submerged cortex called the insula or island of Reil will be exposed.

The basal ganglia

The basal ganglia are the large masses of gray matter embedded in the cerebral hemispheres. They include the caudate and lenticular nuclei; the subthalamic nucleus of the thalamus and the substantia nigra of the midbrain are often included. The lenticular nucleus is divided into the putamen and the globus pallidus. The caudate nucleus and putamen are called the striatum, or neostriatum, while the globus pallidus is also called the paleostriatum or pallidum. The caudate and lenticular nuclei form the corpus striatum. The basal ganglia are integrated through complex neural connections with the cerebral cortex and thalamus. They are involved with the basic regulation of stereotype movements basic to the performance of complex volitional movements. The malfunctioning of the basal ganglia is contributory to a variety of motor disturbances. These include paralysis agitans (Parkinson's disease), athetosis, and choreas. A general term for some of these abnormal movements (called dyskinesias) is cerebral palsy. The precise role of each of these structures in these conditions was not fully understood in the early 1970s. For example, degenerative changes in the globus pallidus and substantia nigra are present in patients with Parkinson's disease; in addition the neostriatum and substantia nigra have reduced amounts of dopamine within them. Paralysis agitans is characterized by rigidity and tremor. In athetosis, the movements are slow and those of the extremities have a wormlike action. Choreas (from the Greek meaning dance) are featured by jerky and brisk movements of the extremities and involuntary twitchings of the facial and neck muscles.

The cerebral cortex is the stratum of gray matter that covers the central white matter of the cerebral hemispheres. It is comprised of nerve cells, nerve fibres, blood vessels, and a supporting tissue, the neuroglia. It exhibits a definite stratification into layers of nerve cells and nerve fibres. In sections of fresh brain, the main strata and various zones are easily recognizable by the unaided eye, and more especially so in the visual cortex, or area striata. These appearances correspond closely with those revealed by microscopic preparations. In recent years it has become apparent that there is also a columnar arrangement of the cells that can easily be seen in properly cut sections.

Although the cerebral cortex is conventionally described as being organized into layers oriented parallel to the surface of the cortex, the basic functional unit is now conceived as being a physiologically defined vertical column extending from the cortical surface to the white matter. These columns are: (1) the terminal regions for fibres (input) that originate in the thalamus and other cortical columns of the same and opposite cerebral hemispheres, and (2) the source of output fibres that terminate in other cortical areas of the same and opposite cerebral hemispheres and in neural centres in the basal ganglia, thalamus, brainstem, and spinal cord. The neurons of the cerebral cortex include: (1) pyramidal cells (cell body in the shape of a pyramid), with axons that extend from the column to the white matter as the output fibres just noted; (2) small cells with short axons that interact with pyramidal cells and other small cells in the immediate vicinity. The axons of the pyramidal cells leave the column of origin, join the white matter, and terminate in other columns by re-entering the cortex, or the gray matter, of other regions of the brain and spinal cord. The axons of the small neurons are integrated into intracortical neuronal circuits that tend to sustain the vertical columnar organization. The complex synapetic connectivity among

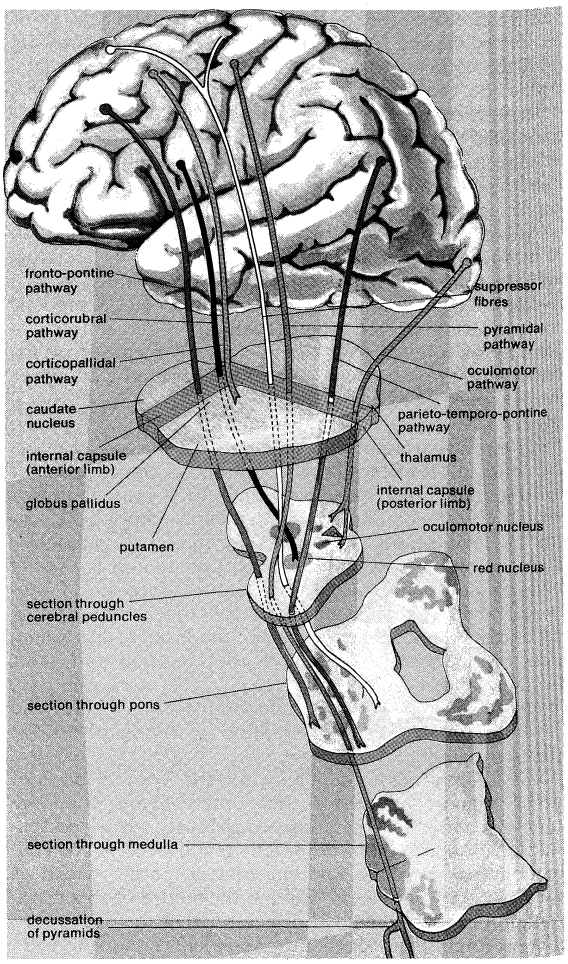


Figure 6: Cerebral cortex efferent–afferent and associate pathways.

After Frank H. Netter, M.D., *The CIBA Collection of Medical Illustrations;* Copyright © 1957, 1972 by CIBA Pharmaceutical Company, Division of CIBA-GEIGY Corporation; reproduced with permission. All rights reserved

these small neurons and the pyramidal neurons is basic to the intricate activity within the cortex. These small neurons are numerous in the cortex of man; hence these neurons are considered to be the "structural expression" of the delicacy of the function of the human brain.

In animals the degree of lamination and differentiation of the nerve cells in the cortex appears to correspond to the stage of evolution attained by the particular species. Thus, the cortex is thicker and more highly evolved in man and the higher types of mammals than it is in lower forms (see NERVES AND NERVOUS SYSTEMS). In the development of the human cortex, stratification begins about the sixth month of fetal life, when the convolutions first appear, and differentiation continues not only during the later months of fetal life but for a considerable period after birth.

The thalamus is strategically located to receive influences from a variety of sources and to send information to the cerebral cortex. In brief, the massive complex of neural processing centres receives input from all sensory systems, with the exception of the olfactory system, and from the basal ganglia, brainstem, and cerebellum, and sends neural influences to the cerebral cortex. The thalamus has a role in the conscious appreciation of sensation; many of the crude sensory modalities (pain, touch, and, possibly, sound detection) are brought to consciousness in this structure. The affect of an individual (*e.g.*, state of contentment and well-being) and the phase of the sleep–wake cycle are largely mediated through the thalamus. The rhythmic brain wave activity as observed in the electroencephalogram is synchronized and integrated in the thalamus and expressed in the cerebral cortex.

The hypothalamus is a small structure, weighing four grams, involved with numerous functional activities associated with the autonomic nervous system. It has crucial roles in influencing the functional activities of the pituitary gland (hypophysis) and in the regulation of water balance through the so-called "thirst centre," of body temperature by maintaining the balance between heat production and heat loss, and of food intake and feeding responses. In addition, the hypothalamus is involved with many objective physical expressions associated with emotion and behaviour; for example, alteration in the heart beat (palpitations), blushing, cold sweat, and tears of sadness. The sleep–wake cycle is influenced by activity of the hypothalamus. A portion of the thalamus, called the subthalamus, is functionally integrated with the basal ganglia.

The brainstem. The brainstem comprises the midbrain, the pons, and the medulla oblongata. The diencephalon—the thalamus and hypothalamus—is often included within the brainstem. The diencephalon and midbrain are referred to as the upper brainstem and the pons and medulla as the lower brainstem. The brainstem is dealt with separately, as a unit, because of its special roles in reflexes, sensory and motor conduction, the regulation of the internal environment of the body, and regulation, or control, of the rest of the nervous system. In man the embryonic central nervous system undergoes differentiation at its interior end into the brain, while the remainder becomes the spinal cord. The original part of this differentiating brain becomes the brainstem. The stem, from its lateral and dorsal aspects, gives off "buds" that develop into the cerebral hemispheres and cerebellum. These overgrow and cover their stalk, so that, when the adult brain is removed from the cranium, nothing of its initial component is seen until, viewed from the underside, a lumpy column about as large as a man's thumb from tip to wrist, and not greatly different in shape, can be distinguished as the brainstem.

The brainstem is interposed between and interconnected with the more differentiated parts of the brain on the one hand and the spinal cord on the other, and it bears certain similarities to each of these structures. By means of its entering and exiting nerves, the spinal cord is concerned largely with gathering sensory impressions from and effecting motor behaviour to what were originally serially arranged segments of skin or muscles of the body. The skin of the head and the muscles of the eyeball and tongue are likewise supplied with nerves from the brainstem. Sensory endings and muscles develop around the nasal, oral, and pharyngeal cavities, and their derivatives are also supplied by brainstem nerves. The specialized sense organs—the ear, eye, and olfactory receptors—which develop in the head, are also innervated by nerves connecting with the brainstem. These nerves of the head are collectively designated as cranial (see below *The peripheral nervous system*).

Ascending pathways from the spinal cord continue upward through the brainstem and are joined there by others, conveying afferent messages to be relayed to the cerebellum or to the cerebral hemispheres. Similarly, from these latter structures descend into the brainstem, in part to terminate in relation to its motor outflows, in part to continue uninterruptedly through it toward motor outflows from the cord, and in part to reach these cranial or spinal outflows indirectly after relay by collections of brainstem neurons. The brainstem therefore resembles the spinal cord, of which it is the rostral continuation, both in supplying nerves to the periphery and in accommodating ascending pathways to, and existing connections from, the cerebral hemispheres and cerebellum.

The development called encephalization, which carried the cerebellum and, to a great degree, the cerebral hemispheres to such heights of structural complexity and functional achievement, led the brainstem also to assume a dominating and integrating role over lower lying outflows from the nervous system—in general, those innervating the viscera, or skeletal structures related to viscera—which preserve the internal environment of the body in a state of optimum activity; in addition, and less clearly perceived, this development also led the brainstem to assume the role of maintaining a state of optimum activ-

ity within the interior of the central nervous system itself, both in the spinal cord below it and in the cerebral hemispheres above.

The functional aspects of the brainstem are as follows:

Brainstem reflexes. A reflex is a relatively invariable and stereotyped response to a specific stimulus. Those reflexes mediated by the brainstem and its cranial nerves are concerned with adjusting the related musculature for optimum performance of the body's distance receptors, the eye and the ear, with orienting parts of the body in spatial relation to the head, and with managing the complicated acts involved in ingesting food.

Reflexes related to the eye and ear

Of the reflexes related to the eye, coordinated deviation of the eyeballs, often combined with turning of the head to perceive some aspect in the field of vision, occurs almost constantly during an individual's waking hours. Alterations for near vision are also frequent and involve contraction both of muscles converging the eyeballs and of intraocular muscles, which alter the curvature of the lens; the latter manoeuvre, called accommodation, serves to focus light rays on the retina. Associated with it is a constriction of the pupil cutting off light from the periphery of the lens, whose shape is not changed as much. The light reflex is a constriction of the pupillary aperture also evoked, as the term implies, by visual stimulation. Its purpose is to reduce excessive illumination of the retina. A so-called Argyll-Robertson pupil (extremely contracted and fixed to light but constricting in accommodation to various distances) is a frequently encountered sign in neural syphilis, for the brainstem path of the light reflex is unusually susceptible to injury by the spirochete of this disease (see also NERVOUS SYSTEM DISEASES).

More intense light evokes a reflex blinking of the lids, but the corneal reflex, a blink caused by stimulation of the cornea by touch or by drying, is more important not only for protection but also because loss of lachrymal secretion by frequent wiping of this structure leads to its injury, the cornea being the last portion of the body surface of terrestrial animals still to require a watery environment (see also EYE AND VISION, HUMAN).

Two reflexes related to the ear are evoked by auditory stimulation. One directs the head toward the source of a sound and appears designed to capture the maximal intensity of sound stimuli. The other, the auditory reflex, involves the two smallest muscles of the body, buried in the middle ear, the tensor tympani and stapedius, whose contractions reduce undue movement of the ossicular chain in response to loud sounds. In its damping action to loud sound, the auditory reflex is analogous to the pupillo-constrictor reflex of the eye to bright light.

Also located in the ear is the vestibular organ, which initiates reflexes involving (1) the muscles that orient the eyeballs in effective positions for vision when the head is moved and (2) muscles of the trunk and extremities that hold the body in a definite spatial relationship to the force of gravity. The most easily examined of these vestibular reflexes is nystagmus, a repetitive deviation of the eyeballs designed to keep them abreast of the rotating head. It may be evoked by spinning in a chair or by irrigating the external ear with hot or cold water, thus creating convection currents in the fluid-filled canals of the vestibular organ, simulating those induced by rotation. Prolonged repetitive stimulation of the balancing part of the vestibular organ, as by the pitching of a ship at sea, evokes a battery of visceral alterations not normally induced by vestibular nerve impulses, accompanied by sensations of nausea and distress—an incapacitating syndrome described as motion sickness (see also EAR AND HEARING, HUMAN).

Reflexes related to food ingestion

Reflexes associated with the nasal and oral cavities are concerned with conveying solid and liquid food into the gastrointestinal tract and with expelling these or other foreign objects from the closely adjacent air passages that they sometimes enter. The first of these objectives is accomplished by the swallowing reflex, in which stimulation of the back of the mouth cavity evokes sequential contractions of the muscles of the lips, jaw, tongue, pharynx, and part of the neck, propelling food or liquid into the esophagus. Ingested material is prevented from regurgitating into the nose, during swallowing, by a tensing of the soft palate and from entering the larynx, by a drawing up of this latter structure beneath the back of the tongue, so that the swallowed material cascades past it (see also DIGESTION, HUMAN). Impairment or absence of swallowing, called dysphagia, from injury to the swallowing reflex arc, is a frequent symptom of poliomyelitic involvement of the brainstem.

Reflexes expelling material from the nasopharynx are of two sorts; those in which the local musculature tries to accomplish this act unaided, as in gagging or choking; and those in which the respiratory musculature is also invoked to contribute a blast of expired air to the process, as in sneezing or coughing. Swelling of the mucosa in allergies or infections of the nasopharynx causes persistent stimulation of its receptor endings bringing about recurrent sneezing or coughing.

Related to these reflexes are vomiting and hiccupping, the latter appearing to be a repeated but ineffective effort to vomit. Both reflexes involve cramplike contractions of the abdominal and thoracic respiratory musculature—particularly the diaphragm—designed to empty the stomach by regurgitation. Because hiccupping is personally annoying and can become painful, the attempt is usually made to stop it by various manoeuvres. In cases of prolonged and intractable hiccupping, surgical crushing of the phrenic nerve, innervating the diaphragm, has sometimes been necessary to relieve an exhausted patient of the annoyance.

Sensory conduction through the brainstem. There are three general aspects of sensation with regard to the brainstem.

First, afferent nerve impulses gathered from all parts of the body and conducted to the cerebral hemispheres are somehow converted there into consciously perceived sensations, in which the intermittent features of nerve conduction fuse into the familiar picture of the world as normally seen and heard. Though the nature of this transmutation is not understood, it is clear that the delivery of afferent impulses to the hemispheres is prerequisite to it. If a receiving area of the hemisphere is destroyed, for example, or if the afferent pathway to it is interrupted at some lower level, the modality of sensation it subserves becomes lost to consciousness. Because paths conducting afferent messages toward the hemispheres are topographically distinct from one another, sensations can be lost singly. It is possible in cases of intractable pain that can no longer be relieved with drugs for the surgeon, by an appropriately placed incision either in the spinal cord or in the brainstem, to interrupt that pathway mediating impulses from pain receptors and thus eliminate pain sensations from consciousness, while preserving other afferent pathways and leaving unaffected the sensations that they mediate.

Next, if the central course of any single sensory pathway is examined, it will be found that its component nerve fibres, mediating impulses from different portions of the receptor field, whether for vision, hearing, pain, or touch, possess a definite spatial relationship to one another. This relationship is preserved in the relay of these impulses from the anterior end of the brainstem to the various sensory areas of the cerebral hemispheres. This orderly arrangement of conducting elements, an obtrusive structural feature of afferent pathways at each level of the central nervous system, may be without functional significance. It seems likely, however, that it underlies the orientation that objects are seen to bear to one another in the field of vision or the orderly relation that sounds of different pitch hold to one another in audition or the sense of localization that normally is attached to painful or tactile impressions.

The "feeling quality" of sensations

Lastly, another general feature of afferent conduction through the brainstem is the "feeling quality" that sensations often possess. Many sensations are perceived with no reaction but indifference, but most individuals can recall some dramatic instance when perception was accompanied by a warm suffusion of pleasure or by a shuddering chill of unpleasantness. Because lesions in or near the anterior end of the brainstem are sometimes followed

by states collectively called the thalamic syndrome, in which the most prosaic sensations are accompanied by intense and unmerited feelings of pleasure or unpleasantness, it has been suggested that this region of the brain is particularly concerned with endowing sensation with these qualities.

Motor conduction through the brainstem. The activity of skeletal musculature, whose contraction holds body parts in position or moves them about, is integrated at both upper and lower levels of the nervous system (the stereotyped contraction of some of the head muscles in cranial reflexes has already been considered). The finer, more labile and adaptive contractions of skeletal muscle, particularly in voluntary movement, are managed by a motor area in the cerebral hemispheres from which the pyramidal tract descends on each side through the brainstem to pass, with crossing, directly to the levels of motor outflow whose discharge it initiates.

A part of this pyramidal projection, concerned with voluntary contraction of the head musculature, terminates in relation to the brainstem cells of origin of the cranial motor nerves and is designated as the corticobulbar system. (The pons and medulla are called the bulb.) Though corticobulbar connections subserve a number of functions, the most important of which is speech, relatively little is known of their neural organization. After corticobular fibres are given off, the remainder of the pyramidal projection descends into the spinal cord for voluntary contraction of the muscles of the body, in particular those manipulating the distal portions of the extremities. Injury to the motor area of the hemispheres or to its descending pyramidal tract, as frequently occurs at the junction of the brainstem and hemispheres in a stroke or apoplectic attack, is followed by a paralysis of voluntary movement.

Not all of the cerebral influences upon movement are exerted over the pyramidal tract, however. A significant part reach lower outflows over a series of shorter paths, some of the constituent relays of which are provided by collections of brainstem neurons. Because this group of motor connections lies outside the pyramidal projection, it is called the extrapyramidal system. This latter system is able to initiate contraction of a postural nature, involving chiefly muscles of the trunk or proximal parts of the extremities. It serves also in managing grosser types of movements, such as those of progression. One of the major functions of the extrapyramidal system is, however, that of modifying activity initiated from other neural sources and already in progress. Because different components of this system respectively augment or reduce motion, their injury is followed either by depression or exaggeration of motor activity. Other symptoms that may be present following an injury to a component of extrapyramidal system are those of involuntary movement, such as tremor.

The extrapyramidal system

After Frank H. Netter, M.D., *The CIBA Collection of Medical Illustrations*; Copyright © 1957, 1972 by CIBA Pharmaceutical Company, Division of CIBA-GEIGY Corporation; reproduced with permission. All rights reserved

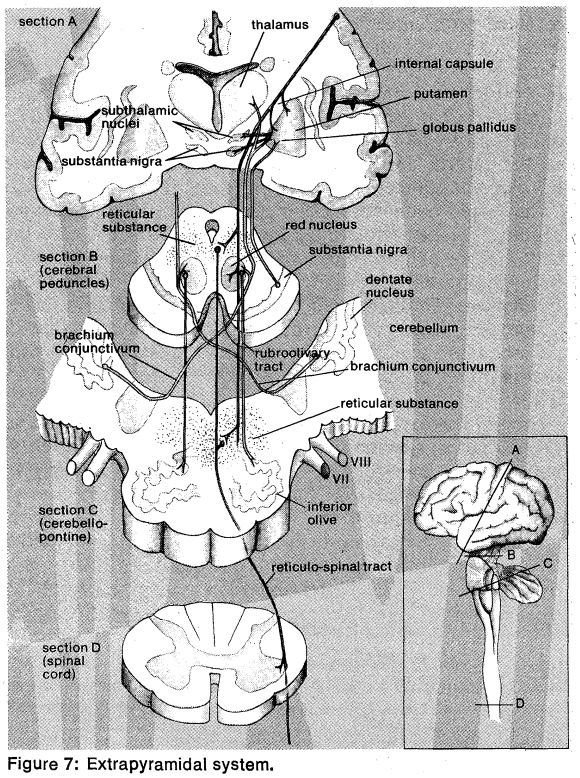


Figure 7: Extrapyramidal system.

Important modulating influences on motion are also exerted by the cerebellum and are mediated by connections passing either upward, via the brainstem, to the cerebral hemispheres or downward, over brainstem relays, which are apparently identical with those of the extrapyramidal system, to lower motor outflows. The brainstem thus plays an important part in motor conduction both from the cerebral hemispheres and from the cerebellum.

Regulation of the internal environment. In serving local reflexes and providing relays for pathways to and from other levels, the brainstem is analogous to the spinal cord. In exerting a dominating and regulating influence over collections of lower outflows concerned with certain general functions of the body, it resembles the cerebral hemispheres. The hemispheres are broadly involved with preserving optimum relations of the body with the external world, and, to a degree, the brainstem is comparably concerned with managing a number of the processes that maintain the internal environment of the body in an optimum condition. Parts of the brainstem so concerned are its anterior and posterior extremities.

The hypothalamus of the upper brainstem, overlies the pituitary, or master gland of the endocrine system. It has connections with this gland that control the liberation of pituitary hormones and so exert widespread influences throughout the body. A second group of efferent connections from the hypothalamus descends, with relays to the spinal cord to terminate in relation both to outflows for skeletal muscle and, in larger part, to outflows of the autonomic nervous system, innervating smooth muscle and exocrine glands throughout the body. By exciting parts of the autonomic system in varying combinations, or throwing the whole system into activity at once, this second outflow from the hypothalamus may likewise alter body states profoundly.

Innervation of the posterior pituitary

Variations in the osmotic tension of blood circulating through the hypothalamus cause its neurons to innervate the neurohypophysis (posterior pituitary) and to elaborate and to secrete that gland's hormones. One of these, antidiuretic in action, passes in the bloodstream to the kidney, where it facilitates the reabsorption of water from the kidney tubules. About 170 litres of water filter from man's blood every 24 hours and start down the kidney tubules toward the bladder. With the aid of a normal complement of an antidiuretic hormone from the neurohypophysis, about 169 litres of this water are reabsorbed into the bloodstream before leaving the kidney, and only the remaining litre, conveying waste materials in solution, continues to the bladder and thence voided as urine.

If the hypothalamus is injured, or if its path to the pituitary is interrupted, the neurohypophysis becomes paralyzed and atrophic, so that the secretion of its antidiuretic hormone fails. Reabsorption of body water from the kidney is then incomplete, and in record cases as many as 40 litres may be lost daily and passed as urine, so that, were not a compensating quantity of water drunk, the body would rapidly become dehydrated. This disturbing syndrome is called diabetes insipidus. Administration of an extracted antidiuretic hormone constitutes a specific therapy (see also ENDOCRINE SYSTEM, HUMAN; ENDOCRINE SYSTEM DISEASES AND DISORDERS; HORMONE).

Although the secretion of the anterior lobe of the pituitary appears to be somewhat more autonomous, appropriate injury to the anterior brainstem causes deficiency in production of gonadotropic hormones, with consequent atrophy of the sex glands, regression of secondary sexual characteristics, and loss of potency. Brainstem in-

juries causing deficiencies in the sexual sphere are often associated with the development of pronounced obesity, the combined disorder referred to clinically as adiposogenital dystrophy, or Fröhlich's syndrome. A contrasting disturbance, the precocious onset of puberty, which may result from involvement of the anterior brainstem, depends on either overactivity of the mechanism just considered or malfunction of the pineal gland, connected with the dorsal, rather than the ventral, aspect of the brainstem.

The anterior brainstem is also involved in the increased secretion of the adrenocorticotropic hormone (ACTH) from the anterior lobe of the pituitary in response to stress stimulation. Regulation of the adrenal cortex thus appears to be under the control of the brain. Stress stimuli have been found to cause increased discharge of the anterior brainstem, and direct stimulation of this structure has been shown to increase ACTH secretion. Conversely, after injury to the anterior brainstem, stress is no longer capable of leading to increased pituitary–adrenal activity.

Relationship of the anterior brainstem with the autonomic nervous system

By way of its connections with autonomic outflows, the anterior brainstem is similarly able to regulate or widely alter conditions in the interior of the body. Excitation of this neural part evokes a mass discharge of the sympathetic portion of the autonomic system, with rise in blood pressure, increase in heart rate, cessation of gastrointestinal peristalsis, erection of hair, sweating, and other changes brought about by emotional excitement known as the fight or flight reaction.

Many of the changes just mentioned are also those that are employed in preserving a relatively stable internal body temperature in the face of wide variations in environmental conditions. Heat, produced in the body by oxidative reactions, is conserved by peripheral vasoconstriction (and erection of the hair in fur-bearing animals), reducing the body's irradiation; when necessary, heat conservation is also augmented by shivering. It is dispelled by increased irradiation, consequent to peripheral vasodilation, and by the evaporation of an increased sweat secretion. These opposing processes (heat retention and heat release) are regulated by two antagonistic thermal mechanisms in the anterior part of the brainstem, influenced by alterations in the temperature of the blood circulating through the brainstem.

These neural systems by which the brainstem preserves a constant body temperature may sometimes undergo derangement. Circulating toxins or breakdown products, incident to infection, paralyze that neural mechanism dispelling body heat and periodically excite the other, initiating heat production and conservation. The affected individual then shivers and exhibits peripheral vasoconstriction, and, although he subjectively experiences a chill, his internal temperature rises to fever height. Elevated body temperature together with increase in the count of white blood cells, enemies of foreign agents in the body, thus constitute the two most reliable diagnostic signs of infections.

The lower brainstem is concerned with managing internal processes, even more vital in nature, that clear the blood of carbon dioxide, provide it with a fresh supply of oxygen, and circulate it through the body (see also RESPIRATION, HUMAN).

Neural control of respiration

The contraction of the thoracic and abdominal musculature, which expands the chest cavity, permitting air to flow into the lungs, is initiated by lower brainstem neurons, the efferent fibres of which descend to the spinal outflows of the inspiratory muscles. These brainstem neurons are constantly being excited by a multiplicity of factors associated with oxygen deficiency in the circulation. The rhythmicity of respiration is affected by the periodic prevention of inspiration, both by inhibitory afferent impulses initiated by lung stretch during inspiration and conveyed to the lower brainstem by cranial afferent nerves and by central inspiratory–inhibitory or expiratory connections. In respiration more forceful than that of ordinary quiet breathing, the active contraction of expiratory muscles is added to the relaxation of inspiratory muscles in expelling air.

Though the motor nerve supply to the respiratory muscles proceeds entirely from the cord, spinal neural levels are, by themselves, incapable of maintaining respiration. Participation of the brainstem mechanisms is essential for this activity.

The lower brainstem also contains regulating collections of neurons whose efferent fibres influence outflows (from both the brainstem and spinal cord) concerned with the rate of heartbeat and those (from the cord alone) that innervate smooth muscle in the walls of arterial blood vessels. Contraction of the latter maintains an optimum pressure in the circulatory system, enabling blood to be conveyed to all parts of the body.

Two receptor organs, informing the lower brainstem of both the pressure and the chemical composition of circulating blood (via cranial afferent nerves), play an important part in the regulation of cardiovascular activity. The first of these is strategically located in the arch of the aorta, where it can test the blood leaving the heart and starting on its way to all parts of the body. Because the body apparently is particularly concerned with maintaining an optimum circulation through the cerebral hemispheres of the brain, another, similar receptor is found in the wall of the internal carotid arteries, which supply blood to these structures. Injury to the brainstem cardiovascular mechanism or interruption of its connections to autonomic outflows is followed by vasomotor collapse and fall of blood pressure. Resultant circulatory disturbances, if unrelieved, are fatal.

Regulation of the central nervous system. The brainstem plays an important role in maintaining an optimum state of activity within the interior of the central nervous system. Injury to the rostral brainstem is followed by deterioration of function of the cerebral hemispheres; injury to more caudal parts of the stem results in impairment of activity within the spinal cord.

Results of injury to the anterior brainstem

The conscious perception of sensation, the voluntary control of motion, and other features that characterize the activity of the cerebral hemispheres in what is called the waking state are somehow dependent upon the functional integrity of the upper brainstem. Injury to this region is followed by states varying in intensity from somnolence to coma. In the most severe states, though the patient's respiration and circulation are preserved, and he may be nourished and kept free of infection for years, he has no more rapport with the external world than a vegetable. The electroencephalogram (EEG) has revealed pronounced differences in the spontaneous electrical activity of the cerebral cortex in wakefulness and sleep. In sleep the electrical record is composed largely of high, slow waves and is described as being synchronized. During wakefulness, in contrast, the EEG is made up of low-voltage, fast activity. Direct stimulation of the central core of the brainstem will reproduce the normal waking record.

Many side connections or collateral branches from ascending sensory paths have been shown to turn into the central core of the brainstem and to exert their arousing influences indirectly by modifying its activity. Structurally, the portion of the brainstem involved is reticular (like a network), and research has identified the existence of an ascending reticular activating system concerned with regulation of the background activity of the cerebral cortex.

Though the lower brainstem's influence upon the spinal cord is not as specific and is abetted by both the cerebral hemispheres and the cerebellum, it is important in preserving the normal function of this part of the nervous system. Transection of the cord from the brain, even at thoracic levels, leaving respiratory and vasomotor activity unimpaired, is followed, in the isolated cord segments, by a specifically neural depression called spinal shock. In this condition, spinal reflexes for a period can no longer be evoked by afferent stimulation and, even when recovered, may exhibit abnormality. Though other factors may be concerned, spinal shock depends in part upon loss of excitatory influences descending to the cord over pyramidal and extrapyramidal connections, to the latter of which the brainstem makes contributions.

In addition to exerting ascending influences upon the cerebral cortex, the central reticular formation of the brainstem has been shown to give rise to descending connections to motor outflows from the spinal cord. The more cephalic of these augment or facilitate spinal motor discharge, while the more caudal region exerts an inhibitory or reducing action. An imbalance in these extrapyramidal motor connections is thought to be responsible for spasticity, in which condition inhibitory influences are no longer active, while facilitatory ones are preserved and exert an augmented effect since they are no longer opposed. The involuntary movements of other types of extrapyramidal disease may also have their basis in malfunction of these descending connections from the brainstem.

Oxygen requirements of the brain

Nutrition of the brain. The oxygen requirement of the brain, even at rest, is high. The human brain comprises only about 2 percent of the total body weight, yet it uses approximately 25 percent of all the oxygen taken up by the body under conditions of complete mental and physical rest. The purpose underlying the oxygen consumption by the brain undoubtedly is the same as in any other living system, viz., to provide energy to be used in doing work. Yet, as Otto Warburg, a pioneer in studies of brain metabolism, pointed out in 1912, nerve cells do no work that is measurable in mechanical or chemical units, since they do not contract, move, divide, or grow. Nevertheless, the brain obviously needs oxygen, as indicated by (1) its extraordinarily high oxygen consumption; (2) its vulnerability to deprivation of oxygen; and (3) the fact that functional activity of the brain runs parallel to its oxygen consumption, exhibiting cerebral depression from such conditions as anesthesia, diabetic coma, insulin shock, and brain tumors.

The foodstuffs and oxygen necessary to meet the requirements of the brain can reach it only by way of the blood. An increase in the functional activity of any organ entails a simultaneous increase in the amount of blood that it receives. It is noteworthy, however, that the way in which this compensation is brought about differs markedly among organs. In the case of the skeletal muscles, for example, increased activity is associated with both an increase in the amount of blood expelled by the heart and a widening of the blood vessels supplying the exercising muscles, brought about largely via the brainstem. In the salivary glands, increased secretion of saliva is elicited by nerve impulses carried by special nerves, and these impulses also cause dilation of the blood vessels supplying the gland. In the brain, however, although nerves are known to be able to mediate either constriction or dilation of blood vessels, these effects are much weaker than those in other tissues, and nerve control appears to play no important part in the normal regulation of the cerebral circulation. It is likely that the blood vessels of the brain possess the unusual property of tending to constrict if left to themselves and that this tendency is constantly opposed by the dilating action of the products of cell metabolism either arising in the brain itself or carried to it in the blood from other tissues. Any increase or decrease in the oxygen consumption of the brain would then be associated with a corresponding change in the calibre of its blood vessels and, consequently, in the volume of blood flowing through them, because the amount of vessel-dilating material had changed. Increase or decrease in the concentration of material such as this in the blood reaching the brain would have similar effects on cerebral blood flow.

Mechanism of brain-blood-flow action

The evidence in favour of this conception is of two main types: (1) all the characteristic products of nerve metabolism tested (increased carbon dioxide, decreased oxygen, acids, potassium salts, acetylcholine, phosphates, heat) have been found to dilate cerebral blood vessels; and (2) increase in the carbon dioxide or decrease in the oxygen content of the blood reaching the brain (by inhalation of appropriate gas mixtures) causes dilation of cerebral blood vessels and increase in cerebral blood flow, while decrease in carbon dioxide (by overbreathing) or increase in oxygen has the opposite effect. It is probable that the ordinary adjustments of the cerebral circulation to the changing requirements of the brain do not necessitate corresponding changes in the general circulation. When the latter is depressed by such influences as hemorrhage, shock, or drugs, cerebral blood flow will follow this change, apparently quite passively. When the general circulation is stimulated (as by blood transfusions or drugs), cerebral blood flow will be increased. No constricting or dilating drugs have been shown to be able to produce a change in cerebral blood flow that is contrary to changes in the general blood pressure, and in this respect the cerebral circulation is unique. Even the coronary vessels in the heart, which resemble the cerebrals in being obedient to changes in general blood pressure and in lacking important control by vasoconstrictor or vasodilator nerves, can be strongly constricted by an extract of the posterior lobe of the pituitary gland, but administration of the extract has no direct effect on the cerebral circulation. Spasm of cerebral vessels has long been suspected to be the cause of certain conditions, such as convulsions, temporary depression or loss of consciousness, memory, or speech, and transitory disturbances of vision, but the cause of the spasm, or indeed the fact of its occurrence, has not been clearly demonstrated. The only factor shown to be capable of causing an important degree of constriction of cerebral vessels is hyperventilation, which can produce mental confusion, or even complete unconsciousness, in normal persons and can elicit a typical convulsion in persons with latent epilepsy. Diminution in cerebral blood flow produced by constriction of cerebral vessels may be partly the cause of these conditions.

Factors affecting cerebral blood pressure

The commonest and most important derangements of the cerebral circulation are produced by either a decrease in the arterial blood pressure or an impediment to flow in the cerebral vessels themselves. The former can be caused by an appropriate decrease in any of the three main factors concerned in the regulation of blood pressure, namely, efficiency of the heart, resistance of the blood vessels, and volume or viscosity of the circulating blood. The latter may be caused by obstruction of cerebral arteries by a clot carried by the blood (embolism) from another part of the body (usually the heart), by formation of a clot in a cerebral artery or vein (thrombosis), or by actual rupture, with hemorrhage into the brain substance. In all such cases the part of the brain supplied by the involved vessel or vessels loses its functions, and, from the consequent symptoms, it is usually possible to locate the lesion with considerable precision. The loss of function is a result of the fact that the distribution of the arteries in the brain does not overlap significantly. Therefore, there is no alternative source of supply of blood when one terminal artery is occluded, and the tissue supplied loses its function because it undergoes degeneration.

Another source of interference with the cerebral circulation lies in the peculiar location of the brain inside a rigid container, the skull. Because the container is practically indistensible, an increase in the volume of any of the cranial contents will lead to an increase in pressure involving all of them and will be reflected in an increase in pressure of the cerebrospinal fluid. This pressure operates on the blood vessels, as well as on other cranial contents, and, when it rises to a level higher than the pressure inside the vessels, the latter collapse and flow of blood through them ceases. A change of this type in the vessels supplying the vasomotor centre in the brainstem causes increased activity there and thus brings about a rise in blood pressure. When the rise is sufficient to overcome the abnormal pressure on the outside of the vessels, flow of blood through them is restored. Thus the arterial blood pressure characteristically rises in cases of high intracranial pressure whether the latter depends on an intracranial hemorrhage or a tumour, on interference with drainage of the cerebrospinal fluid, on swelling of the brain itself (as in uremia or eclampsia), or on inflammatory disease (encephalitis, meningitis). If this rise in blood pressure did not occur, all functions of the brain would cease as soon as the cerebrospinal pressure exceeded the normal pressure in the cerebral veins and capillar-

ies, and death would occur from respiratory failure and circulatory collapse.

Anatomical arrangements of cerebral circulation

The anatomical arrangements are among the unusual features of the cerebral circulation. Instead of depending on a single artery to supply most of the blood and a single vein to carry it back to the heart, the brain gets its blood from four major arteries and is drained by three separate sets of veins. Any two of the arteries and any one of the veins could meet the needs of the brain if time were given for adjustment, and even one artery is enough to supply the vital centres of the brainstem and thus prevent failure of circulation and respiration. The arteries (the two internal carotids and the two vertebrals) join at the base of the brain to form a circle (the circle of Willis) from which arterial branches are given off to course either through the brain substance or over its surface in the pia mater. The carotid arteries run in the neck and can be occluded by deep pressure, but the vertebrals arise in the chest and soon enter the vertebral column, where they are protected from pressure or injury short of disruption of the spine itself. The cerebral veins follow the arteries; they do not meet in a circle, however, but in a number of wide thin-walled sinuses from which blood leaves by way of the two large internal jugular veins, in a variable network of veins located at the base of the skull, and in large sinuses found inside the vertebral canal. Even if the entire venous return from the upper portion of the body is obstructed (as by thrombosis of the superior vena cava), there is no serious impediment to the drainage of the cerebral circulations.

Weight of the brain. The weight of the human brain varies with age, stature, body weight, sex, and race. It is also influenced by congestion of the blood vessels, degenerative changes, and atrophy. At birth the brain weighs approximately 380 grams (about 13 ounces) and is 12.4 percent of the body weight. The entire brain, with the pia-arachnoid, of an adult male weighs approximately 1,400 grams (about three pounds); that of a female, 1,260 grams (44.6 ounces). The average stature and body weight of the female, however, are less than those of the male, and, when these facts are considered, the size and weight of the brain in the two sexes are approximately equal. The influence of age on brain weight is considerable. The growth of the brain is rapid during the first three years but slightly less so up to the seventh year, when it is almost its full weight. After this the increase is very gradual, its prime weight usually attained in males by the 20th year and in females somewhat earlier. From this period onward, in both sexes, there is a continuous diminution in the average brain weight of approximately one gram per year.

Tall persons have heavier brains than short persons, but, relative to their height, short persons have larger heads and brains than tall persons. Many men of conspicuous ability have had brains of large size, but averages calculated from groups—*e.g.*, scholarship winners and persons of average and below average ability—show statistically that there is only a slight correlation between head size and intelligence.

THE SPINAL CORD

The spinal cord in man is the elongated cylindrical portion of the central nervous system located in the upper two-thirds of the vertebral canal of the vertebral column. Spinal nerves and many of the reflexes stemming from spinal connections are described in detail below in the major sections on the peripheral and autonomic nervous systems.

As with the brain, the spinal cord of man closely resembles that of other higher animals, particularly the primates. Additional detail may be found in the article NERVES AND NERVOUS SYSTEMS.

Anatomy. The spinal cord is about 45 centimetres (18 inches) long. It extends from just above the first cervical vertebra (atlas) at the base of the skull to the level of the first or second lumbar vertebra, where it is known as the conus medullaris. Small groups of nerve fibres attached to the spinal cord unite to form the segmental series of dorsal and ventral roots. One dorsal root and one ventral root combine in the region of an opening, the intervertebral foramen, between the two vertebrae to form one of the 31 pairs of spinal nerves. The dorsal roots contain the nerve fibres that transmit sensory impulses from the body to the spinal cord, whereas the ventral roots are composed of the motor fibres that transmit nerve impulses from the cord to the muscles and glands.

Dorsal and ventral roots

The spinal cord does not occupy the entire length of the vertebral canal because the vertebral column continues to grow after the cord has reached its full length. Since the column is longer than the cord, the nerve filaments of each dorsal root leave the cord above the site where the root enters each intervertebral foramen.

Each thoracic, lumbar, sacral, and coccygeal spinal nerve is named from the vertebra that is just above its exit from the vertebral column. Since there are 12 thoracic vertebrae, five lumbar vertebrae, five fused sacral vertebrae and one (to three) vestigial coccygeal vertebra, there are 12 thoracic, five lumbar, five sacral, and one coccygeal spinal nerves.

There are eight cervical nerves, although only seven cervical vertebrae, because, in addition to a nerve passing below each cervical vertebra, one exists above the first cervical vertebra. The long roots of the lower third of the vertebral canal below the conus medullaris form a structure called the cauda equina ("horse's tail").

Meninges. The cord is enveloped by three coverings, the meninges. From within outward, as with the brain, they are the pia mater, arachnoid, and dura mater. These meninges are continuous with comparable layers surrounding the brain. The pia mater is closely attached to the spinal cord and contains the many blood vessels that supply the cord. The pia mater continues as a strand in the middle of the cauda equina known as the filum terminale, which extends to the coccyx.

The arachnoid layer is connected to the pia mater by delicate strands. The region between the two layers is the subarachnoid space, which is filled with the cerebrospinal fluid. Formed mainly in the brain, the fluid cushions the spinal cord. In the region below the conus medullaris, the arachnoid forms a cul-de-sac—containing the cauda equina and cerebrospinal fluid—into which lumbar puncture needles may be inserted to withdraw cerebrospinal fluid for diagnostic examination and to inject spinal anesthetics. This procedure avoids damage to the spinal cord, and injury to the spinal roots is slight.

The outer dura mater is a tough fibrous tissue surrounding the cord and continuous with the sheaths of the spinal roots and spinal nerves. It extends caudally to the lower sacral region, where it is reduced to a strand, the filum durae matris spinalis, attached to the coccyx. A thin subdural space is located between the dura mater and the arachnoid. Between the dura mater and the vertebral column is the epidural space, which contains many blood vessels, especially veins, and fat. A hypodermic needle may be inserted into this space in the lower sacral region for the injection of certain anesthetics (caudal anesthesia).

Cervical and lumbosacral enlargements. The spinal cord has two enlargements: the cervical enlargement and the lumbosacral enlargement. These regions are large because these segments are involved in the innervation of the upper and lower extremities, respectively. The fifth through eighth cervical segments and the first thoracic segment comprise the cervical enlargement. The spinal nerves arising from the cervical enlargement form the brachial plexus.

Cross section. In cross section the spinal cord consists of an H-shaped gray matter surrounded by white matter. The gray matter is made up of cell bodies of nerve cells (the part of the cell containing the nucleus and other constituents) and nerve fibres that are oriented in the cross-sectional plane because the fibres enter and leave the gray matter from the white matter and spinal roots. The white matter consists of nerve fibres that are oriented parallel to the long axis of the spinal cord. These fibres ascend to and descend from the brain and other levels of the spinal cord. The whiteness is caused by the fatty sheath (myelin sheath) that surrounds many of the fibres.

The gray and the white matter of the cord

Some non-neural neuroglia cells are scattered throughout both the gray and the white matter. The arms of the H that extend dorsally (toward the back of the body) are called dorsal or sensory horns, while the ventrally (toward the front) directed arms are called ventral or motor horns. The crossbar of the H is the gray commissure and the intermediate gray. The gray matter is largest in the regions of the cervical and lumbosacral enlargements because of their relation to the innervation of the limbs. In the middle of the gray commissure is the small central canal, which is continuous with the ventricles of the brain. The gray matter has been subdivided by Rexed into nine lamina, counting from dorsal to ventral. The incoming sensory impulses end in lamina four, five, and six. In lamina six end the proprioceptive messages; in lamina five, the protopathic messages; and in lamina four, the epicritic messages. The white matter is subdivided into three columns: the dorsal column, between the dorsal horns; the ventral column, between the ventral horns; and the lateral columns, at the sides of the gray matter.

Anatomical and functional organization of the spinal cord. The functional anatomy of the spinal cord may be visualized by describing the orientation of the nerve cells of the spinal reflex arcs. The withdrawal of the hand upon touching a hot stove before it is felt is an example of a spinal reflex. The heat stimulates the terminal endings of the thermal neurons in the hand. The resulting impulses are transmitted by the fibres of these neurons through the peripheral nerves and dorsal roots into the dorsal horn. Each sensory nerve cell has its cell body in a dorsal-root ganglion. Thus each nerve cell is a unit structure extending from its peripheral ending in the hand to its termination in the gray matter of the spinal cord.

In the dorsal horn the nerve fibres have synapses (interconnections) with one or more other neurons. The second nerve cells in the arc are called interneurons (internuncial or intercalated neurons); each has a cell body in the gray matter. Some interneurons send axons into the ventral horn and form synapses with the motor cells whose cell bodies are in this horn. The axons of the motor cells reach the voluntary muscles via the ventral roots and peripheral nerves. In short, an exteroceptive stimulus transmitted by a nerve goes to the spinal cord via a dorsal root into the gray matter, then to an interneuron, and finally to the motor cell and the voluntary muscle via the ventral root and the peripheral nerve. This is known as a three-neuron reflex. If the interneuron is not involved, only two cells are in the series, and it is thus a two-neuron reflex.

In the knee-jerk reflex, the sensory cells synapse directly with the motor neurons in the gray matter; it is, therefore, a two-neuron reflex. The interneuron may cross over from the dorsal horn of one side to the ventral horn of the opposite side and synapse with motor cells. This is one means of coordinating one side with the other. An intersegmental reflex is a mechanism relaying impulses to higher and lower spinal cord levels to coordinate large areas of the body. The fibres of such interneurons leave the gray matter and go up, down, or both (fibres may divide into two or more branches) to other spinal levels in that part of the white matter called the fasciculus proprius and re-enter the gray matter to synapse with motor cells.

The white matter of the spinal cord contains nerve fibres that transmit most of the nerve impulses from the body (below the head) to the brain and from the brain to the body. Microscopic examination of the white matter does not give visible evidence of functional subdivisions. On the basis of numerous experimental data, however, and examinations of spinal cords from persons with spinal injuries, the white matter can be partitioned into functional regions called tracts or pathways. A tract, or fasciculus, is a collection of nerve fibres having the same origin, function, and termination. Those fibres transmitting impulses toward the brain form sensory, or ascending, tracts; those transmitting impulses from the brain form motor, or descending, tracts.

Pathways of sensation. So-called epicritic sensations are transmitted by the fibres in the dorsal columns. They include position sense, pressure sense, vibratory sense, sense of movement, and stereognosis. Position sense is the quality that enables a person to know where the parts of his body are in space when his eyes are closed; pressure sense is the quality of sensation that results from a deformation of the skin and may, thus, be called deep touch; vibratory sense is experienced when the stem of a vibrating tuning fork is placed in contact with a bone; and stereognosis, or two-point discrimination, is the faculty of perceiving the weight, form, and shape of objects and of distinguishing the distance between two touch stimuli. Epicritic senations

These senses are transmitted by fibres that pass from the peripheral nerves and dorsal roots into the dorsal horn of the spinal cord. These fibres with their cell bodies in the dorsal-root ganglia have several branches in the gray matter. The shorter branches (collateral fibres) are utilized in the spinal reflex arcs. The main axon continues into the dorsal column to ascend to the lower medulla of the brain. Two interesting features are that (1) a fibre in the dorsal column is a part of the same neuron that was stimulated in the body, and (2) this fibre ascends on the same side of the spinal cord as it enters; *i.e.*, it does not cross the midline of the cord.

As a consequence, the sense of position of the toes, for example, is carried by nerve cells that extend from the toes to the medulla oblongata without a synapse. In addition, injury to the dorsal column on one side of the cord will result in the impairment of the proprioceptive senses on the same side of the body below the level of injury. Each half of the dorsal column is divided into a fasciculus gracilis and a fasciculus cuneatus. The epicritic impulses from the lower half of the body, including the lower extremities, are carried in the fasciculus gracilis, while those from the upper half of the body, including the upper extremities (but excluding the head) are carried in the fasciculus cuneatus.

Some light-touch sensations (*e.g.*, gentle brushing of skin by cotton) are transmitted by fibres in the dorsal column. In the medulla the fibres of each half of the dorsal column synapse with neurons in the nucleus gracilis and nucleus cuneatus. The neurons of these nuclei cross over to the opposite side of the medulla, pass upward as the medial lemniscus, and terminate in the thalamus, from which they relay to the cerebral cortex.

Pain and temperature sensations are transmitted by another pathway (see also SENSORY RECEPTION, HUMAN). The fibres from the periphery, in common with other sensory fibres, enter the dorsal horn of the spinal cord, where their terminal branches establish reflex connections with other neurons. At the spinal level, at which the fibres enter the cord, they synapse with dorsal-horn neurons. Some of these cells are interneurons in the reflex arcs. Other neurons of the dorsal horn have axons that cross the midline to the opposite side through the ventral white commissure (ventral to the gray commissure) to join the lateral spinothalamic tract in the lateral columns. This tract contains more fibres as it ascends in the spinal cord, the new fibres being added from each spinal level. The spinothalamic tract's name is based on the fact that the fibres of the pathway originate in the dorsal horn of the spinal cord and terminate in the thalamus on the opposite side of the brain. Crude pain and awareness of touch may be brought to consciousness in the thalamus. Impulses from the thalamus are relayed to the cerebral cortex. Pain and temperature sensations

In addition to the light-touch fibres of the dorsal column noted above, some light-touch impulses take a different pathway. After entering the dorsal horn, the axons transmitting light touch synapse with neurons whose axons cross the midline to join the ventral spinothalamic tract in the ventral column.

The symptoms resulting from injury to the spinal cord are a consequence of the course of these pathways. If the right half of the spinal cord is cut at a midthoracic level, for example, conscious proprioception would be lost on the right side of the body below the level of the lesion of the fasciculus gracilis (right abdomen and the right

lower extremity). Pain and temperature sensations would be lost on the left side of the body below the level of the lesion of the lateral spinothalamic tract. Because of the dual course of their fibres, light touch would be diminished, but not necessarily lost, below the level of the lesion. Perception of sensations from regions above the cut would be unimpaired.

Some unconscious proprioceptive impulses relay from the gray matter of the cord via the spinocerebellar tracts of the same and opposite sides and terminate in the cerebellum.

The precise course of the pathways for sensations from the viscera (heart, intestinal tract, kidney, and so forth) through the white matter of the cord is not known.

Motor pathways. In the ventral horn of the gray matter are large motor cells whose axons pass through the ventral roots to the spinal nerves. These neurons, which innervate the voluntary muscles, are called lower motor neurons; if they become nonfunctional, as in poliomyelitis, the muscles are deprived of their nervous stimulation and do not contract. These muscles waste away (atrophy) and do not respond to stimulation (areflexia). This paralytic condition is known as a lower motor neuron paralysis.

Lower motor neurons

The lower motor neuron cells are stimulated and influenced by neurons of the spinal reflex arcs and by the descending motor tracts of the white matter. These descending tracts from the brain may synapse directly with the motor cells, but usually interneurons of the gray matter are interposed between the descending tracts and the lower motor neurons.

Collectively the descending motor tracts from the brain are known as upper motor neurons. Their cell bodies are in the brain and their axons pass through the brainstem and white matter of the spinal cord before terminating in the gray matter of the spinal cord. Two major subdivisions of these motor tracts are recognized: the pyramidal tracts, which are phylogenetically recent, and the phylogenetically old extrapyramidal tracts.

The lateral pyramidal tract, or lateral corticospinal tract, arises in the cerebral cortex of the opposite side and crosses over just as its fibres enter the spinal cord. The direct pyramidal tract, or ventral corticospinal tract, originates in the cerebral cortex and crosses over in the spinal cord. The extrapyramidal tracts include the motor tracts that arise in regions of the brain other than the cerebral cortex. These tracts do, however, receive neural influences from the cerebral cortex as well as from other parts of the brain.

The extrapyramidal tracts include (1) the vestibulospinal tract, which transmits impulses from the vestibular (balance) mechanisms of the inner ear; (2) the reticulospinal tracts, which arise from the medulla and pons; (3) the medial longitudinal fasciculus, a collection of fibres from several regions of the brainstem; and (4) the rubrospinal tract from the red nucleus in the midbrain.

Spastic paralysis

When a number of these motor tracts are damaged, an upper neuron paralysis may result. This contrasts with the lower motor neuron paralysis previously mentioned. Here the reflexes are exaggerated; there is loss of voluntary motion; and the extremities become spastic. The condition, called spastic paralysis, results because the muscles are under control of a spinal cord that still mediates spinal reflex arcs but is deprived of its normal innervation from the brain. Spastic paralysis resulting from an injury of the spinal cord is observed on the same side of the body as the lesion.

The complexity of the nervous system may be illustrated by spinal reflex arcs innervating the voluntary muscles. In probably all muscles (since all muscles have not been investigated) there are small fibres arranged in bundles of two to eight fibres that receive their motor innervation from small neurons in the anterior horn and that are endowed with sensory fibres. These fibres go through the spinal ganglia, where their cell bodies are situated, to the posterior horn (lamina six), where they make direct contact with the motor cells in the anterior horn. These motor cells stimulate the voluntary muscles to contract. When the voluntary muscles contract sufficiently, the sensory endings in the tendons are stimulated to send impulses via sensory fibres through the dorsal root back to the spinal cord to inhibit the motor cells of the ventral horn. In brief, the sensory endings in the muscle and tendon have a critical role in muscle activity.

Autonomic nervous system. The autonomic nervous system will be considered in detail in a later section. Briefly, it is concerned with the motor innervation of glands, blood vessels, and the viscera, and it has neurons in the gray matter. The nerve cells of the sympathetic division of this system have their cell bodies in the lateral part of the intermediate gray, known as the intermediolateral cell column, of the thoracic and upper lumbar levels. Their axons leave the spinal cord through the ventral roots. The nerve cells of the spinal part of the parasympathetic division of this system have their cell bodies scattered in the gray matter of the midsacral levels. Their axons leave the spinal cord through the ventral roots. The sympathetic nervous system assists the body in meeting emergencies and stress. This includes increasing the rate and force of the heartbeat, the dilation of the bronchial tubes of the lungs, and the inhibition of the motility and the secretory activity of the digestive system. The parasympathetic system contributes to the conservation of bodily energy. This includes decreasing the rate and force of the heartbeat, the constriction of the bronchial tubes, and the stimulation of the motility and secretory activity of the digestive system.

Intermediolateral cell column

Physiology. *The reflex.* The simplest complete reaction of the nervous system is a reflex. There are many reflexes that are extremely complex, being built up of a number of simpler reflexes combined. A reflex is a reaction started by the environment acting as a stimulus upon some nerve that communicates the excitement to other nerves by connections with them in the central nervous system. The excitement so generated and transmitted finally travels outward from the central nervous system by one or more of the efferent nerves and through these reaches muscles or galnds, producing in them its final effect. The muscles and glands are from this point of view termed effector organs. The reaction is therefore "reflected" from the central nervous system. The nerve structures that include its trajectory are spoken of as a nervous arc.

The whole purpose of the central nervous system is to bring afferent neurons into touch with efferent neurons. The whole purpose of reflex arcs is to bind one part of the organism to another part in such a way that what the environment is doing to the organism at one place may appropriately call forth or restrain movement or secretion in the muscles or glands wherever situated in the organism.

Receptor cells. There is one condition for the performance of these reactions that is not provided by the nervous system itself. The afferent neurons are not in most cases so constituted as to be excitable themselves directly by the environment—they cannot, for example, be stimulated by light. Their amenability to the environment, their sensitization to environmental stimuli, is effected by special cells adjunct to their peripheral ends. These cells form organs called receptors. They are delicately adapted to be stimulated by specific agents and are classifiable into various species, so that each species is easily excited by a particular stimulus that is "adequate" for it and is inexcitable or excitable only with difficulty by stimuli of other kinds. Thus in the skin some receptors are adapted for mechanical stimuli (touch) and not for thermal stimuli, while others (cold spots, warm spots) are adapted for thermal stimuli and not for mechanical. As far as it is known, each afferent neuron is connected with receptors of one species only. The receptors thus confer upon the reflex arcs selective excitability. Each arc is tuned to respond to certain stimuli, while other arcs not having that kind of receptor do not respond. The receptors, therefore, while increasing the responsiveness of the organism to the environment, prevent confusion of reactions (incoordination) by limiting to particular stimuli a particular reaction.

Receptors

Proprioceptors. By the conditions just described, the

body's system of neurons is made accessible to the external environment; in addition to those receptors, there are others lying within the mass of the organism itself that are excitable by actions occurring within the organism. These are called proprioceptors. They are distributed principally in the muscles and in structures functionally adjunct to muscle, such as joints, ligaments, and fasciae. The reactions, both passive and active, induced in such motor structures reflexly by environmental stimuli tend, therefore, secondarily to be followed and accompanied by reflex reactions initiated from proprioceptors.

Conduction. The process by which the excitement generated in the afferent neuron travels along the reflex arc is known as conduction. Conduction along afferent and efferent nerves differs in some important respects from that in the nerve centre—*i.e.*, in the piece of the central nervous system connecting the afferent nerve with the efferent nerve. In a nerve trunk the excited state set up by a stimulus travels along its fibres as a wavelike disturbance, called an impulse, at a speed up to 0.5 and 100 metres (1.5–330 feet) per second. An impulse does not alter in intensity or speed as it travels. A nerve trunk when excited at some point along its length transmits the impulse in both directions (*i.e.*, both up and down each fibre) from the point stimulated. This is true whether the fibre is afferent or efferent. The speed of travel of the nervous impulse along the nerve trunk is the same whether the nervous impulse is weak or intense. The nerve trunk shows practically no delay in its response to an effective, even though weak, stimulus, and its response ceases practically at once on cessation of the exciting stimulus. When excited by repeated brief stimuli, the rhythm of the response corresponds closely with that of the stimuli, even when the frequency of the latter is as high as 500 per second. Finally, nerve-trunk conduction is singularly resistant to fatigue, to impoverished blood supply, and to many drugs that powerfully affect reflex actions.

Through the central nervous system the travel of the impulse differs from these features. Its intensity is likely to be altered in transit. Its transit time is much longer than that for a similar length of nerve trunk. Its direction of transmission becomes polarized—*i.e.*, confined to one direction along the nervous path. To a strong stimulus the central reaction, instead of being as brief as an impulse, may endure for a whole second or more. A stimulus not capable of evoking a response from a centre when applied once may, by simple repetition, become effective—a process known as temporal summation.

Conduction in the gray matter differs in these respects from conduction in nerve trunks. In the gray matter each afferent fibre breaks up into branching threadlets that ramify in various directions and terminate in close apposition with other neurons. The point of junction of one neuron with another is called the synapse. The distance between two neurons at the synapse is about 200 Å (angstrom units; 1 Å = 1/10,000,000 millimetre, or about 1/250,000,000 inch). The absence of cytoplasmic continuity between the neurons is the anatomical basis that accounts for the slight delay in the speed of the impulse at the synapse and for the polarization of the direction of the impulse.

Reflex reactions. When the spinal cord is severed at any point, the reflex arcs of the portion of the body below the transection are cut off from the rest of the nervous system above, including the brain. The reflex reactions elicited from the isolated region cannot, therefore, be modified by the action of the higher nervous centres. In the head the local centres are overlaid by higher centres that cannot by any simple severance be separated from them. By studying the powers of the cord behind a complete spinal transection, information has been obtained as to the powers of the purely local, or segmental, reflex mechanisms.

The flexion reflex

The so-called flexion reflex of the limb is one of the most accessible of the local reflex reactions that can be studied by means of an isolated portion of the spinal cord. If the isolated portion serves the leg, the following phenomena are noted. The three main joints of the leg are the hip, the knee, and the ankle. Each of these joints is provided with muscles that flex (bend) it and with others that extend (straighten) it. It is found that reflex mechanisms throw into contraction the flexor muscles of each of these joints. It matters little which of all the various afferent nerves of the limb is stimulated; whichever afferent nerve may be involved, the discharge goes to practically the same muscles, namely, always to the flexors of the joints.

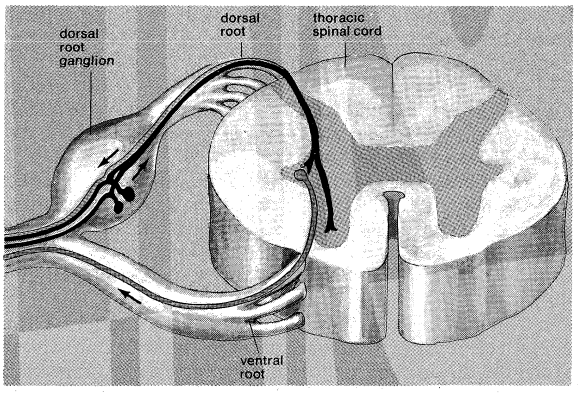


Figure 8: The spinal reflex.

After Frank H. Netter, M.D., *The CIBA Collection of Medical Illustrations;* Copyright © 1957, 1972 by CIBA Pharmaceutical Company, Division of CIBA-GEIGY Corporation; reproduced with permission. All rights reserved

Not only does the reflex action not discharge motor impulses into the nerves of the extensor muscles, but, if the spinal cord happens to be discharging impulses into these nerves when the reflex is evoked, this discharge is suppressed or diminished (inhibited). In this way the latter muscles are prevented from impeding the action of the contracting flexors. This inhibition prevents other reflexes from upsetting, for the time being, the due action of the flexion reflex, for it renders the muscles opposing that reflex less accessible to motor discharge through the spinal cord, whatever the source of that discharge.

A feature of this reflex is its graded intensity. A weak stimulus evokes in the flexor muscles a weak contraction and in the extensor muscles a slight relaxation. In the case of the muscle fibre, any strength of stimulus, if sufficient to excite it at all, totally excites the fibre and evokes its full contraction; and the contraction remains the same, quantitatively, as long as the mechanical and nutritive conditions of the fibre remain constant. With each muscle fibre reacting on this all-or-none principle, and each nerve fibre similarly responding, different amounts of contraction yielded by the muscle in response to different strengths of stimulation to its nerve signify differences in the number of muscle fibres the stimulus excites. In other words, the weaker stimuli excite fewer of the nerve fibres innervating the muscle.

When the tension developed by the muscle reaches the maximum obtainable from it, that fact indicates that all the muscle fibres are then in contraction; when the tension development is less than maximal, some of the muscle fibres are not in action. A stimulus to the nerve that evokes maximal tension from the muscle is one that succeeds in exciting all the nerve fibres innervating the muscle, and any stimulus that evokes less than maximal tension is having some of the nerve fibres idle.

The tension developed by a muscle under reflex excitation falls short of its maximal tension, even when the afferent nerve is stimulated maximally. The reflex activates only a proportion of the total aggregate of the muscle's fibres and, therefore, only a proportion of the spinal motor neurons innervating the muscle. But, within the range of the limit imposed by the maximal number of muscle fibres that it can activate, the grading of the number activated in accordance with the grading of the strength of stimulation of the afferent nerve is delicate, much more so than in the case of the motor nerve itself. The actual fractional proportion of the muscle's total aggregate of motor units that an afferent nerve, maximally stimulated, can activate differs from one afferent nerve to another and for the same afferent nerve under different conditions. The threshold value of stimulus for the reflex

is usually higher than for the contraction of the muscle from the motor nerve.

Spinal reflex action no more regards "muscles" than does a motor centre of the cerebral cortex. Both ignore muscles as entities and have in view only movements. A weak flexion reflex activates a few fibres of each of the several flexor muscles of the limb. The reflex treats a complex composed from some (*i.e.*, those of similar threshold for the particular afferent nerve) motor units of hip flexor, of knee flexor, and of ankle flexor as being functionally more homogeneous than the total group of fibres composing any one of these involved muscles alone.

Reflex inhibition

Reflex inhibition has its seat in the gray matter of the spinal centre. It acts at some point in the reflex arc above from the all-or-none motor unit (motor nerve fibre with its group of muscle fibres). The inhibitory reflex is, like the excitatory, capable of delicate grading; its effect in diminishing or precluding a specific amount of an excitatory reflex is effected entirely by regulating the number of the motor units thrown out of action or kept out of action. Intensity of contraction or of relaxation is thus in both cases simply a question of number of motor units reacting. The reflex centre continues discharging impulses for a certain time after its exciting stimulus has ceased. This after-discharge succeeding a strong stimulus may persist even for several seconds.

Refractory phase. Besides characteristics common to all or many spinal reflexes, certain spinal reflexes have features peculiar to themselves or exhibited by them in degrees not obvious in other reflexes. One of these features is a refractory phase, exemplified by the scratch reflex. In the dog, cat, and many other animals, the hindlimb often performs a rapid scratching movement, the foot being applied to the skin of the shoulder or neck as if to groom the hairy coat in that region. This movement is, in the intact animal, under control of the brain and can be executed or stopped at will. When certain of the higher centres in the brain have been destroyed, this scratching action occurs very readily and in an uncontrolled way. When the spinal cord has been severed in the neck, this scratching movement of the hindlimb can be elicited with regularity as a spinal reflex by merely rubbing, or applying a weak electric current to, the skin of the side of the neck or shoulder.

Mechanism of the scratch reflex

In this reflex the stimulus excites afferent nerves connected with the hairs in the skin, and these convey impulses to the spinal centres in the neck or shoulder segments; these in turn discharge impulses into nerve fibres entirely intraspinal, passing backward along the cord to reach motor centres in the hindlimb region. These motor centres in turn discharge impulses into the muscles of the hindlimb of the same side of the body as the shoulder that is the seat of irritation. The motor discharge is peculiar in that it causes the muscles of the hindlimb to contract rhythmically at a rate of about four contractions per second, and the discharge is peculiar further in that it excites the flexor and extensor muscles of the joints alternately, so that at the hip, for instance, the limb is alternately flexed and extended, each single phase of the movement lasting about an eighth of a second.

This rhythmic discharge remains the same in rate whether the exciting stimulus applied to the skin is continuous or one of many various rates of repetition. Evidently at some point in the reflex arc there is a mechanism that, after reacting to the impulses reaching it, remains for a certain brief part of a second unresponsive, and then becomes once more for a brief period responsive, and so on. This phasic alternation of excitability and inexcitability repeats itself throughout the continuance of the reflex. The phase of inexcitability is called the refractory phase. Its seat lies in the spinal centre. A similar element probably forms part of the coordinating mechanism for many other cyclic reflexes, including those of the stepping of the limbs, the movement of the jaw in mastication, the action of the eyelids in blinking, and perhaps the respiratory movements of the chest and larynx.

Fatigue. Nerve trunks do not tire easily under even the most prolonged stimulation. Reflex actions, on the other hand, tire relatively soon, some being more resistant to fatigue than others. The flexion reflex may be continued for ten minutes at a time. As a reflex tires, the muscular contraction that it causes tends to become less intense and less steady. The relatively rapid onset of fatigue in reflexes is counterbalanced by speedy recovery in repose. A fatigued long-flexion reflex, although weak, tremulous, and irregular, after 30 seconds' rest will recommence with almost the same vigour and steadiness that it had previously.

Reflex postures. Reflex postures are specific reflexes that persist for long periods at a time. The hindlimbs of a spinal frog (one in which the spinal cord has been surgically isolated from the brain) assume an attitude that is reflex, for it ceases on severance of the afferent spinal roots. This attitude is one of flexion at hip, knee, and ankle, resembling the well-known natural posture of the frog as it squats when quiet. Similarly, in the decerebrate mammal certain muscles (*e.g.*, the limb extensors) exhibit a persistent contraction. These tonic reflexes are related to attitudes. In the dog and cat they are exhibited by those muscles whose action antagonizes gravity in postures that are usual in the animal; thus the extensors of the knee and hip and shoulder and elbow are in tonic contraction during standing. The postural contraction is accompanied by electrical "action currents" like those of other contractions, though weak and indicative of self-smothering from the asynchronous rhythms in the contributory fibre groups. The postural contraction of the extensor muscles in the maintenance of the erect posture is traceable to a reflex, called the stretch reflex. A stretch (*e.g.*, by a pull upon the tendon) lengthening the muscle by even so little as 0.5 percent excites reflex contraction of some of its fibres. The passive stretch is a mechanical stimulus to some of the proprioceptive organs (perhaps the muscle spindles) in the muscle. The reflex is therefore unobtainable when the afferent nerve fibres of the muscle itself have been severed. The essential centre for the stretch reflex is spinal, but midbrain and cerebellar centres greatly reinforce it. If the "antigravity" muscles are not checked from yielding by the stretch reflex—which their own passive yield induces in themselves—the head, neck, tail, and jaw droop, and the limbs fold up under their own weight and that of superincumbent parts. Hence the reflexly maintained standing posture is a multiple stretch reflex. Considerable latitude of actual pose is allowed to the individual parts because each muscle involved acts for itself and develops and regulates its own stretch reflex.

The stretch reflex

Characteristic of reflex postural contraction is its continuance for long spells at a time without obvious fatigue. Attempts to reproduce this feature by electrical stimulation of afferent nerves or of the muscle's motor nerve are uniformly unsuccessful. Thus, the suggestion has been made that the postural contraction of muscle is brought about by a process in the muscle fibre different from that of the ordinary contraction so easily studied under electrical stimulation in the laboratory. The explanation has been offered that the postural contraction, although the same process as that of ordinary reflex and other contraction, usually involves relatively few motors units at a time and these, as they fatigue, are replaced by others. The stretch, which is the essential stimulus, is automatically shifted to fresh motor units as soon as those already in operation begin to fatigue.

The eyeballs are organs that employ active postures (the gaze being a posture) throughout the waking day. The proprioception of the neck muscles excites from the cervical spinal cord reflex influences acting on the nerves of the eyeball muscles. For example, turning the neck toward one side excites turning of the eyeballs toward the opposite side. Or again, a clockwise partial rotation of the neck round its long axis excites from the neck proprioceptor a compensatory counterclockwise shifting of the eyeballs. These reflex postures of the eyeballs are sometimes called compensatory because they tend to keep the gaze level despite displacements of the head by the movements of the neck.

When the cord or the medulla and cord of a mammal are surgically isolated from the brain, the animal's com-

posite stretch reflex enables it when placed in the erect posture to maintain that attitude, even for hours at a time; but it is incapable of assuming that posture when laid, for instance, on its side. But, when the midbrain is retained in addition to hindbrain and cord, the animal has that power of righting itself. These righting reflexes are likewise postural and largely evoked by the otolith organs in the vestibular portion of the ear. Although not truly spinal reflexes, their field of nervous operations overruns into the spinal cord.

Interaction between reflexes. At the start of every reflex arc is a receptive neuron extending from the receptive surface to the central nervous system. This neuron forms the only avenue that can be used by impulses generated at its receptive point, regardless of their destination. The neuron is, therefore, a path exclusive to the impulses generated at its own receptive point; receptive points other than its own cannot employ it. A single receptive point may play reflexly upon a considerable number of different effector organs. It may be connected through its reflex path with muscles and glands in many different regions. Nevertheless, all its reflex arcs arise from the one single shank or stem—*i.e.*, from the one afferent neuron that conducts from the receptive point at the periphery into the central nervous system.

Reflex arcs as receptive neurons

But at the termination of every reflex arc is a final neuron, the ultimate conductive link to an effector organ (muscle or gland). This last link in the chain (*e.g.*, the motor neuron) differs, obviously, in one important respect from the first link of the chain. It does not subserve exclusively impulses generated at one single receptive source but receives impulses from many such sources situated in various regions of the body. It is the only path that all impulses, no matter what their source, must travel if they are to act on the muscle fibres to which that path leads.

Therefore, while the receptive neuron forms a private route exclusively serving impulses of one source only, the final or efferent neuron is, so to say, a public path common to impulses arising at any of many sources of reception. A receptive field (*e.g.*, an area of skin) is analyzable into receptive points. One and the same effector organ is in reflex connection not only with many individual points but even with many various receptive fields. Reflexes generated in manifold sense organs can pour their influence into one muscle. Thus a limb muscle is the end point of many reflex arcs arising in various parts of the body. Its motor nerve is a path common to all the reflex arcs that reach that muscle.

Reflexes arcs show, therefore, that the initial neuron of each is a private path belonging exclusively to a single receptive point (or small group of points), that finally the arcs discharge into a path leading to an effector organ, and that their final path is common to all receptive points wherever they may lie in the body, so long as they have connection with the effector organ in question. Before finally converging upon the motor neuron, the arcs converge to some extent. Their private paths join interneuronal paths common in various degrees to groups of private paths. The terminal path may, to distinguish it from interneuronal common paths, be called the final common path. The motor nerve to a muscle is a collection of final common paths.

The final common path

There are certain consequences from this arrangement. One of these is the preclusion of essential qualitative difference between nerve impulses arising in different afferent nerves. If two conductors have a tract in common, there can hardly be essential qualitative difference between their modes of conduction.

A second consequence is that, each receptor being dependent for final communication with its effector organ upon a path not exclusively its own but common to it with certain other receptors, such junction necessitates successive and not simultaneous use of the common path by various receptors using it to different or opposed effect.

In the simultaneous correlation of reflexes, some reflexes combine harmoniously, being reactions that mutually reinforce. These are called allied reflexes, and the neutral arcs that they employ are allied arcs. On the other hand, some reflexes, as mentioned above, are antagonistic one to another and are incompatible. These do not mutually reinforce; instead, one of them inhibits the other.

Allied reflexes. When trains of impulses, travelling by convergent afferent arcs, meet at the same common path (as frequently happens), certain phenomena occur. For example, an afferent path, *A*, discharges into the intraspinal motor neurons of the knee flexor and can activate 75 percent of them. An afferent path, *B*, also discharges into the motor neurons of the same muscle and can activate 75 percent of them. If the muscle is reflexly contracting to *A* with a tension 75 percent of that of the muscle's maximal contraction, and stimulation of *B* is added, the contraction tension will rise to 85 percent of the maximal. The inference is that, for 65 percent of the total motor units of the muscle, the excitation from *A* acts upon the same motor units as would that from *B*. In other words, 65 percent of the total aggregate of the motor units of the muscle are common to *A* and to *B*. The stream of impulses arriving at these common motor units from *B* when *A*'s stream is already activating them leaves their activity unchanged, failing either to increase or to diminish it. *A*'s stream, which is engaging them, precludes *B*'s stream from engaging them. This is called occlusion.

Occlusion

The final common path under the competition of allied arcs is comparable with a telephone line already busy for one subscriber and for that time engaged and unavailable to others. The action of the principle of the final common path may be instanced in regard to allied arcs in the scratch reflex as follows. If, while the scratch reflex is being elicited from a skin point at the shoulder, a second point one centimetre (0.4 inch) from the first point, but also in the receptive field of skin, is stimulated, the stimulation at this second point favours the reaction from the first point. This is clear when the stimulus at each point is of subminimal intensity. The two stimuli, each unable separately to invoke the reflex, do so when both are applied at the same time. The receptive field of a reflex is really the common area of commencement of a number of allied arcs.

Antagonistic reflexes. Not all reflexes connected to one and the same common final path supplement one another. If, for instance, during the scratch reflex, a stimulus is applied to the foot not of the scratching side but of the opposite side, then the left leg, which is executing the scratch reflex in response to stimulation of the left shoulder skin, is cut short in its movement by the stimulation of the right foot, although the stimulus at the shoulder to provoke the scratch movement is maintained unaltered all the time. The stimulus to the right foot will temporarily interrupt a scratch reflex or will cut it short or will delay its onset, depending on the time relations of the stimuli. The inhibition of the scratch reflex occurs sometimes when the contraction of the muscles innervated by the reflex conflicting with it is very slight. There is interference between the two reflexes, and the one is inhibited by the other.

The reflex from the right foot straightens (extends) the opposite leg at the knee; in doing this it causes steady excitation of extensor neurons of that knee and steadily inhibits the flexor neurons. But the scratch reflex causes rhythmic excitation of the flexor neurons. In this conflict the flexor neurons, as a final common path, therefore lie under the influence of two antagonistic reflexes, one of which would excite them to rhythmical discharge four times a second, while the other would continually repress all discharge in them.

In all these forms of interference there is a competition, as it were, between the excitatory stimulus used for the one reflex and the excitatory stimulus for the other. Both stimuli are in progress together, and the one in taking effect precludes the other's taking effect as far as the final common path is concerned.

Competition between stimuli

Again, if, while stimulation of the skin of the shoulder is evoking the scratch reflex, the skin of the hindfoot of the same side is stimulated, the scratching may be arrested. Stimulation of the skin of the hindfoot in a way that has the character of threatening the part with damage (nociceptive stimulus) causes the leg to be flexed, drawing the

foot up by steady maintained contraction of the flexors of the ankle, knee, and hip. Here, therefore, is an arc that discharges into a final path, common to it and to the scratch reflex arc; both these arcs employ the same effector organ, namely, the knee flexor, and employ it by the common medium of the final pathway. But, though the channels for both reflexes discharge into the same final common path, the excitatory flexor effect, specific to each, differs strikingly in the two cases. In the scratch reflex, the flexor effect is an intermittent effect; in the nociceptive flexion reflex, the flexor effect is steady and maintained. The scratch reflex is set aside by that of the nociceptive arc from the homonymous foot. (The term homonymous is used here and below to indicate a common locus or other intimate experimental relationship.)

The stimulation that previously sufficed to provoke the scratch reflex is no longer effective, though it is continued all the time. But, when the stimulation of the foot is discontinued, the scratch reflex returns. In that respect, although there is no enforced inactivity, there is an interference tantamount to, if not the same thing as, inhibition. Though there is no cessation of activity in the motor neuron, one form of activity being impressed upon it is cut short, and another takes its place. A stimulation of the foot, too weak to cause more than a minimal reflex, often will suffice completely to interrupt, cut short, or prevent onset of the scratch reflex.

The whole point of the interference between the homonymous flexion reflex and the scratch reflex is that both employ the same final common pathway to different effect—just as in the interference between the crossed extension reflex and the scratch reflex. The motor neuron to the flexor of the knee being taken as a representative of the final common path, the homonymous flexion reflex inhibits it from discharging. Hence if, while the direct flexion reflex is in progress, the crossed foot is stimulated, the reflex of the knee flexor is inhibited. The crossed extension reflex, therefore, inhibits not only the scratch reflex but also the homonymous flexion reflex.

Further, in all these interferences between reflexes, the direction taken by the inhibition is reversible. Thus the scratch reflex is not only likely to be inhibited by, but is itself able to inhibit, either the homonymous flexion reflex or the crossed extension reflex; likewise, the homonymous flexion reflex is not only capable of being inhibited by the crossed extension reflex but conversely in its turn can inhibit the crossed extension reflex. These interferences are therefore reversible in direction. Certain conditions determine which reflex among two or more competing shall obtain mastery over the final common path and thus obtain expression.

Reversibility of interferences

The intimate nature of the mechanism that thus, by summation or by interference, gives coordination where neurons converge upon a common path is unknown. In the central nervous system of vertebrates, afferent neurons *A* and *B* in their convergence toward and impingement upon another neuron, *Z*, toward which they conduct, do not make any lateral connection directly with one another—at least, there is no clear evidence that they do. It seems, then, that the only structural link between *A* and *B* is neuron *Z* itself. *Z* itself should, therefore, be the field of coalition of *A* and *B* if they transmit allied reflexes.

Factors determining the sequence. The formation of a common path from tributary converging afferent arcs is important; the dominant action of one afferent arc or of a set of allied arcs is subject to supersession by another (afferent arc or set of allied arcs), and the supersession normally occurs without intercurrent confusion.

Whatever may be the nature of the physiological process occurring between the competing reflexes for dominance over the common path, the issue of their competition, namely, the determination of which one of the competing arcs shall, for the time being, reign over the common path, is largely conditioned by four factors: spinal induction, relative fatigue, relative intensity of stimulus, and the functional species of the reflex.

1. *Spinal induction.* This refers to the ability of one spinal reflex to lower the threshold of another. Induction occurs in two forms: immediate and successive. In immediate induction the stimulus that excites a reflex tends by central spread to facilitate and lower the threshold for reflexes allied to that which it particularly excites. There thus tends to be formed a constellation of reflexes that reinforce each other, so that a reflex figure results. If the stimulus shifts, allied arcs are by the induction particularly prepared to be responsive to it or to a similar stimulus.

Immediate induction occurs only between allied reflexes. Its tendency in the competition between afferent arcs is to fortify the reflex just established or, if transition occurs, to favour transition to an allied reflex. Immediate induction seems to obtain with highest intensity at the outset of a reflex, or at least near its commencement.

Successive induction is in several ways the reverse of immediate induction. If the crossed extension reflex of the limb of a spinal dog (one whose spinal cord has been severed from the brain) is elicited at regular intervals, say once a minute, by an electrical stimulus of defined duration and intensity, the resulting reflex movements are repeated each time with almost constant character, amplitude, and duration. If in one of the intervals a strong, prolonged (*e.g.*, 30 seconds) flexion reflex is reduced from the limb yielding the extensor reflex movement, the latter reflex becomes intensified after the intercurrent flexion reflex. The intercalated flexion reflex lowers the threshold for the aftercoming extension reflexes and especially increases their afterdischarge. This effect may endure, progressively diminishing, through four or five minutes, as tested by the extensor reflexes at successive intervals. Now, during the flexion reflex, the extensor arcs were inhibited; after the flexion reflex, they are in this case evidently in a phase of excitability. The phenomenon presents obvious analogy to visual contrast. The aftereffect may ensue with such intensity that simple discontinuance of the stimulus maintaining one reflex is immediately followed by spontaneous appearance of the antagonistic reflex.

The so-called mark-time reflex of the spinal dog is an alternating stepping movement of the hindlimbs that occurs on holding the animal up so that its limbs dangle. The reflex can be inhibited by stimulating the skin of the tail. On cessation of that stimulus, the stepping movement sets in more vigorously and at a quicker rate than before. This afterincrease might be explicable in either of two ways. It might be due to the mere repose of the reflex centre, the repose so recruiting the centre as to strengthen its subsequent action. But a similar period of repose obtained by simply supporting one limb—which causes cessation of the reflex in both limbs, the stimulus being stretch of the hip flexors under gravity—is not followed by afterincrease of the reflex. Or the afterincrease might result from the inhibition being followed by a rebound to superactivity. This latter seems to be the case. The afterincrease occurs even when both hindlimbs are passively lifted from below during the whole duration of the inhibitory stimulus applied to the tail. And the reflex inhibition of the knee extensor by stimulation of the central end of its own nerve is followed by marked rebound to superactivity of the extensor itself. Again, the knee jerk, after being inhibited by stimulation of the hamstring nerve, is more brisk than before the inhibition.

Inhibition followed by superactivity

By virtue of this spinal contrast, therefore, the extension reflex predisposes to, and may actually induce, a flexion reflex, and conversely the flexion reflex predisposes to and may actually induce an extension reflex. This process may play a part in linking reflexes together in alternating sequence.

A reflex movement in its progress must generate a number of further stimuli and throw up a shower of impulses from the moving muscles and joints into the spinal cord. Squeezing of muscles and stimulation of their afferent nerves and those of joints and other structures elicit reflexes. The primary reflex movement might be expected, therefore, of itself to initiate further reflex movement, and that secondarily to initiate further still, and so on. Yet on cessation of the external stimulus to the foot in the flexion reflex, the whole reflex comes to an end, usually at once.

2. *Relative fatigue.* Another condition influencing competition among reflexes of different sources for possession of one and the same final common path is fatigue. A spinal reflex under continuous excitation or frequent repetition becomes weaker, and it may cease altogether. This decline is progressive and takes place earlier in some kinds of reflexes than in others. In the spinal dog, for example, the scratch reflex under ordinary circumstances tires much more rapidly than does the flexion reflex.

3. *Relative intensity of stimulus.* In the transition from one reflex to another, a final common path changes hands and passes from one control to another. A fresh set of afferent arcs becomes dominant on the supersession of one reflex by the next. Of all the conditions determining which one of competing reflexes shall, for the time being, reign over a final common path, the intensity of reaction of the afferent arc itself relative to that of its rivals is probably the most powerful. Other things being equal, an afferent arc that strongly stimulates is more likely to capture the common path than is a weaker one. A stimulus can establish its reflex and inhibit an opposed one only if it has intensity. This explains why, in order to produce examples of spinal inhibition, recourse has so frequently been made to strong stimuli. A strong stimulus will inhibit a reflex in progress although a weak one will fail. In the same way, any condition that raises the excitability and responsiveness of a nervous arc will give it power to inhibit other reflexes, just as it would if it were excited by a strong stimulus.

Crossed reflexes

Crossed reflexes are usually less easy to provoke, less reliable of obtainment, and less intense than are direct reflexes. Consequently, crossed reflexes usually are more easily inhibited and replaced by direct reflexes than are the latter by the former. Thus the crossed stepping reflex is easily replaced by the scratch reflex, though its stimulus may be continued all the time and though the scratch reflex itself is not a very potent reflex. But the reverse can occur with suitably adjusted intensity of stimuli.

Again, in the dog's leg, the flexion reflex when fully developed is accompanied by extension in the opposite leg. This crossed extensor movement, though often vigorous, may be considered as an accessory and weaker part of the whole reflex, of which the prominent part is flexion of the homonymous limb. When the flexion reflex is poorly elicitable—as, for instance, in spinal shock or under fatigue or weak excitation—the crossed extension does not appear. But, where the flexion reflex is well developed, if not merely one foot but both are stimulated simultaneously with stimuli of fairly equal intensity, steady flexion at knee, hip, and ankle occurs in both limbs, and extension in neither. The part of each reflex that affects the opposite side is inhibited by the flexion on its own side. In other words, the more intense part of each reflex obtains possession of the final common paths at the expense of the less intense portion of the reflex. But if the intensity of the stimuli applied to the right and left feet is not balanced closely enough, the crossed extension of the reflex excited by the stronger stimulus is found to exclude even the homonymous flexion that the weaker stimulus would otherwise evoke from the leg to which it is applied.

As was pointed out previously, in a number of cases the transference of control of the final common pathway from one afferent arc to another is reversible. Other factors being equal, the direction of the transference can be easily governed by making the stimulation of one receptor or another the more intense. A factor largely determining whether a reflex will succeed another or not is, therefore, intensity of stimulus.

4. *Functional species of reflex.* The fourth main determinant for the issue of the conflict between rival reflexes is the functional species of the reflexes. Reflexes initiated from a species of receptor apparatus that may be termed nociceptive, or pain receptors, appear to dominate particularly the majority of the final common paths issuing from the spinal cord. In the simpler sensations experienced from various kinds of stimuli applied to the skin, there can be distinguished those of touch, of cold, of warmth, and of pain. The pain ending may be regarded as adapted to a whole group of excitants that have, in relation to the organism, one feature in common, namely, a noxious, or injurious, character.

Domination of the final common path by nociceptive reflexes

Considering the organism's liability to various kinds of mechanical and other damage, in a world beset with dangers amid which the individual and species have to win their way in the struggle for existence, noxious stimuli may be regarded as part of a normal state of affairs. The skin has evolved a specific sense of its own injuries. As psychical adjunct to the reactions of that apparatus, a strong displeasurable effective quality is found in the sensations they evoke. This may be a means for branding upon memory, of however rudimentary a kind, a feeling from past events that have been perilously critical for the existence of the individuals of the species. In other words, if it is admitted that damage to such an exposed sentient organ as the skin must in the evolutionary history of animal life have been sufficiently frequent in relation to its importance, then the existence of a specific set of nerves for skin pain seems to offer no genetic difficulty, any more than does the clotting of blood or innate immunity to certain diseases. That these nerve endings constitute a distinct species is argued by their all evoking not only the same species of sensation but the same species of reflex movement as regards such attitudes as purpose, intensity, and resistance to shock. Their evolution may well have been unaccompanied by evolution of any specialized end organ, since the naked free nerve endings would better suit the wide and peculiar range of stimuli to which reaction is in this case required. A low threshold was not required because the stimuli were all intense, intensity constituting their harmfulness; but response to a wide range of stimuli of different kinds was required, because harm might come in various forms. That responsive range is supplied by the naked nerve itself and would be cramped by the specialization of an end organ. Hence, these nerve endings remained free.

Domination of pain reflexes

In the spinal animal—in which pain is of course nonexistent—the prepotent (dominant) reflexes are excited by stimulation of areas in which pain can be excited most intensely in intact animals and, likewise, by stimuli that are most fitted to excite pain in intact animals. The nervous arcs of pain nerves, broadly speaking, dominate the spinal centres in peculiar degree. Pain is thus the psychical adjunct of an imperative protective reflex. The term pain nerves is not used, however, since no sensation of any kind can be shown to enter into the merely spinal and reflex aspect of the reaction of these nerves. Because the feature common to all this group of stimuli is that they threaten or actually commit damage to the tissue to which they are applied, a convenient term for application to them is nocuous. Thus, the nerves that, from the point of view of sense, are cutaneous pain nerves are, from the point of view of reflex action, conveniently termed nociceptive nerves.

In the competition between reflexes the nociceptive usually dominate. This explains why such stimuli have been used so much to evoke reflexes in the spinal frog and why, judging from them, such importance for the fate of the organism belongs to spinal reflexes.

One and the same skin surface will evoke in the hindlimb of the spinal dog one or other of two diametrically different reflexes depending on whether the mechanical stimulus is a harmful or a harmless touch. A needle prick to the sole of the foot invariably causes the drawing up of the limb: the flexion reflex. A harmless smooth contact, on the other hand, causes extension: the extensor thrust described above. This flexion is therefore a nociceptive reflex. But the scratch reflex—which is so readily evoked by simple light irritation of the skin of the shoulder—is relatively mildly nociceptive. When the scratch reflex and the flexion reflex are in competition for the final neuron common to them, the flexion reflex more easily dispossesses the scratch reflex from the final neuron than does the scratch reflex the flexion reflex. If both reflexes are fresh, and the stimuli used are such that, when employed separately, they evoke their respective reflexes with some intensity, it is the flexion reflex that is usually prepotent. Yet if a strong stimulus suitable for producing the scratch reflex is applied while the flexion reflex is being moder-

ately evoked, the steady flexion is replaced by the rhythmic scratch reflex, and this occurs even though the stimulus for the flexion reflex is maintained unaltered. When the stimulus producing the scratch reflex is discontinued, the flexion reflex reappears as before.

Skin receptors concerned with sex

Besides those receptors that react to direct noxious stimuli, the skin has others concerned with functions of vital importance to the species and bound up with sensations of similarly intensive affective quality—*e.g.*, those concerned with sexual functions. It would seem to be a general rule that reflexes arising in species of receptors that, considered as sense organs, provoke strongly affective sensation prevail (other things being equal) over reflexes of other species when in competition with them for the use of the final common path.

One great function of the tonic reflexes is to maintain habitual attitudes and postures. They form, therefore, a nervous background of active equilibrium. It is obviously advantageous that this equilibrium should be easily upset, so that the animal may respond agilely to the passing events.

Summary of spinal cord physiology. Intensity of stimulation, fatigue and freshness, spinal induction, functional species of reflex—all these physiological factors influence the result of the interaction of reflex arcs at a common path. All resolve themselves ultimately into intensity of reaction. Thus intensity of stimulus means, as a rule, intensity of reaction. Those species of reflex that are habitually prepotent in interaction with others are those that are habitually intense, and those specially impotent in competition are those habitually feeble—*e.g.*, the tonic reflexes of skeletal muscle. The tonic reflexes of attitude are thus easily interfered with and temporarily suppressed by intercurrent reflexes (those set up by passing events), which have the higher intensity.

Causes of variability in reflex reactions

The high variability of reflex reactions from one experiment or observation to another is often attributed to changes in nutrition and local blood supply. Variability, however, seems far more often due to changes produced in the central nervous organ by its own functional conductive activity apart from fatigue. This activity itself causes, from moment to moment, the temporary opening of some connections and the closure of others. The chains of neurons, the conductive lines, have been clearly revealed by the microscope. Anatomical tracing of nerve fibres may be likened to tracing the distribution of blood vessels before the vasomotor mechanism was discovered. (The blood vessels of an organ may be turgid at one time, constricted almost to obliteration at another.) With the conductive network of the nervous system the temporal variations are even greater, for they extend to absolute withdrawal of nervous influence. Under reflex inhibition a skeletal muscle is relaxed to its postmortem length *i.e.*, there may then be no evidence of even a tonic influence on it by its motor neuron.

The final common path is handed from some group of a plus class of afferent arcs to some group of a minus class or of a rhythmic class, and then back once more to one of the previous groups, and so on. The conductive web changes its functional pattern to and fro—within certain limits. It changes its pattern at the entrances to common paths. The changes in its pattern occur there as interactions between rival reflexes: occlusion, substitution by equivalence, inhibition, immediate induction, successive induction, and fatigue are factors. A new stimulus striking the receptive surfaces causes in the central organ a shift of functional pattern of the linkage. The central organ is a vast network whose lines of conduction follow a certain scheme or pattern, but within that pattern the details of connection are, at the entrance to each common path, mutable. The gray matter may be compared with a telephone exchange, where, from moment to moment, though the end points of the system are fixed, the connections between starting points and terminal points are changed to suit passing requirements. To comprehend the exchange at work, it is necessary to add to its purely spatial plan the temporal one—that within certain limits the connections of the lines shift to and fro from minute to minute. An example is the reciprocal innervation of antagonistic muscles: when one muscle of the antagonistic pair is thrown into action, the other is thrown out of action.

Reciprocal innervation

Reciprocal innervation is a widespread case of the general rule that antagonistic reflexes interfere where they discharge into the same final common paths. And that rule is part of the general principle of the mutual interaction of reflexes that impinge upon the same common path: unlike reflexes have successive but not simultaneous use of the common path; like reflexes mutually reinforce each other on their common path. The principle may be expressed teleologically: the common path, although economically subservient for many and various purposes, is adapted to serve but one purpose at a time; hence it is a coordinating mechanism and prevents confusion by restricting the use of the organ to but one action at a time.

In the case of simple antagonistic muscles, and in the instances of simple spinal reflexes, the shifts of conductive pattern due to interaction at the mouths of common paths are of but small extent. The coordination covers, for instance, one limb or a pair of limbs. But the same principle extended to the reaction of the great arcs arising in the proficient receptor organs of the head—*e.g.*, the eye—which deal with wide tracts of musculature as a whole, operates with more multiplex shift and wider ambit. Releasing forces acting on the brain from moment to moment shut out from activity whole regions of the nervous system, as they conversely call vast other regions into play. The resultant singleness of action from moment to moment is a keystone in the construction of the individual, whose unity it is the specific office of the nervous system to perfect. The interference of unlike reflexes and the alliance of like reflexes in their action upon their common paths seem to lie at the very roots of the great psychical processes.

The spinal cord not only accommodates reflexes centring wholly within the cord itself but also supplies conducting paths for nervous reactions initiated by impulses derived from afferent spinal nerves—even those involving mechanisms situated altogether headward of the cord, in the brain. Many of these reactions affect consciousness, occasioning sensations of various kinds.

Besides the paths followed by headward-running impulses, the spinal cord contains paths for impulses passing along it backward from the brain. These paths lie in the lateral and ventrolateral columns of the cord. The fibres of which they are composed cross but little in the cord. Their sources are various; some of them come from the hindbrain and some from the midbrain, and, in man, a large tract of fibres in the lateral column (the crossed pyramidal tract) comes from the cortex of the forebrain. This last tract is the main medium by which impulses initiated by electrical stimulation of the motor cortex reach the motor neurons of the cord and through them influence the activity of the skeletal muscles.

Of the function of the other tracts descending from the brain into the cord little is known except that they excite or inhibit the spinal motor neurons at various levels. How they harmonize with one another in their action or what their purpose in normal life may be is little more than conjecture at present.

(G.v.B./H.W.M./C.F.Sc./C.S.S./C.R.N./Ed.)

III. The peripheral nervous system

The peripheral nervous system is made up of three parts: (1) the cranial nerves, (2) the spinal nerves, and (3) that part of the autonomic (involuntary) nervous system that is outside the brain and spinal cord. The cranial nerves arise from, or are attached to, the brain and enter the skull via openings in the bone, called foramina. These nerves are concerned with the special senses and with the innervation of structures in the head and neck. The spinal nerves, attached segmentally to all regions of the spinal cord, enter and leave the vertebral canal by intervertebral foramina (openings between the vertebrae). These nerves innervate the cutaneous surfaces and muscles of most of the body and the deep structures in the trunk and limbs. The autonomic nervous system is a system of nerve

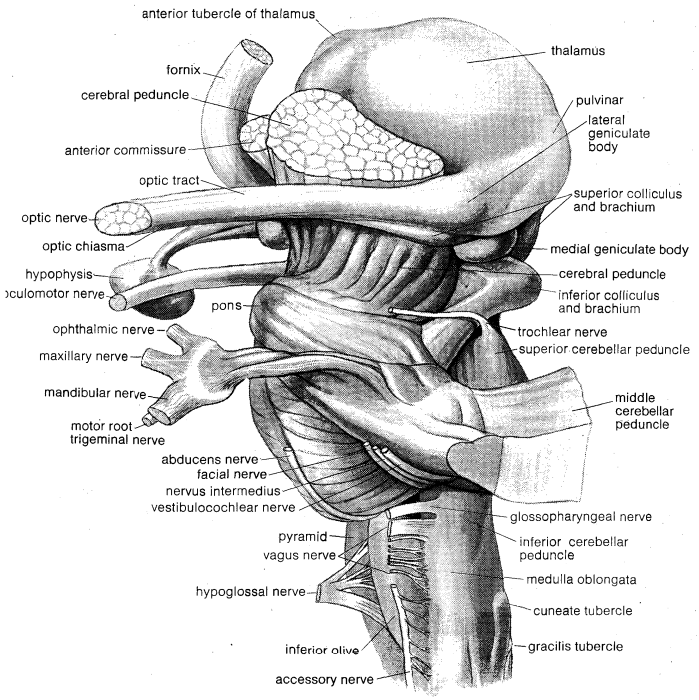


Figure 9: View of the lateral surface of the brainstem and attached cranial nerves.
From Fred A. Mettler, *Neuroanatomy*, 2nd ed. (1948); The C.V. Mosby Co., St. Louis

fibres and masses of nerve cells, called ganglia, arranged in trunks and plexuses; some fibres in this system pass peripherally with certain cranial and spinal nerves even though they subserve separate functions. (The autonomic nervous system is treated as a separate entity in another section of this article.)

The peripheral nervous system has the same essential structure as the central nervous system in that it consists of neurons surrounded by supporting cells (Figure 9). The most striking difference is that the cranial and spinal nerves are composed primarily of axons and dendrites invested by connective tissue sheaths. (Axons carry impulses from a nerve cell body; dendrites, impulses to it.) Cells of the peripheral nervous system lie in spinal, cranial, and autonomic ganglia. Virtually all afferent fibres, both somatic and visceral, have their cell bodies in spinal ganglia or ganglia associated with cranial nerves. Afferent fibres transmit impulses from peripheral receptors into the central nervous system. Autonomic ganglia contain neurons that distribute fibres peripherally that terminate in smooth muscle, heart muscle, and glandular epithelium. These cells give rise to efferent fibres; that is, fibres that convey impulses only to peripheral effector structures. Somatic efferent fibres emerging via spinal and cranial nerves arise from neurons within the spinal cord or brainstem. These fibres pass with peripheral nerve fibres directly to striated, or voluntary, muscles. In a strict sense, they actually are extensions of the central nervous system that link it with peripheral effectors. Autonomic ganglia also receive impulses from cells located within the central nervous system. These fibres, referred to as preganglionic, synapse on (*i.e.*, interconnect with) neurons within the autonomic ganglia. Thus, two neurons are involved in transmitting visceral efferent impulses to peripheral structures. They are: (1) a preganglionic fibre arising in the central nervous system, which synapses upon one or more autonomic ganglion cells, and (2) a postganglionic fibre, projecting peripherally, that arises from a cell in an autonomic ganglion.

Impulses, generated in receptors on body surfaces and within muscles, joints, tendons, and various viscera in response to external and internal stimuli, are conveyed by peripheral nerves to the brain and spinal cord. Receptors, in essence, constitute an elaborate set of transducers capable of converting physical and chemical stimuli (energy) into nerve impulses that are transmitted centrally. The central nervous system exerts its regulating and controlling influences upon striated muscle and viscera by impulses transmitted ultimately by peripheral nerves to effector structures. Virtually the entire input and output of the central nervous system is dependent upon the integrity of the peripheral nervous system. Conversely, the peripheral nervous system could not function as an isolated system.

CRANIAL NERVES

There are 12 pairs of cranial nerves that arise from, or are attached to, the brain. These symmetrically distributed nerves are concerned with the special senses and the innervation of muscles, glands, skin, and the mucous membranes in the head and neck. The vagus nerve is an exception in that it sends fibres to visceral structures in the thorax and abdomen (Figure 10).

Summary of cranial nerve functions

The olfactory and optic nerves, referred to as the first and second cranial nerves, are not actually peripheral nerves but tracts belonging to the central nervous system. The third, fourth, sixth, and 12th cranial nerves are exclusively motor and innervate the extraocular muscles and the muscles of the tongue. The fifth, seventh, ninth, tenth, and 11th cranial nerves innervate muscles derived from the branchial arches. The fifth cranial nerve, the largest of all the true cranial nerves, consists of a large

After Frank H. Netter, M.D., *The CIBA Collection of Medical Illustrations*; Copyright © 1957, 1972 by CIBA Pharmaceutical Company, Division of CIBA-GEIGY Corporation; reproduced with permission. All rights reserved

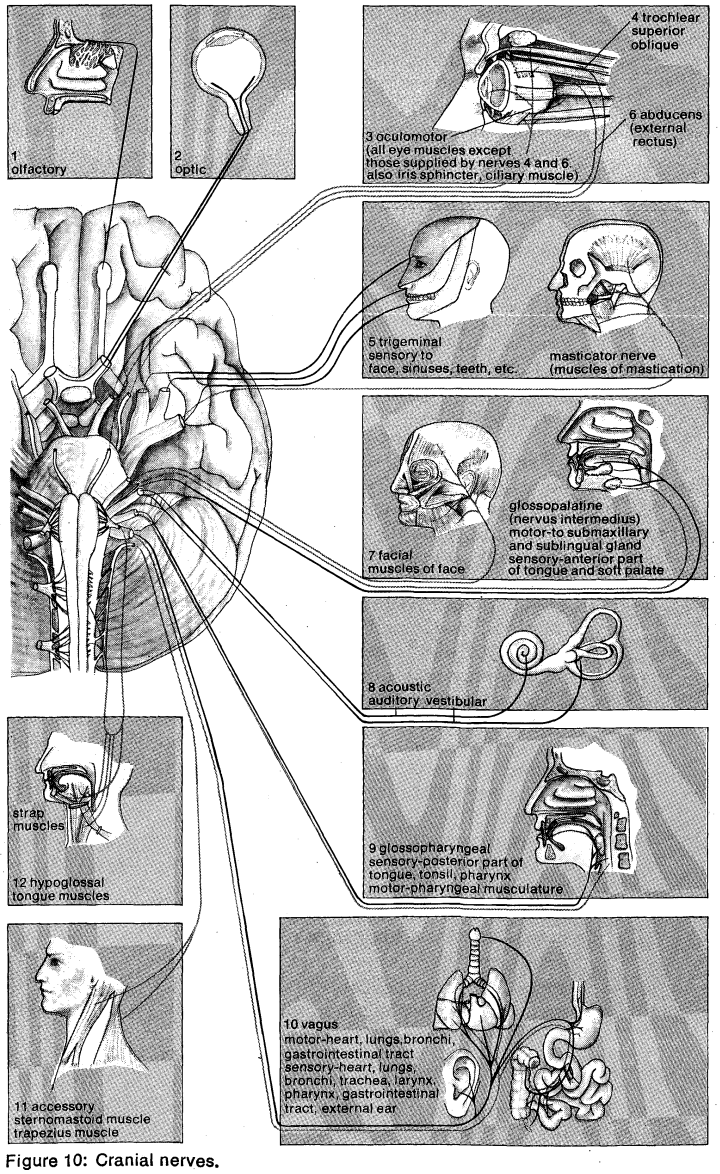


Figure 10: Cranial nerves.

sensory component supplying the face and scalp and a smaller motor component that innervates the muscles of mastication. The seventh, ninth, and tenth cranial nerves contain motor, sensory, and autonomic components. The 11th cranial nerve consists of two parts: (1) a cranial part functionally associated with the vagus nerve, and (2) a spinal part which has a distinctive origin and a separate distribution. The vestibulocochlear nerve (eighth) consists of two parts: the vestibular division concerned with equilibrium and the cochlear division concerned with audition. The first (olfactory), second (optic), eighth (vestibulocochlear), and parts of the seventh, ninth, and tenth (taste) cranial nerves convey impulses concerned with the special senses.

While four functional components of a typical spinal nerve are recognized (*i.e.*, general somatic afferent, general visceral afferent, general somatic efferent, and general visceral efferent), three additional special functional components are represented in the cranial nerves. These additional functional components are designated as: (1) special somatic afferent fibres, (2) special visceral afferent fibres, and (3) special visceral efferent fibres. The special somatic afferent cranial nerves are the optic and the vestibulocochlear nerves. The olfactory nerve and components of the facial, glossopharyngeal, and vagus nerves concerned with taste are classified as special visceral afferent. Motor components of the branchiomeric cranial nerves (*i.e.*, trigeminal, facial, glossopharyngeal, vagus, and accessory) are classified as special visceral efferent because the muscles they innervate are derived from the branchial arches. The oculomotor, the trochlear, the abducens, and the hypoglossal cranial nerves give rise to somatic efferent fibres that innervate muscles derived from somites. These nerves are classified as general somatic efferent cranial nerves. Portions of the oculomotor, facial, glossopharyngeal, and vagus nerves contain preganglionic parasympathetic fibres. These components are designated as general visceral efferent fibres.

Olfactory nerve. The first nerve arises from olfactory neurons scattered among supporting cells in a yellowish-brown patch of specialized epithelium in the upper part of the nasal cavity. Approximately 20 fine, olfactory filaments, not sheathed in myelin, arise from these receptor cells and pass from the nasal cavity into the anterior cranial fossa through foramina in the ethmoid bone. Olfactory filaments, representing the central processes of bipolar cells in the olfactory epithelium, collectively constitute the olfactory nerve. These filaments form a plexus before terminating in the olfactory bulb, a flattened, ovoid body resting on the cribriform plate of the ethmoid bone. Axons of cells in the olfactory bulb form the olfactory tract, which enters the brain.

Olfactory discrimination

The ability of the olfactory receptors, in concert with the brain, to discriminate thousands of different odours is well-known, but the physiological and psychological bases for these discriminations are unknown. Olfactory discrimination does not appear to be based upon morphologically distinct types of receptors. Spatial and temporal factors may play important roles in the neural coding of olfactory responses.

Optic nerve. The second cranial nerve consists of nerve fibres that arise from the ganglion cells of the retina. These fibres converge to form the optic disk, pierce the sclera, and combine to form the optic nerve (see also EYE AND VISION, HUMAN). Optic nerve fibres are axons of connector neurons which receive fibres from receptor neurons. Thus, the optic nerve is regarded as a central nervous system tract. Fibres of the optic nerve lack the neurolemmal sheath common to peripheral nerves, but the optic nerve is surrounded by the three meningeal membranes and a subarachnoid space, the space between the central and innermost layers of the membranes covering the brain and spinal cord. The optic nerves enter the optic canals and pass into the middle cranial fossa, where they form the optic chiasm, or chiasma. The optic chiasm, which lies at the base of the brain immediately in front of the tuber cinereum and infundibulum, is the site of a partial decussation (crossover) of optic nerve fibres. In the optic chiasm, fibres arising from the nasal halves of the retinas cross, while those from the temporal halves of the retinas do not. Thus, each optic tract, the uninterrupted continuation of optic nerve fibres, consists of uncrossed fibres from the temporal half of the ipsilateral retina (*i.e.*, the retina on the same side) and crossed fibres from the nasal half of the opposite retina. Fibres of the optic tract terminate upon cells of the lateral geniculate body, a relay nucleus of the thalamus.

Fibres of the optic nerve convey visual impulses from the retina, which are projected via relays to visual centres in the cerebral cortex. Light reflexes and reflexes resulting in lens accommodation also are mediated by the optic nerves and optic tracts.

Oculomotor nerve. The third cranial nerve is an efferent nerve that consists of two components: (1) general somatic efferent fibres, which innervate all the extraocular muscles except the lateral rectus (lateral straight muscle) and the superior oblique muscles of the eyeball, and (2) general visceral efferent fibres (preganglionic parasympathetic), which pass to the ciliary ganglion. The oculomotor nuclear complex is located ventral to the central gray matter of the midbrain. The complex consists of paired lateral somatic cell columns and an unpaired median cell group called the caudal central nucleus. The lateral somatic cell columns are composed of motor neurons, the fibres of which innervate the superior, medial, and inferior recti muscles and the inferior oblique muscle of the eyeball. The caudal central nucleus gives rise to crossed and uncrossed fibres that innervate the levator palpebrae, the muscle that raises the eyelid. The visceral nuclei of the oculomotor complex arise from small ovoid or spindle-shaped cells located rostral and dorsal to (*i.e.*, in front of and above) the somatic cell columns. These nuclei give rise to preganglionic parasympathetic fibres, which emerge with the somatic fibres and pass to the ciliary ganglion. The ciliary ganglion is a small reddish body, about the size of a pinhead, which lies near the back of the orbit. Preganglionic parasympathetic fibres synapse on cells of this ganglion. The short ciliary nerves arise from the ganglion and, as postganglionic parasympathetic fibres, convey impulses to the ciliary muscle and the sphincter of the iris.

The two components of the oculomotor nerve

Fibres of the oculomotor nerve exit from the brainstem and course forward to enter the lateral wall of the cavernous sinus. Fibres of this nerve enter the bony orbit of the eye via the superior orbital fissure. The nerve divides into superior and inferior branches. The superior branch of the nerve innervates the superior rectus and the levator palpebrae muscles. The inferior branch supplies the medial and inferior recti muscles and the inferior oblique muscle.

The extraocular muscles innervated by the oculomotor nerve serve to elevate the eyelid and move the eye up, down, and toward the midline. Parasympathetic impulses conveyed by the short ciliary nerves cause: (1) contraction of the ciliary body, which effects a thickening of the lens for accommodation, and (2) contractions of the sphincter of the iris, which constricts the pupil.

Trochlear nerve. The fourth nerve supplies motor fibres to the superior oblique muscle of the eyeball. This nerve arises from a small compact cell group in the midbrain. Root fibres of these cells curve dorsolaterally and caudally in the outer margin of the central gray matter and reach the superior medullary velum, where the nerves of the two sides cross (*i.e.*, decussate). The nerves emerge from the dorsal surface of the brainstem just below the inferior colliculus, a rounded eminence on the dorsal surface of the midbrain. This nerve traverses the lateral wall of the cavernous sinus and enters the orbit via the superior orbital fissure.

The superior oblique muscle, which is innervated by the trochlear nerve, turns the eye inward (*i.e.*, intorts) when abducted and depresses the eye when adducted. The trochlear nerve is the only totally crossed cranial nerve and the only one to issue from the dorsal surface of the brainstem.

Trigeminal nerve. The fifth nerve, the largest of the cranial nerves, contains both sensory and motor fibres.

Sensory and motor fibres in the fifth cranial nerve

The sensory division of this nerve conveys tactile sense, pain and thermal sense from the skin of the face and forehead, from the mucous membranes of the nose, nasal sinuses, and mouth, from the teeth, and from large parts of the cranial dura. The sense of deep pressure and movement (kinesthesis) is conveyed by branches of this nerve from the teeth, periodontium, hard palate, and receptors in the temporomandibular joint (articulation or "corner" of the jaw). Additional afferent fibres convey impulses from stretch receptors in the muscles of mastication.

Most, but not all, primary sensory neurons lie in the trigeminal ganglion situated in a dural cavity in the middle fossa of the skull. The ophthalmic, maxillary, and mandibular nerves arise from this ganglion.

Ophthalmic nerve. The ophthalmic nerve passes forward in the wall of the cavernous sinus and divides into three branches, which enter the orbit via the superior orbital fissure. These branches are the lacrimal, frontal, and nasociliary nerves. The lacrimal nerve supplies the lacrimal (tear) gland, the conjunctiva, and the skin of part of the upper eyelid. The frontal nerve passes through the orbital cavity and divides into two branches: the supraorbital nerve supplying the forehead and scalp and the supratrochlear nerve supplying part of the forehead and part of the adjacent upper eyelid. The nasociliary nerve enters the nasal cavity, where branches supply the mucous membrane, nasal septum, and certain areas of skin near the tip of the nose.

Maxillary nerve. The maxillary nerve leaves the middle fossa of the skull via the foramen rotundum and emerges on the face below the eye. The nerve gives off branches to the dura in the middle cranial fossa and to the mucous membranes of the soft and hard palate. The posterior and anterior superior alveolar nerves supply the teeth, gums, and the mucous membrane of the maxillary sinus. The infraorbital nerve, the terminal branch of the maxillary nerve, supplies the skin and conjunctiva of the lower eyelid, the skin on the side of the nose, the cheek, and the upper lip.

Mandibular nerve. The mandibular nerve contains sensory fibres from the trigeminal ganglion and a smaller number of motor fibres, which leave the middle fossa of the skull via the foramen ovale. The motor fibres supply the muscles of mastication. The sensory fibres in the mandibular nerve give rise to the auriculotemporal, lingual, and inferior alveolar nerves. The auriculotemporal nerve gives branches to the mandibular joint, the parotid gland, the upper half of the auricle, and the skin and fascia of the temple and scalp. The lingual nerve supplies the mucous membrane over the sides and dorsum of the anterior two-thirds of the tongue, the floor of the mouth, and the mucosa overlying the sublingual gland. The inferior alveolar nerve distributes branches to the lower teeth, including the incisors. The mental nerve, a cutaneous terminal branch of the inferior alveolar nerve, distributes fibres to the skin of the chin and the lower lip.

Sensory fibres concerned with pain, touch, thermal sense, and some related to pressure sense represent central processes of cells in the trigeminal ganglion. These fibres are distributed to central nuclei that extend throughout most of the lower brainstem. Sensory fibres that convey impulses from stretch receptors in the muscles of mastication, and pressure and kinesthetic sense from the teeth, periodontium, and joint capsules arise from cells within the central nervous system. These cells form the mesencephalic nucleus of the trigeminal nerve, an anomalous group of primary sensory neurons within the brainstem. The motor fibres of the trigeminal nerve arise from typical motor cells in the pontine tegmentum, located in the dorsal part of the pons, one of the basic subdivisions of the brainstem.

Abducens nerve. The sixth cranial nerve contains general somatic efferent fibres that innervate the lateral rectus muscle of the eyeball. The nucleus of the abducens nerve forms a column of typical somatic motor cells in the floor of the fourth ventricle. Root fibres of the abducens nerve emerge from the brainstem at the caudal border of the pons, just lateral to the medullary pyramid. The nerve enters the orbit via the superior orbital fissure and supplies the lateral rectus muscle. This is the cranial nerve most frequently injured. Injury results in horizontal diplopia (double vision) on attempted gaze to the side of the paralyzed muscle.

Facial nerve. The seventh cranial nerve is a mixed nerve containing efferent and afferent components. The largest part of this nerve is the motor component that innervates the muscles of facial expression, the superficial muscles of the scalp and neck, and certain deep muscles. The above component of the facial nerve contains special visceral efferent fibres. The smaller part of the nerve, known as the intermediate (glossopalatine) nerve, contains afferent taste fibres and efferent parasympathetic fibres that convey impulses to the lacrimal gland and certain salivary glands. Cell bodies of afferent fibres are located in the geniculate ganglion.

The motor component of the facial nerve

The motor component of the facial nerve arises from a distinct collection of motor neurons in the ventrolateral pontine tegmentum. Fibres of the nerve emerge from the brainstem at the cerebellopontine angle, formed at the junction of pons, medulla, and cerebellum. The nerve emerges from the skull at the stylomastoid foramen. The mastoid process covers this region in the adult but not in the child. After entering the region of the parotid gland, the nerve breaks up into numerous branches that supply the facial muscles. Temporal branches of the nerve supply the frontalis muscle, the muscles around the eye, and the superior auricular muscles. The lower zygomatic branches supply lower facial muscles including those of the nose and upper lip. Buccal branches supply the buccinator muscle in the cheek, which serves to push food onto the occlusal surfaces of the teeth. A marginal mandibular branch passes between the lower lip and the chin and innervates the muscles surrounding the mouth and those that depress the corner of the mouth. Cervical branches of the facial nerve sweep downward behind the angle of the jaw and supply the platysma muscle, a thin, superficial muscle that extends from the chest over the sides of the neck to the lower jaw.

Special visceral afferent fibres arise from neurons in the geniculate ganglion located at the external genu (bend) of the facial nerve. These fibres convey taste sensation from the anterior two-thirds of the tongue and form part of the intermediate nerve. General visceral efferent fibres (preganglionic parasympathetic) arising from cells of the superior salivatory nucleus in the pons, in part, pass to the lacrimal gland and to nasal and palatine glands. The continuation of the intermediate nerve distal to the geniculate ganglion is known as the chorda tympani nerve. This nerve, containing taste and parasympathetic fibres, separates from the facial nerve and becomes incorporated in the lingual nerve, a branch of the trigeminal nerve. Taste fibres remain intimately associated with the lingual nerve, while the preganglionic parasympathetic fibres synapse upon postganglionic neurons the fibres of which supply the submandibular (submaxillary) and sublingual salivary glands.

Chorda tympani nerve

Lesions of the facial nerve produce a paralysis of facial muscles known as Bell's palsy. Disturbances of taste, salivation, and lacrimation may accompany the facial weakness.

Vestibulocochlear nerve. The eighth nerve is a special somatic afferent cranial nerve consisting of two functionally distinct parts. The vestibular nerve transmits impulses from the labyrinth, a specialized end organ concerned with static and kinetic equilibrium. The cochlear nerve transmits impulses from the cochlear apparatus concerned with audition. Although these nerves run together from the internal auditory meatus to the cerebellopontine angle, where they enter the brainstem, they have distinctive peripheral receptors and central connections (see also EAR AND HEARING, HUMAN).

Vestibular nerve. The vestibular part of the inner ear consists of the three semicircular canals, the utricle, and the saccule. These receptors are concerned with equilibrium and orientation in three-dimensional space. The semicircular canals are arranged at right angles to one another, and one end of each canal has a dilated portion,

the ampulla, containing a transversely oriented ridge. These ridges (cristae ampullaris) are covered by specialized neuroepithelial hair cells, which constitute the vestibular receptor. The utricle and saccule each have a similar patch of sensory epithelium: the macula utriculi and macula sacculi. Hair cells in the maculae are in contact with a gelatinous covering containing small calcareous particles, known as the otoliths.

The cristae of the semicircular canals are stimulated by angular acceleration, which causes movement of the endolymphatic fluid and deflection of hairs in the sensory epithelium. The utricular macula, concerned with static equilibrium, responds to changes in gravitational forces and to linear acceleration. The functional role of the saccule in equilibrium has not been resolved.

The maculae and cristae of the labyrinth are innervated by cells of the vestibular ganglion (ganglion of Scarpa), an aggregation of bipolar cells located in the internal auditory meatus. Peripheral processes of bipolar cells pass to the hair cells of the maculae and cristae. The central processes of these cells form the vestibular nerve the fibres of which pass between the inferior cerebellar peduncle and the spinal trigeminal tract and are distributed to vestibular nuclei in the brainstem.

Cochlear (auditory) nerve. The cochlea is a spiral bony tube with two and three-quarter coils arranged around a central pillar called the modiolus. The coils are referred to as basal, middle, and apical. The cochlea is partitioned by two membranes, the vestibular membrane and the basilar membrane, creating three divisions in all coils. These divisions are the scala vestibuli, the cochlear duct, and the scala tympani. The organ of Corti, the auditory end organ, rests on the basilar membrane in the cochlear duct. This receptor consists of three outer rows of hair cells and one inner row of hair cells, surrounded by supporting cells. A delicate membrane, the tectorial membrane, overhangs the inner half of the spiral organ. Sound waves induce movements of the tympani membrane, which are transmitted via the ear ossicles to the scala vestibuli via the oval window. The endolymph and basilar membrane respond to these waves, which alter relationships between the hair cells of the organ of Corti and the tectorial membrane. Sound waves of different frequencies cause hair cells in different regions of the cochlea to respond. Tones of low frequency cause cells in the apical part of the cochlea to respond, while those of high frequency selectively stimulate cells in the base of the cochlea.

Fibres of the cochlear nerve arise from bipolar neurons situated around the modiolus of the cochlea. The short peripheral processes of these cells end upon the hair cells of the organ of Corti. The central processes of these cells form the cochlear nerve, which enters the brainstem at the cerebellopontine angle.

Glossopharyngeal nerve. The ninth cranial nerve is a mixed nerve, the large afferent component of which supplies the pharynx, part of the tongue, and the carotid sinus and body. A smaller efferent component innervates the stylopharyngeus and middle pharyngeal constrictor muscles and sends preganglionic parasympathetic fibres to the otic ganglion.

Sensory components of the ninth cranial nerve

Sensory components of this nerve arise from two peripheral ganglia: the superior and the inferior. The superior ganglion is small, lies in the jugular foramen, and gives rise to general somatic afferent fibres that supply small cutaneous areas back of the ear; these fibres are carried in the auricular branch of the vagus nerve. The larger inferior ganglion of the glossopharyngeal nerve gives rise to both general and special (taste) visceral afferent fibres. General visceral afferent fibres convey tactile sense, pain, and thermal sense from the mucous membranes of the pharynx, the posterior third of the tongue, the tonsil, and the eustachian tube. A special sensory branch, the carotid sinus nerve, conveys afferent impulses from the carotid sinus and body. The carotid sinus is a dilatation of the common carotid artery and the root of the internal carotid branch. Elevation of carotid arterial pressure initiates afferent impulses in carotid sinus receptors, which are conveyed centrally to vagal nuclei and produce a slowing of the heart rate and a fall in blood pressure. The carotid body, near the branching of the common carotid artery, is a small neurovascular structure that is sensitive to changes in the hydrogen ion concentration of the blood.

Lingual branches of the glossopharyngeal nerve convey taste and general sensation from the posterior third of the tongue.

Preganglionic parasympathetic fibres in the ninth nerve arise from the inferior salivatory nucleus in the medulla and pass via the lesser petrosal nerve and the tympanic branch to the otic ganglion. Postganglionic parasympathetic fibres arising in the otic ganglion pass to the parotid gland and provide its secretomotor and vasomotor supply.

The glossopharyngeal nerve is attached to the lateral surface of the medulla by five or six rootlets. These rootlets lie caudal to those of the facial nerve and are in series with the rootlets of the vagus nerve.

Vagus nerve. The tenth cranial nerve is a mixed nerve containing a large number of parasympathetic fibres that are widely distributed (vagus from the Latin "wandering"). This cranial nerve is unique in that a large part of its fibres are distributed within the thorax and abdomen. The nerve supplies afferent and efferent fibres to the pharynx, larynx, esophagus, stomach, gut, trachea, and lungs. Special fibres of the nerve pass to and from the heart and great vessels. Numerous rootlets of the nerve are attached to the lateral surface of the medulla in series with those of the glossopharyngeal nerve above and the accessory nerve below. The cranial portion of the accessory (11th) nerve contributes efferent fibres to the vagus nerve, most of which pass to the muscles of the larynx. Rootlets of the vagus nerve unite to form a single trunk, which enters the neck via the jugular foramen.

Two ganglia associated with the vagus nerve give rise to afferent fibres. The superior ganglion of the vagus nerve is a small spherical swelling in the root that lies in the jugular foramen. Cells in this ganglion give off general somatic afferent fibres that pass to meningeal and auricular branches. The meningeal branches supply the dura mater of the posterior fossa of the skull. The auricular branch distributes fibres to the skin back of the ear, to the floor of the external auditory meatus, and to the lower part of the tympanic membrane. The inferior ganglion of the vagus nerve, large and spindle-shaped, lies immediately below the jugular foramen. Cells in the inferior ganglion give rise to general visceral afferent fibres and special visceral afferent fibres. The latter convey impulses from taste buds in the region of the epiglottis.

Dorsal motor nucleus of the vagus

Efferent fibres of the vagus nerve arise from cell columns within the medulla. Preganglionic parasympathetic fibres arise from cells of the dorsal motor nucleus of the vagus nerve. This cell column, composed of spindle-shaped cells, lies in the floor of the fourth ventricle dorsolateral to the hypoglossal nucleus. Fibres from the cell column join other fibres of the vagus nerve. Special visceral efferent fibres of the vagus nerve arise from the nucleus ambiguus, a column of typical multipolar motor neurons embedded in the reticular formation. These fibres join other fibres of the vagus nerve and emerge from the lateral surface of the medulla dorsal to the inferior olivary nuclear complex. Fibres arising from the nucleus ambiguus innervate the muscles of the pharynx and larynx. Those that innervate the larynx emerge from the caudal part of the nucleus ambiguus and form the cranial part of the accessory nerve; after joining the spinal part of the accessory nerve, these fibres return to form an integral part of the vagus nerve.

Two branches of the vagus nerve are the pharyngeal and the superior laryngeal nerves. The pharyngeal nerve contains efferent fibres that, along with branches of the glossopharyngeal nerve and sympathetic nerve fibres, form the pharyngeal plexus. Fibres from this plexus supply the muscles of the pharynx and the soft palate. The superior laryngeal nerve divides into a larger internal and a smaller external laryngeal branch. The internal laryngeal nerve supplies the mucous membrane of the larynx, the epiglottis, and the most posterior part of the tongue. The

external laryngeal nerve supplies motor fibres to the inferior constrictor muscle of the pharynx and the cricothyroid muscle.

In the neck the vagus nerve follows a vertical downward course. It enters the thorax on the right side after crossing the subclavian artery; on the left side it crosses between the left common carotid and subclavian arteries. In the thorax the courses of the nerves differ on the two sides. The right vagus nerve continues along the side of the trachea to the root of the lung. The left vagus nerve passes across the aortic arch and then proceeds to the root of the left lung. The vagus nerves form two plexuses: the pulmonary and the esophageal. The pulmonary plexuses are formed behind the root of each lung. From the lower part of the pulmonary plexuses two nerves emerge that reach the esophagus and form the esophageal plexus.

Recurrent laryngeal nerves

The recurrent laryngeal nerves, which supply all of the intrinsic muscles of the larynx except the cricothyroid, pursue different courses on the right and left sides. The right recurrent laryngeal nerve arises in the root of the neck as the vagus nerve crosses the first part of the subclavian artery. This nerve hooks around the artery and passes upward and medially in a groove between the esophagus and the trachea. In this location the nerve lies along the medial side of the right lobe of the thyroid gland. The left recurrent laryngeal nerve separates from the vagus nerve as it crosses the arch of the aorta and, after hooking under the arch, passes upward into the neck in a fashion similar to that of the nerve on the right side.

In the neck both vagus nerves supply cardiac branches, designated as upper and lower cardiac nerves. On the right side both cardiac branches pass downward behind the subclavian artery into the thorax to reach the deep cardiac plexus. On the left side the nerves separate in the thorax; the upper nerve passes near the trachea to join the deep cardiac plexus. The lower nerve passes across the aortic arch to terminate in the superficial cardiac plexus.

In the thorax the vagus nerves form the cardiac, pulmonary, and esophageal plexuses. The cardiac plexus is formed by sympathetic and vagal cardiac branches that meet to produce a large plexus that lies around the great vessels at the base of the heart. The whole plexus is divided into superficial and deep plexuses that distribute fibres to the heart and to the pulmonary plexuses.

The pulmonary plexuses, formed behind the roots of the lung, give off numerous branches which extend into the lungs and are distributed to the bronchi, the pulmonary vessels, and the pleura. The esophagus is supplied by vagal nerve fibres that emerge from the posterior pulmonary plexuses to form the esophageal plexus. Branches of this plexus supply the muscle and mucous membranes in the wall of the esophagus.

From the lower part of the esophageal plexus, anterior and posterior vagal trunks arise and pass into the abdomen. The posterior vagal trunk distributes fibres to the posterior surface of the stomach as well as to the coeliac, splenic, and renal plexuses. The anterior vagal trunk distributes fibres to the lesser curvature of the stomach and branches to the pylorus and the first part of the duodenum. Thus, vagal fibres are distributed directly to the stomach and indirectly to the great autonomic plexuses in the abdomen. Fibres of the vagus nerve extend in the gut as far as the splenic flexure of the colon. These fibres innervate the muscular wall of the gut and provide secretomotor fibres to the glands of the stomach and intestine.

The vagus nerve is the most complex and the most important cranial nerve. Unilateral lesions of the vagus nerve produce ipsilateral paralysis of the soft palate, pharynx, and larynx, which results in hoarseness of the voice and difficulty both in breathing and in swallowing fluids. Vagal cardiac fibres serve to slow the heart rate and probably constrict the coronary vessels, while pulmonary vagal influences constrict the bronchi. Vagal fibres distributed to abdominal viscera stimulate peristalsis and gastrointestinal secretion.

Accessory nerve. The 11th cranial nerve consists of two distinct parts with separate origins and distributions. The cranial part arises from the caudal part of the nucleus ambiguus by several rootlets and emerges below the vagus nerve from the lateral surface of the medulla. These fibres innervate the intrinsic muscles of the larynx and perhaps some muscles of the pharynx. The spinal part of the accessory nerve originates from a cell column in the anterior horn of the spinal cord extending from the fifth (or sixth) cervical segment to the lower part of the medulla. Root fibres from these cells unite to form a common trunk. This common trunk ascends in the spinal canal, enters the skull through the foramen magnum, and joins the cranial part of the accessory nerve. Both parts of the nerve emerge from the skull via the jugular foramen. The cranial part becomes incorporated in the vagus nerve, while the spinal part passes into the neck, where it supplies the sternocleidomastoid and the upper part of the trapezius muscles. The sternocleidomastoid muscle turns the head; the trapezius muscle braces the shoulder and rotates the scapula during elevation of the upper limb.

The two parts of the 11th cranial nerve

Hypoglossal nerve. The 12th cranial nerve is a motor nerve that innervates the muscles of the tongue. The nerve arises from a long column of multipolar motor neurons that occupies the central gray matter of the medulla, which forms the floor of the fourth ventricle. The rootlets of the nerve pierce the dura mater separately and unite to form the hypoglossal nerve in the hypoglossal canal, the bony opening through which the nerve emerges from the skull. At the base of the skull, the nerve is closely associated with the glossopharyngeal, vagus, and accessory nerves. The nerve runs forward and passes into the floor of the mouth, where its fibres are distributed to the hypoglossus, styloglossus, geniohyoid, and genioglossus muscles, and all the intrinsic muscles of the tongue.

The intrinsic muscles of the tongue alter the shape of the tongue, while the extrinsic muscles alter its shape and position. The genioglossus muscle protrudes the tongue. Injury to the hypoglossal nerve results in paralysis of one half of the tongue with loss of muscle tone and substance (atrophy). When the tongue is protruded, it deviates to the side of the injury.

SPINAL NERVES

The spinal nerves are arranged in pairs, with each pair symmetrically attached to the sides of its respective spinal cord segment. Paired spinal nerves are symmetrical in their course and peripheral distribution. There are 31 pairs of spinal nerves, which for descriptive purposes are separated into eight pairs of cervical nerves, 12 pairs of thoracic nerves, five pairs of lumbar nerves, five pairs of sacral nerves, and one pair of coccygeal nerves. The paired coccygeal nerves are rudimentary and occasionally absent. Spinal nerves emerge from the vertebral canal via the intervertebral foramina. The first cervical nerve emerges between the occiput and the atlas, while the eighth cervical nerve exits between the seventh cervical and the first thoracic vertebrae. All thoracic, lumbar, sacral, and coccygeal nerves emerge below the vertebrae of their corresponding number. Spinal nerves vary in size; the largest nerves are those which form the great nerve trunks innervating the lower extremities (Figure 11).

Attachments of spinal nerves to the spinal cord

Origin of spinal nerves. Each spinal nerve is attached to the spinal cord by two nerve roots: (1) a dorsal sensory root (afferent) and (2) a ventral motor root (efferent). The dorsal roots are larger than the ventral roots, contain more fibres, and have fibres of larger diameter. Each dorsal root has interposed in its course an ovoid mass of nerve cells that form the spinal (dorsal root) ganglion. Nerve fibres forming the dorsal root arise from these cells and thus are of peripheral origin. Dorsal root ganglion cells give rise to a single fibre that undergoes a T-shaped bifurcation. One branch of this bifurcation passes peripherally, while the central branch enters the spinal cord. These cells are referred to as unipolar neurons. Central branches of the dorsal root ganglia at first form a single compact bundle, which passes through a sleeve of dura mater and then breaks up into a series of root filaments. These filaments spread vertically in a fanlike manner to enter the spinal cord. Peripheral fibres

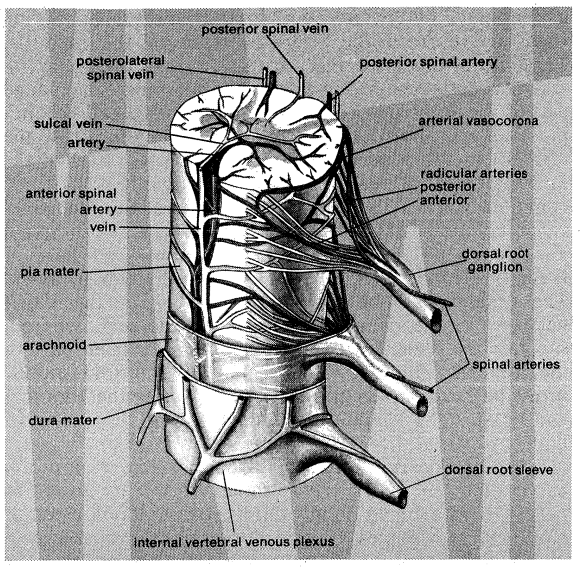


Figure 11: Portion of the spinal cord showing the nerve roots, the meninges, and the blood supply.
From R.C. Truex and M.B. Carpenter, *Human Neuroanatomy* (© 1969); The Williams and Wilkins Co., Baltimore

from the ganglia pass laterally and join the ventral root in the intervertebral foramina.

Fibres composing the ventral root are of central nervous system origin. These fibres arise from collections of large motor cells in the anterior horn of the spinal gray matter. The ventral root has no ganglion. Fibres of the ventral root emerge from the anterior and anterolateral surfaces of the spinal cord as a series of fine nerve fibres. These fibres are irregular in their arrangement and sometimes overlap. Certain spinal ventral roots contain efferent autonomic nerve fibres. Spindle-shaped cells in the lateral horns of all thoracic (chest) segments and in the first two or three lumbar (mid-lower back) spinal segments give rise to preganglionic sympathetic fibres that emerge via the corresponding ventral roots. The cells of origin of these fibres are known as the intermediolateral nucleus. Small scattered neurons located along the lateral surface of the anterior gray horn in the second, third, and fourth sacral (low-back) segments give rise to preganglionic parasympathetic fibres that exit via the ventral roots and pass to autonomic ganglia in or near the walls of pelvic viscera. Cells of these sacral autonomic nuclei appear similar to those of the intermediolateral nucleus.

The mixed spinal nerve

Dorsal and ventral roots pass laterally from their attachments to the spinal cord toward the intervertebral foramen, where they unite to form a mixed spinal nerve, or common nerve trunk. In the mixed nerve, motor and sensory fibres are blended. The first two cervical nerves pass directly laterally and somewhat upward. The roots of other spinal nerves course obliquely downward and laterally. This oblique downward course increases in caudal regions; lower lumbar, sacral, and coccygeal nerve roots pursue a vertical course downward in the vertebral canal. This progressively increasing obliquity of spinal roots is a consequence of the more rapid rate of growth of the vertebral column. During early development the spinal cord extends to the lower sacrum, but after the fourth month of fetal life the vertebral column elongates more rapidly than the spinal cord. At birth the conus medullaris (*i.e.*, terminal part of the spinal cord) lies opposite the third lumbar vertebra, and in the adult it reaches the lower border of the first lumbar vertebra. Because the spinal cord is attached to the brainstem above, there is a progressive lengthening of the spinal roots from their attachments to the spinal cord to their entrance into the intervertebral foramina. This lengthening of spinal roots is much greater in the lumbar and sacral regions than in the cervical region. The collection of spinal nerve roots below the first lumbar vertebra that surrounds the filum terminale is the cauda equina.

As spinal root filaments leave the spinal cord they become invested with pia mater, which becomes continuous with the arachnoid where the roots penetrate the dura. The two roots pierce the dura separately, and a dural investment is added. Distal to the spinal ganglion the meningeal sheaths are blended and become continuous with the connective tissue covering of the spinal nerve.

Functional components of spinal nerves. Four functional components are recognized in a typical spinal nerve. These components are: (1) general somatic afferent fibres, (2) general visceral afferent fibres, (3) general somatic efferent fibres, and (4) general visceral efferent fibres. General somatic and general visceral afferent fibres arise from cells in the spinal ganglia. General somatic afferent fibres convey impulses from skin receptors concerned with pain, thermal sense, pressure and touch, and from receptors in muscles, joints, tendons, and subcutaneous tissues. General visceral afferent fibres convey impulses from visceral structures, glands, and blood vessels. General somatic efferent fibres arise from neurons within the central nervous system and convey motor impulses to striated muscles that initiate and maintain contractions of these muscles. General visceral efferent fibres arise from cell groups in restricted regions of the spinal cord and project to autonomic ganglia which, in turn, give rise to postganglionic fibres. Sympathetic visceral efferent fibres arise from thoracic and upper lumbar spinal segments and constitute the sympathetic outflow for the entire body. Sacral segments two, three, and four provide the spinal parasympathetic outflow that passes to lower abdominal and pelvic viscera via parasympathetic ganglia.

Typical spinal nerve. As a spinal nerve emerges from an intervertebral foramen, it divides into two main branches—a dorsal and a ventral ramus. The dorsal ramus, usually smaller than the ventral ramus, divides into medial and lateral branches that innervate the skin, fascia, ligaments, and axial muscles of the back. The larger ventral ramus courses laterally and ventrally in the body wall as an intercostal (between the ribs) nerve. A lateral cutaneous branch innervates the skin on the lateral and ventral surface of the body. The principal ventral ramus continues ventrally, giving off branches that innervate intercostal muscles; the terminal part of this nerve gives rise to anterior cutaneous branches.

The dorsal and ventral rami

The ventral rami of the first thoracic to the third lumbar spinal nerves give rise to white rami communicantes that pass to sympathetic ganglia. These are preganglionic (efferent) fibres that arise from cells within the spinal cord and pass to sympathetic ganglia. The sympathetic ganglia and trunk give rise to gray rami communicantes that enter the ventral rami; these postganglionic fibres are distributed to blood vessels, glands, and piloerector muscles.

Distribution of spinal nerves. The dorsal rami of the spinal nerves innervate the skin, fascia, and muscles over dorsal parts of the head and trunk. The muscular branches of these rami are limited to the longitudinal postvertebral muscles associated with the axial skeleton. Cutaneous branches of these rami supply areas on the back of the head, the dorsum of the trunk, and in the shoulder and gluteal (buttock) regions. None of these branches supplies cutaneous areas or muscles in the extremities. The first cervical spinal nerve has a small or frequently absent dorsal root; muscular branches, mainly in the dorsal ramus, supply posterior neck muscles attached to the skull and cervical vertebrae. The second cervical nerve gives rise to small muscular branches and the greater occipital nerve, which innervates cutaneous areas over the posterior part of the scalp. The dorsal rami of the lower cervical nerves are small, and those of the seventh and eighth cervical nerves have no cutaneous branches.

The dorsal ramus of each thoracic spinal nerve divides into medial and lateral branches. In upper thoracic nerves the medial branches are cutaneous, while the lateral branches are distributed to the muscles. The reverse is found in the lower six thoracic nerves where the medial branches innervate the larger longitudinal muscle masses of the back. The dorsal rami of the first three lumbar nerves are distributed in the same fashion as the lower

thoracic nerves. Lower lumbar nerves supply only muscular branches to the longitudinal muscles of the back. The first three sacral dorsal rami give rise to medial muscle branches and lateral cutaneous branches that supply skin and fascia over the sacral and gluteal regions. The fourth and fifth sacral dorsal rami distribute only cutaneous branches to the coccygeal region.

The ventral rami of spinal nerves consist of afferent and efferent fibres and are distributed to lateral and anterior parts of the body, including the extremities. Only in the thoracic region are the ventral rami distributed in a regular manner. Other ventral rami combine to form plexuses. In these plexuses components are rearranged and distributed in a more complex manner. The principal plexuses are the cervical, brachial, lumbar, sacral, and coccygeal. Ventral rami of the cervical nerves and parts of the upper two thoracic nerves are distributed to the head, neck, and upper extremity by the cervical and brachial plexuses.

Principal plexuses of spinal nerves

Cervical plexus. The cervical plexus is formed by the ventral rami of the first four cervical nerves. From the loops of this plexus, branches are distributed to the cutaneous regions of the head, neck, and shoulder; muscles of the neck and diaphragm; and certain cranial and sympathetic nerves. Six cutaneous nerves of the plexus radiate from the posteriad border of the sternocleidomastoid muscle. Among these are the following ascending branches: the lesser occipital nerve, which innervates the upper part of the neck and the scalp behind and above the ear; the great auricular nerve, which innervates the lower part of the auricle, the scalp immediately behind the ear, and skin overlying the parotid gland and the angle of the mandible; and the transverse nerve of the neck, which supplies the skin over the anterior triangle of the neck. The supraclavicular branches radiate laterally from the neck to be distributed to skin and fascia over the inferior part of the neck, the front of the chest, and the shoulder. Lateral muscular branches of the cervical plexus supply parts of the sternocleidomastoid, the trapezius, and the levator scapulae muscles (muscles of the neck and shoulder). Medial muscular branches of the cervical plexus supply prevertebral muscles, the geniohyoid and infrahyoid muscles, and the diaphragm. The phrenic nerve, which arises from the fourth cervical nerve and to a lesser extent from the third and fifth, courses downward in the neck, passes between the subclavian artery and vein, and enters the thorax. This nerve traverses the middle mediastinum, lying between the pericardium and the pleura, and reaches the diaphragm where it separates into numerous branches.

Brachial plexus. The brachial plexus is formed from the ventral rami of the four lower cervical and the first thoracic nerves with a small contribution from the fourth cervical nerve. Considerable variations occur. If the contribution from the fourth cervical is appreciable and that from the first thoracic negligible, the plexus is called prefixed. It is called postfixed if the fourth cervical does not contribute to the plexus and the first thoracic makes a strong contribution. The ventral rami forming the plexus give rise to three primary trunks: (1) fifth and sixth cervical unite to form the upper trunk, (2) seventh cervical forms the middle trunk, and (3) eighth cervical and first thoracic unite to form the inferior trunk. Each trunk divides into an anterior and a posterior division. The three posterior divisions unite to form the posterior cord. The anterior division of the upper trunk becomes the lateral cord, while the anterior division of the lower trunk becomes the medial cord.

Prefixed and postfixed plexuses

Nerves supplying the shoulder muscles arise directly from the ventral rami or from the trunks. The dorsal scapular nerve, innervating the rhomboid muscles, arises from the dorsal surface of the fifth cervical. The long thoracic nerve, supplying the anterior serratus muscle, arises from the dorsal surfaces of the fifth, sixth, and seventh cervical. The suprascapular nerve to the supraspinatus and infraspinatus muscles arises from the upper trunk. The medial and lateral pectoral nerves, which innervate the pectoralis major and minor muscles, arise from the ventral surfaces of the upper and middle trunks and, in part, from the medial cord. The thoracodorsal and subscapular nerves arise from the posterior cord and supply, respectively, the latissimus dorsi muscle and the teres major and subscapular muscles—shoulder muscles which control movements of the upper arm.

The three largest peripheral nerves in the upper extremity (*i.e.*, radial, median, and ulnar) are formed in the following manner. The posterior cord divides into its two terminal branches: the larger radial nerve and the smaller axillary nerve. The medial and lateral cords each split into two branches. The two middle branches, one from the lateral cord and one from the medial cord, unite to form the median nerve. The outer branch of the lateral cord becomes the musculocutaneous nerve. The inner branch of the medial cord gives off two purely sensory nerves, the medial brachial cutaneous nerve and the medial antebrachial cutaneous nerve, and is continued peripherally as the ulnar nerve.

Axillary, radial, median, and ulnar innervation

The axillary nerve innervates the deltoid and teres minor muscles. The radial nerve supplies the triceps muscle and extensor muscles of the elbow, hand, and digits. Its cutaneous branches supply the posterior aspect of the arm, forearm, and part of the dorsum of the hand. The median nerve supplies most of the flexor muscles of the forearm and certain muscles in the hand. The sensory innervation of the median nerve is limited to the hand, where it includes the volar surface of the thumb, index, and middle fingers and the radial half of the fourth finger and corresponding parts of the palm. The ulnar nerve supplies flexor muscles in the forearm and certain flexor and adductor muscles in the hand. Cutaneous areas innervated by this nerve include the dorsal and volar surface of the little finger and adjacent half of the ring finger. A mild blow to the medial aspect of the elbow ("funny bone") may cause a sensation of numbness and tingling in the cutaneous distribution of the ulnar nerve.

Lumbar plexus. The lumbar plexus is formed by the ventral rami of the first four lumbar nerves, often with an additional branch from the 12th thoracic. The fourth lumbar nerve contributes to both the lumbar and sacral plexuses. Nerves arising from this plexus supply the loin region as well as part of the lower limb. The arrangement of the lumbar plexus is simpler than that of the brachial plexus in that the ventral rami do not form interlacing trunks but split into posterior and anterior divisions related primarily to primitive posterior and anterior muscle groups in the leg. Peripheral nerves to the lower extremity, lower abdomen, and genital region are formed by the union of variable numbers of either posterior or anterior divisions. Anterior divisions give rise to the following nerves: (1) iliohypogastric (anterior branch), (2) ilioinguinal, (3) genitofemoral, and (4) obturator. The posterior divisions give rise to the following nerves: (1) iliohypogastric (posterior branch), (2) femoral, and (3) lateral femoral cutaneous.

The iliohypogastric, ilioinguinal, and genitofemoral nerves supply muscular and cutaneous branches to the abdominal wall, the inguinal region, and the external genitalia. The obturator nerve supplies muscles on the medial aspect of the thigh, gives articular branches to the hip and knee joints, and sometimes distributes a cutaneous branch to the inner surface of the thigh. The large femoral nerve supplies the muscles on the anterior aspect of the thigh (*i.e.*, the extensor muscles of the thigh), gives articular branches to the hip and knee joints, and has an extensive cutaneous distribution on the medial aspect of the lower extremity to the foot. Branches of this nerve innervate the pectineus, sartorius, and quadriceps muscles. Cutaneous branches of the nerve are the anterior cutaneous nerve, distributed to anterior and anteromedial surfaces of the thigh, and the saphenous nerve, which is distributed to skin and fascia of the medial aspect of the leg and foot. The lateral femoral cutaneous nerve innervates skin and fascia over the lateral part of the front of the thigh as far down as the knee; a smaller posterior branch supplies the lateral aspect of the buttock and skin over the lateral side of the thigh.

Innervation of the legs and feet

Sacral plexus. The sacral plexus is formed by the ventral rami of the fifth lumbar and the upper three sacral

nerves with contributions from the fourth lumbar and the fourth sacral nerves. The sciatic nerve, the largest and thickest nerve in the body, is the chief continuation of all roots of the sacral plexus. This large nerve leaves the pelvis and enters the gluteal region through the greater sciatic foramen. The nerve enters the thigh at the lower border of the gluteus maximus and descends under the long head of the biceps femoris muscle. This nerve usually terminates in the popliteal fossa (*i.e.*, back of the knee) by dividing into the tibial and common peroneal nerves. The division of the nerve may occur at any higher level, as these two nerves frequently are distinct at their origin. The tibial and common peroneal nerves are bound together proximally by an investing sheath.

In addition to the sciatic nerve the sacral plexus gives rise to numerous smaller nerves. These collateral branches innervate the main external rotators of the thigh, the obturator internus, the gemelli, and the quadratus femoris muscles. The pudendal nerve arises from the second, third, and fourth sacral nerves and provides branches distributed to the anal sphincter, the scrotum (or labium majus), and the penis (or clitoris). Other branches from proximal parts of the sacral plexus are nerves to the gluteal muscles and the posterior femoral cutaneous nerve. Parasympathetic fibres from the second, third, and fourth sacral nerves join autonomic plexuses in the pelvis and are distributed to urogenital organs and the lower part of the large intestine. Nerves given off from the sciatic nerve near the lower part of the buttock innervate the hamstring muscles, the principal flexors of the knee. The common peroneal and tibial nerves are the two main trunks resulting from the division of the sciatic nerve.

The peroneal nerves

The common peroneal nerve arises from the posterior cords of the sacral plexus and is homologous with the radial nerve of the upper limb. Two cutaneous nerves arise from the common peroneal nerve in the popliteal fossa. These are the lateral cutaneous nerve on the calf, distributed over the lateral and posterior aspect of the leg, and the peroneal communicating branch, which gives branches to the lateral surface of the leg. The two terminal branches of the common peroneal nerve are the deep and superficial peroneal nerves. The division of the common peroneal nerve takes place immediately below the head of the fibula. Branches of the deep peroneal nerve innervate the tibialis anterior, extensor hallucis longus, and extensor digitorum longus muscles, which serve to dorsiflex (extend) the foot and toes. In the foot the terminal branches are: (1) the medial digital branch, supplying the skin on adjacent sides of the great and second toes; and (2) a lateral muscular branch which innervates the extensor digitorum brevis and gives rise to articular branches for tarsal, tarsometatarsal, and metatarsophalangeal joints. The superficial peroneal nerve passes downward in front of the fibula and innervates the peroneus longus and brevis muscles. Terminal branches of this nerve are cutaneous. Medial branches supply skin and fascia over the dorsum of the foot, the medial aspect of the great toe, and adjacent parts of the second and third toes. Lateral branches of the superficial peroneal nerve supply distal parts of the leg, part of the dorsum of the foot, and regions between the third and fourth, and fourth and fifth toes. Cutaneous nerves on the dorsum of the foot are subject to considerable variation, and nerves supplying the digits are smaller than plantar digital nerves.

The tibial nerve arises from the anterior divisions of the sacral plexus. This nerve is homologous with the median and ulnar nerve trunks in the upper extremity. The nerve lies deep in the posterior thigh, passes into the popliteal fossa, and distally is covered by the gastrocnemicus and soleus muscles. Branches of the nerve arising in the popliteal fossa innervate the gastrocnemicus, soleus, and posterior tibial muscles. The sural nerve, a cutaneous branch, pierces the deep fascia on the back of the leg and descends to reach the lateral surface of the foot. This nerve and its branches supply cutaneous regions on the posterior and lateral aspects of the distal third of the leg, the ankle, the heel, and the lateral border of the foot. Muscular branches of the tibial nerve in the leg innervate the tibialis posterior, the soleus, and the long flexor muscles of the foot. The terminal branches of the tibial nerve are distributed to the plantar surface of the foot by the medial and lateral plantar nerves. The medial plantar nerve is homologous with the median nerve in the hand, while the lateral plantar nerve is homologous with the ulnar nerve. The medial plantar nerve innervates the abductor hallucis and the flexor digitorum brevis muscles and gives rise to four common plantar digital nerves. The lateral plantar nerve gives branches to small muscles of the foot and cutaneous branches to the fifth toe and part of the fourth toe.

Coccygeal plexus. The coccygeal plexus is formed by the ventral ramus of the fourth sacral nerve with those of the fifth sacral and coccygeal nerves. These nerves form a plexiform cord homologous with the inferior caudal trunk of tailed animals. Some of these nerve fibres pierce the coccygeus and overlying ligaments to supply the skin and fascia near the coccyx and anus. (M.B.C.)

IV. The autonomic nervous system

By classical definition, the autonomic nervous system, sometimes called the visceral or vegetative nervous system, is that system of motor nerve fibres that supplies (carries impulses to) the cardiac (heart) muscle, smooth muscle, such as that in the blood vessels, and glands, and that consists, in its simplest form, of nerve pathways each consisting of two succeeding nerve cells. The first cell is located in the brain or spinal cord and the second in a ganglion, or collection of nerve cells, outside of the brain and spinal cord. Anatomically and functionally, however, the autonomic nervous system is much broader than the definition given above would indicate. The motor fibres comprising this system are closely associated with sensory fibres in all of their functions. An integral part of the entire nervous system, they are activated by nerve impulses arising by a reflex action, at local levels, or originating in the higher centres of the brain.

Because a modern treatment of the autonomic system can be presented only within the context of the entire nervous system, some overlap with the other sections of this article is inevitable. Such overlap, however, focusses on autonomic functions.

For the sake of clarity and to facilitate an understanding of certain functions, the autonomic nervous system will be presented as if it were a series of levels of the central and peripheral nervous systems that differ in function in that the higher the level, the more widespread and general its functions, the lower the level, the more restricted and specific its functions. The lowest level consists of the two-nerve-cell pathway mentioned above. Before the various levels are taken up in detail, a general account of the principles of organization and function is given.

GENERAL ORGANIZATION AND FUNCTION

The two-nerve-cell pathway begins with certain nerve cells located in the brainstem and spinal cord (Figure 12). The axons of these cells—that is, the fibres carrying impulses from the cells—are called preganglionic fibres (they are also known as general visceral efferent fibres). With some exceptions, taken up in later sections, the preganglionic fibres leave the brain and spinal cord and transmit impulses to nerve cells in peripheral ganglia. (A circumscribed collection of nerve cells outside of the central nervous system is usually called a ganglion; a collection of nerve cells within the central nervous system is usually called a nucleus—not to be confused with the nucleus of a single cell. If a nucleus gives rise to a cranial nerve or a spinal root, it is often referred to as a nucleus of origin.) The axons of the ganglion cells are commonly called postganglionic fibres. They terminate in cardiac (heart) muscle, smooth muscle, and certain gland cells, all of which, because they can produce actions, are often known as effectors. The most obvious effectors are skeletal muscles, which are supplied by motor (somatic efferent) fibres; these fibres go directly, without interruption, from motor cells in the brain or spinal cord to the muscle fibres.

The locations, arrangements, and connections of the

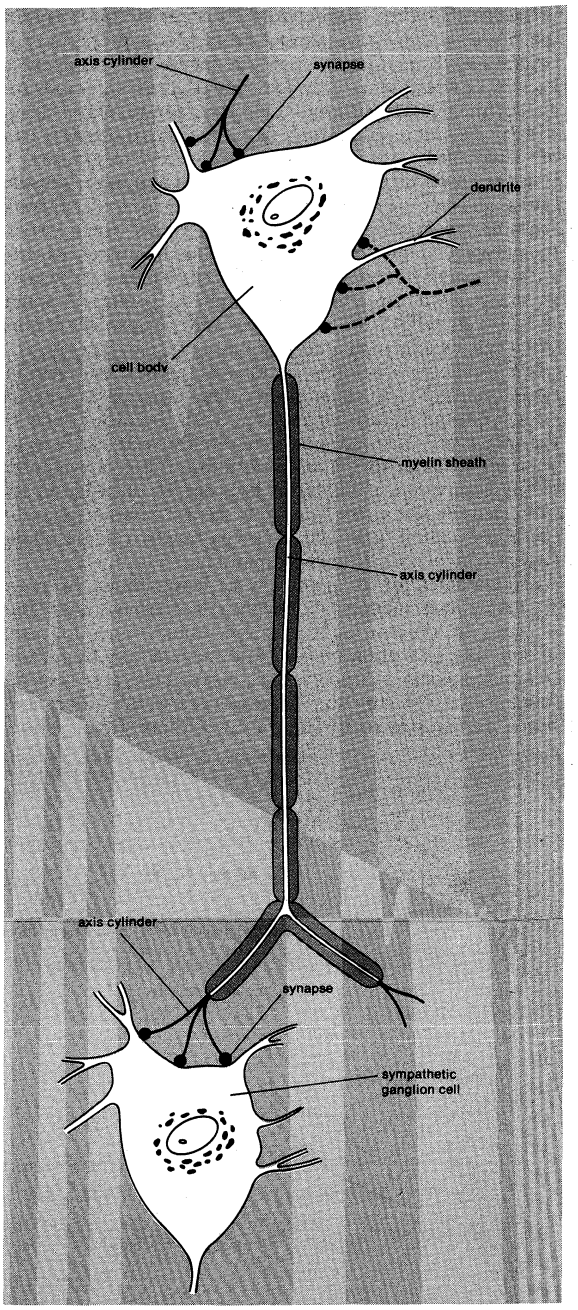


Figure 12: General features of a multipolar sympathetic nerve cell of the spinal cord (as shown by the light microscope), and of one of the many synapses between the axis cylinders axons) of this cell and nerve cells in sympathetic ganglia.

two-nerve-cell pathways of the autonomic nervous system form the basis of an anatomical classification or subdivision of the pathways into sympathetic and parasympathetic systems, or divisions. Before these two divisions are described in greater detail, it is necessary to give a brief account of the effectors they supply, namely, cardiac muscle, smooth muscle, and glands. The activation of an effector by nerve impulses takes place across a specialized nerve–effector relationship known as a neuro-effector junction or, especially in the case of skeletal muscle, a motor nerve ending.

Effectors supplied. *Cardiac muscle.* Cardiac muscle fibres comprise the bulk of the heart. The fibres are multinucleated cells that contain longitudinally disposed minute fibres—myofibrils—that have alternating light and dark segments. This structural arrangement causes the fibres to appear cross-striated when examined under the microscope. Though skeletal muscle fibres are also cross-striated, they differ from cardiac fibres in many important respects. The significant differences are in nerve supply, in the fact that cardiac fibres have an intrinsic rhythmic beat or contraction that begins in the embryo, that cardiac fibres can survive without a nerve supply, that cardiac fibres cannot be voluntarily controlled as most skeletal muscles can, and that some of the cardiac fibres are specialized for conduction rather than contraction.

The specialized fibres form the conducting system of the heart (see CARDIOVASCULAR SYSTEM, HUMAN), consisting of two definitive collections of specialized fibres called nodes (the sino-atrial and the atrioventricular node), together with bundles of fibres spreading out from the nodes to the regular cardiac fibres. Postganglionic autonomic fibres, both sympathetic and parasympathetic, form motor endings on the fibres of the nodes. From there, impulses are conducted to the ordinary cardiac muscle fibres. The two nodes, by virtue of rhythmic activity that begins in the embryo, together set the level of cardiac activity, or rate of heartbeat. The autonomic nerve supply to the heart acts by stimulating or inhibiting the ongoing nodal rhythm, thereby increasing or decreasing ongoing activity of the heart.

Smooth muscle. Many organs contain what is known as smooth muscle. The unit of a layer of smooth muscle is a small spindle-shaped cell; this contains a nucleus and myofibrils that lack alternating light and dark sections and hence are said to be smooth (uniform) rather than cross-striated. Smooth muscle cells are usually arranged in sheets or bundles, within the walls of viscera, around many glands, and in the walls of blood vessels. Contractions of smooth muscle are slower than those of striated muscle and often seem to spread from one fibre to another as a wavelike activity of the sheet or bundle. This spread may be facilitated by the low resistance contacts, or fusions, known as tight junctions, which occur between many of the fibres.

Neuro-effector junctions in smooth muscle seem to be between varicosities—bulging areas—of nerve fibres and individual smooth muscle cells. It is unlikely that every smooth muscle fibre is so supplied, and it is likely that organs differ in the density of the innervation of their smooth muscle. Also, the degree to which a single postganglionic nerve fibre may divide and supply more than one smooth muscle fibre is not known with certainty. Like cardiac muscle, smooth muscle may survive without a nerve supply and in some instances possesses intrinsic, or locally initiated, activity, the level of which may be modified by the autonomic nervous system.

Glands. Glands are composed of epithelial (covering or lining) cells adapted for the formation of materials other than those directly concerned in their intrinsic metabolism. The glands secrete such substances as hormones, enzymes, and mucus. They may be composed of one or of many cells. If the glands discharge their products onto a surface or into the channel of an organ, they are called exocrine glands, or glands of external secretion. If they discharge their contents into the blood, they are called endocrine glands, or glands of internal secretion.

Exocrine and endocrine glands

Generally, only the multicellular exocrine glands are supplied by the autonomic nervous system. The neuro-effector junctions formed by postganglionic autonomic fibres are basically similar to those of smooth muscle. Like smooth muscle, gland cells are also usually relatively unaffected by loss of nerve supply. They may continue to function normally, either by virtue of intrinsic activity or through activation by circulating hormones and other substances.

An important principle emerges from the above, namely, that the autonomic nervous system is not indispensable to the structures it supplies and that it often functions by regulating intrinsic activity. Cutting the nerve supply (denervation) is not without effect, however. Some of the consequences are discussed later (see below *Pathophysiology*).

Viscera and visceral activity. The organs containing cardiac fibres, smooth muscle, or gland cells are sometimes called viscera (hence the term visceral nervous system). The term viscera is a general one, however, and is difficult to define precisely. It is often used to refer to

Meaning of term viscera

organs and tissues that are supposedly not under voluntary control (for example, the heart), in contrast to skeletal, or striated, muscle. The distinction is more apparent than real. Many skeletal muscles, such as the diaphragm, which separates the chest and abdominal cavities, are not subject to complete voluntary control. Organs containing smooth muscle may be as much under voluntary influence or control as many so-called voluntary muscles. Moreover, certain organs, such as the pharynx, or throat, and esophagus, or gullet, contain striated muscle but are classified as viscera.

In its narrow sense, then, the term viscera refers to the heart and to the organs and tissues containing smooth muscle and glands and to a large degree under the control or regulation of the autonomic nervous system. In the broader sense, the term visceral activity is more useful. For example, skin, which is usually not classified as a viscus, contains many glands, as well as smooth muscle attached to hair follicles. Blood vessels, which often contain smooth muscle fibres, are found throughout the body. However ill defined it may be or how unaware of it one may be, visceral activity is a complex pattern involving sensory and motor nerve fibres and many parts of the central nervous system. The regulation of body temperature is an excellent example of such a complex pattern of activity, a fundamental component of which is the autonomic nervous system.

Cold-blooded vertebrates, such as fish, amphibians, and reptiles, have body temperatures that vary with the surrounding temperature. Warm-blooded animals, such as birds and mammals, keep their body temperatures constant within certain limits and are therefore to a certain extent independent of their thermal environment. Heat production and heat loss are relatively balanced. The heat that is produced by metabolic activity is lost chiefly through radiation from the blood vessels that lie near the surface of the body. A certain amount of heat is also lost with the evaporation of sweat, with the exhalation of air and water vapour from the lungs, and with the excreta. If more blood flows through the surface vessels per unit of time, and especially if more vessels are made available for blood to flow through, more heat will be lost by radiation. The autonomic nervous system can regulate the calibre of the blood vessels by, for example, activating the smooth muscle of the vessels with a resultant narrowing of the diameter of the vessels. It can also control special vessels that act as shunts—that is, when open, allow blood to bypass the surface vessels. If these vascular shunts are closed, blood is diverted to the surface vessels. This overall control of blood flow may be initiated in a number of ways. Skin exposed to the cold air, for example, becomes blanched or pale as the result of a reflex narrowing, or constriction, of the vessels in the skin, coupled with an opening of the vascular shunts. This reflex action is brought about as follows: the cold air stimulates nerve receptors that are sensitive to temperatures lower than that of the normal skin surface. The nerve impulses thus initiated reach the spinal cord or brainstem, where they activate motor cells of the autonomic nervous system. These in turn send impulses along fibres to peripheral ganglia, whose postganglionic fibres supply and activate the smooth muscle of the surface vessels, thereby constricting them and narrowing their lumen, or channel. At the same time, some fibres activate the smooth muscle of the shunts so that these vessels open. This reflex mechanism results in a decrease in heat radiation, thereby helping to maintain the internal body temperature, even though the skin may be cold, both subjectively and objectively.

In general terms, then, the autonomic nervous system is an important part of the mechanisms whereby the body keeps its internal environment (*milieu interieur*) constant. That is, it helps to maintain the temperature, fluid balance, and ionic composition of blood. The maintenance of the internal environment is called homeostasis.

In carrying out these general functions, the autonomic nervous system is concerned with a number of specific functions, such as those related to digestion, intermediate metabolism of foods, and excretion. The system is concerned with many specific regulatory functions, involving, for example, body temperature, blood pressure, and respiration. It enables the body to handle emergencies or sudden environmental changes and is thus an important part of the mechanisms by which a person reacts to stress. Finally, the autonomic nervous system is an integral part of behavioral and emotional actions and responses. It must be emphasized that, in all of these activities, sensory fibres and sensory mechanisms are often indispensable features. Most viscera have sensory fibres arising from them, some concerned with pain but many concerned with reflex mechanisms. These fibres, which are termed general visceral afferent fibres, reach the brain and spinal cord by travelling in autonomic and peripheral nerves, often side by side with autonomic fibres.

Specific functions of the autonomic nervous system

The foregoing brief description of temperature regulation illustrates how effectors may be activated as a result of nerve impulses arising reflexly, with the motor side of the reflex including autonomic nerve fibres. It has already been pointed out that these autonomic fibres characteristically comprise a pathway of two nerve cells, one in the brainstem or spinal cord and the second in a peripherally located ganglion, and that these pathways can be classified as sympathetic or parasympathetic.

The classification of pathways as sympathetic or parasympathetic is inadequate because it does not take into account the afferent (sensory) and reflex connections, the higher levels of organization, and certain functional and pharmacological characteristics. Nevertheless, it does indicate some fundamental characteristics and differences. Some are illustrated in Figure 13. Others are described

From E. Gardner, *Fundamentals of Neurology*, 5th edition (1968); W.B. Saunders Company

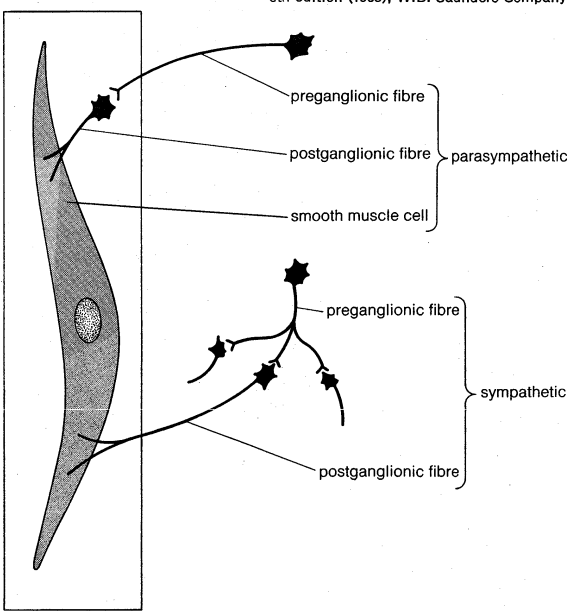


Figure 13: Differences in sympathetic and parasympathetic distribution. Sympathetic fibres synapse with many cells in ganglia of the sympathetic trunk or prevertebral plexuses, usually at some distance from the organ supplied.

later in this account (see also Table 3). Most organs are supplied by both systems, although, as indicated in Table 3, there are major exceptions. Nearly all peripheral blood vessels, for example, are supplied by the sympathetic system alone. When an organ has a double autonomic nerve supply, the two systems tend to have opposite functions. Thus, the parasympathetic system slows the heart rate, and the sympathetic increases it.

Main divisions. *Sympathetic system.* This is also called the orthosympathetic or thoracolumbar system. The nerve cells that give rise to the preganglionic fibres form a narrow column that extends throughout each side of the thoracic (chest) and upper lumbar (small of the back) parts of the spinal cord (hence the term thoracolumbar). The axons leave by way of the ventral (forward) roots of the corresponding spinal nerves. It is char-

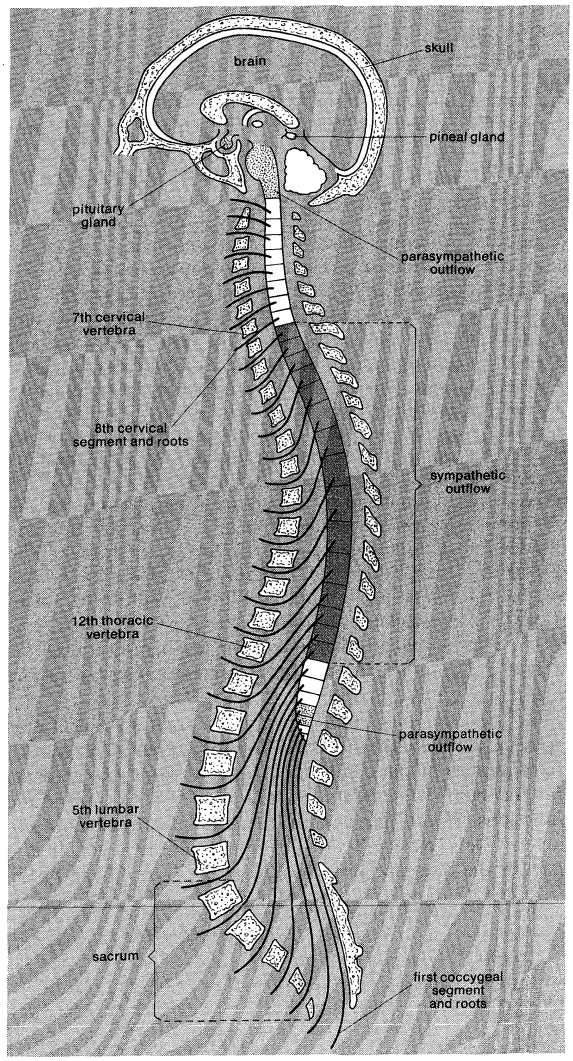


Figure 14: Diagrammatic representation of brain, spinal cord, spinal nerves, and vertebrae and the levels of autonomic outflow.
From E. Gardner, D.J. Gray, and R. O'Rahilly, *Anatomy*, 3rd edition (1969); W.B. Saunders Company

The preganglionic fibres

acteristic of preganglionic fibres, at least in higher vertebrates, that they are small in diameter and myelinated (sheathed), whereas postganglionic fibres are still smaller in diameter and are nonmyelinated. The preganglionic fibres then reach nerve cells in nearby autonomic ganglia by way of communicating branches called rami communicantes. Most of the ganglia are contained in the sympathetic trunks, which lie one on each side of the vertebral column, extending along its whole length, from the base of the skull to the coccyx. A number of ganglia are found in the meshes of the great prevertebral plexuses, in the chest, abdomen, and pelvis, and still others comprise accessory ganglia scattered along certain spinal roots and nerves. Further details of these various arrangements are given later in this article in the sections on sympathetic and parasympathetic outflow.

Some preganglionic fibres synapse (interconnect) with nerve cells in the ganglia of the sympathetic trunks, whereas others continue through to nerve cells in prevertebral ganglia. Still others continue on to the medullae (inner substance) of the adrenal, or suprarenal, glands, where they synapse with gland cells of the medullae. Each preganglionic sympathetic fibre usually synapses with many nerve cells or medullary cells.

The postganglionic fibres that arise in the various ganglia either go directly to adjacent viscera and blood vessels or return to spinal nerves by way of rami communicantes. They usually run relatively long distances to reach the organs or tissues that they supply. Those that return to the spinal nerves are distributed by way of the peripheral nerves and supply (1) the skin with secretory fibres to sweat glands, motor fibres to smooth muscle fibres attached to hairs, (2) vasomotor fibres to the blood vessels of the limbs and trunks, and (3) motor fibres to smooth muscle and gland cells of the head and neck.

Parasympathetic system. This is also called the craniosacral system (Figure 13). The motor cells whose axons form the preganglionic fibres are found in the nuclei of origin of certain cranial nerves (listed below) and also in the sacral part of the spinal cord (hence the term craniosacral).

Cranial parasympathetic outflow

The cranial outflow is from the oculomotor, facial, glossopharyngeal, vagus, and accessory nerves (third, seventh, ninth, tenth, and 11th cranial nerves). The myelinated preganglionic fibres leave the brainstem in the corresponding cranial nerves. Those in the 11th nerve soon enter the tenth nerve and are conducted by it to parasympathetic ganglia. The fibres in the oculomotor, facial, and glossopharyngeal nerves synapse with cells in certain ganglia in the head and neck called cranial or cephalic parasympathetic ganglia. Those in the vagus nerves synapse with ganglia in the walls of the thoracic and abdominal viscera. In most instances, the postganglionic fibres are nonmyelinated; the fibres are usually rather short, and the ganglia are in or near the organ to be innervated. Moreover, each preganglionic fibre synapses with relatively few nerve cells, or just one, and the postganglionic fibres supply fewer structures than do postganglionic sympathetic fibres.

The chief parasympathetic ganglia associated with the cranial nerves are the ciliary (related to the oculomotor nerves and concerned with constriction of the pupil of the eye), the pterygopalatine and submandibular (related to the facial nerves), and the otic (related to the glossopharyngeal nerves and concerned with the secretion of the parotid salivary gland).

The motor cells comprising the preganglionic fibres issuing from the sacral part of the spinal cord form a short column of cells on each side of the middle segments of the sacral cord. The fibres leave by way of the ventral roots of the corresponding spinal nerves and reach sympathetic ganglion cells in the walls of the pelvic organs and some of the organs of the lower abdomen.

Further details of the various arrangements of the craniosacral outflow are given in the sections dealing with sympathetic and parasympathetic outflow.

LEVELS OF STRUCTURAL ORGANIZATION

As indicated before, the autonomic nervous system can be considered as if it were controlled or regulated by a series of physiological and anatomical levels that differ in function in that the higher the level, the more widespread

Table 1: Parasympathetic Ganglia Associated with Cranial Nerves

ganglion	location	parasympathetic root	sympathetic root	chief distribution
Ciliary	lateral to optic nerve	from third cranial nerve (oculomotor)	internal carotid plexus (from superior cervical ganglion)	eye and eyelid (ciliary muscle; sphincter pupillae; dilator pupillae; tarsal muscles)
Pterygopalatine	in pterygopalatine fossa	from seventh cranial nerve (facial)	internal carotid plexus (from superior cervical ganglion)	lacrimal gland and glands in oral, nasal, and palatine mucosa
Otic	below foramen ovale	from ninth cranial nerve (glossopharyngeal)	plexus on middle meningeal artery (from superior cervical ganglion)	parotid gland
Submandibular	on hyoglossus muscle	from seventh cranial nerve (facial)	plexus on facial artery (from superior cervical ganglion)	submandibular and sublingual glands

Table 2: Derivatives of the Primary Vesicles Showing Origin of the Parts of the Autonomic Nervous System

primary vesicle	subdivision	derivatives	lumen
Prosencephalon (forebrain)	telencephalon (endbrain)	cerebral cortex; basal ganglia and limbic system	lateral ventricle and part of third ventricle
	diencephalon (interbrain)	thalamus and hypothalamus; neurohypophysis; pineal body	most of third ventricle
Mesencephalon* (midbrain)	mesencephalon (midbrain)	collicular region; cerebral peduncles	cerebral aqueduct
Rhombencephalon* (hindbrain)	metencephalon (afterbrain)	pons; cerebellum	fourth ventricle
	myelencephalon (marrowbrain)	medulla oblongata	fourth ventricle and part of central canal
Remainder of* neural tube		spinal cord	most of central canal

*Cells migrating from these parts of the neural tube, together with cells migrating from the neural crest, give rise to autonomic ganglion cells, the adrenal medullae, and chromaffin cells.

and general its functions; the lower the level, the more restricted and specific its functions. The lowest levels comprise the autonomic nervous system of classical definition.

In consideration of these levels, the general arrangement of the nervous system is notable. Certain features need to be emphasized in order to pave the way for consideration of their autonomic functions. During the course of development in the embryo, the nervous system becomes a tubular structure. This tubular arrangement is retained in the adult, though in a considerably modified form. The part of the tube in the head region develops into the brain and its various subdivisions (see Table 2), whereas the part of the neural tube that lies in the vertebral canal becomes the spinal cord. The brain and spinal cord constitute the central nervous system. The cranial and spinal nerves, their peripheral combinations, and the peripheral parts of the autonomic nervous system comprise the peripheral nervous system.

The parts of the brain

The brain, from below upward, consists of the hindbrain, the midbrain, and the forebrain, all contained within the cranial cavity. Generally speaking the Latin term *cerebrum* means "brain"; it is also used to mean specifically the forebrain and midbrain. The bulk of the brain is formed by two convoluted (or folded over) cerebral hemispheres and, lying between them, the diencephalon, or posterior part of the forebrain. The cerebral cortex, which is the thin, outer part of the hemispheres, is composed of gray matter (chiefly nerve cells, fibres, and supporting cells), in contrast to the interior of the brain, which is composed partly of white matter (consisting largely of the fibres projecting from nerve cells). The interior of the cerebral hemispheres contains not only white matter but well-demarcated masses of gray matter known collectively as basal ganglia (one of the exceptions to the general rule that collections of nerve cells in the brain and spinal cord are known as nuclei).

The diencephalon, which is composed mostly of gray matter, lies on each side of and below the third ventricle (one of several fluid-filled cavities derived from the lumen, or channel, of the embryonic neural tube; see Table 2). It forms the uppermost part of what is generally called the brainstem. This is an unpaired stalk that descends from the base of the brain and that is formed, from above downward, by the diencephalon, midbrain, pons, and medulla oblongata.

The pons and medulla oblongata, together with the cerebellum, comprise the hindbrain. The cerebellum is a large, fissured structure situated on the back of the brainstem, to which it is attached by three pairs of peduncles, or stalks. Its cortex, like that of the cerebral hemispheres, is composed of gray matter.

In their emergence from, or attachment to, the brain, the first two cranial nerves are associated with the forebrain. The third and fourth are associated with the midbrain, and the fifth to the 12th with the hindbrain.

The spinal cord

The spinal cord is a long, almost cylindrical mass of nervous tissue that occupies about the upper two-thirds of the vertebral canal. In contrast to the cerebral hemispheres, gray matter is found in the interior, somewhat in the form of the letter H, with a central canal in the bar of the H. The spinal cord has essentially the same structure throughout. Attached to the spinal cord on each side is a series of nerves, the spinal roots, named dorsal and ventral according to their position: dorsal root fibres are sensory, and ventral root fibres are motor (Figure 15). Generally, there are 31 pairs of roots on each side, which, from above downward, comprise eight cervical, 12 thoracic, five lumbar, five sacral, and one coccygeal. The part of the spinal cord to which a pair of dorsal roots and a pair of ventral roots are attached is termed a segment of the cord. Each dorsal root presents an ovoid swelling, the spinal ganglion, consisting of sensory cells conveying impulses from the periphery to the spinal cord. Distal to the ganglion (*i.e.*, beyond the ganglion in the direction away from the cord), each dorsal root combines with the corresponding ventral root to form a spinal nerve. This nerve is also connected to the adjacent sympathetic trunk by rami communicantes (Figure 15). Each spinal nerve divides into a dorsal and a ventral ramus, or branch. The dorsal rami supply the skin and muscles of the back. The ventral rami supply the limbs and the rest of the trunk. In the cervical (neck) and lumbosacral (lower back) regions, the ventral rami intermingle to form plexuses, or networks, from which peripheral nerves emerge to supply the limbs.

The various levels that control or regulate autonomic activity begin with the cerebral hemispheres, in which certain parts of the cerebral cortex and some of the basal ganglia control, regulate, or are involved in autonomic functions. These parts of the hemispheres send fibres to lower levels, especially to the hypothalamus, which is a part of the diencephalon located at the base of the brain. The hypothalamus has vascular and nervous connections

From E. Gardner, *Fundamentals of Neurology*, 5th edition (1968); W.B. Saunders Company

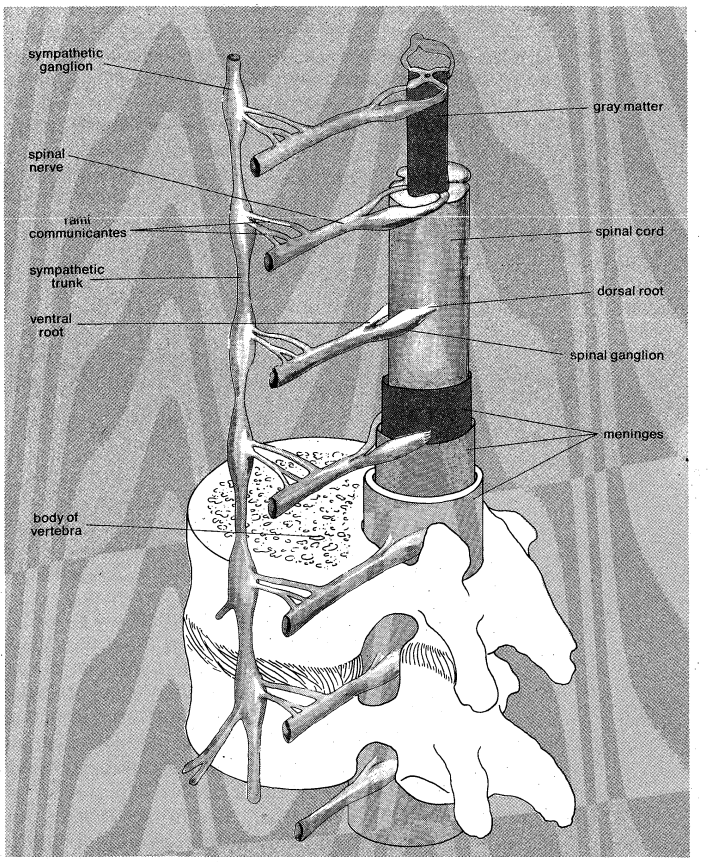


Figure 15: Diagram of the spinal cord, vertebrae, and sympathetic trunk (shown on one side only). Dorsal rami of the spinal nerves are not shown.

with the pituitary gland and also sends nerve fibres to still lower centres in the brainstem. These centres in turn send nerve fibres to the parasympathetic and sympathetic cells in the brainstem and spinal cord. The axons of these cells leave the central nervous system as preganglionic fibres to synapse at the lowest level, namely, in the sympathetic and parasympathetic ganglia. The postganglionic fibres from the cells in these ganglia are distributed to the appropriate effectors.

The purpose of the following paragraphs is to outline briefly the structural characteristics of the various levels before proceeding to a more detailed discussion of functions.

Cerebral hemispheres. As has been mentioned previously, the cerebral cortex, which is the outer part of the cerebral hemispheres, consists chiefly of nerve cells, their fibres, and supporting cells. Extensive studies have led to the conclusion that the nerve cells of the cerebral cortex are arranged in a similar pattern in all animals that have a cortex. With the exception of certain parts of the cortex that developed earlier in the evolution of the race, the nerve cells are grouped into six layers. The earliest studies of the cerebral cortex led to the naming of lobes, or subdivisions; convolutions, or gyri (ridges); fissures, or clefts; and sulci, or furrows. Subsequently, studies of cellular lamination indicated regional differences in cell density, type, and other features in the various layers. This resulted in a system of numbering (also lettering) different cortical areas, several hundred of which have been described. Some experimental and clinical studies tend to indicate specific functions for each area. While it is true that certain functional differences do exist (for example, motor areas, sensory areas, and speech areas), nevertheless, the anatomical distinctions are often not clear-cut, the connections between nerve cells are enormously complicated and not at all well understood, functional specificity is often lacking, and species differences in function are hardly understood at all.

The layers and areas of the cortex

The highest mental and behavioral activities are centred in the cerebral hemispheres, in particular the cerebral cortex. Also, the cerebral cortex controls or influences many aspects of autonomic activity, by way of connections to lower centres. There is no specific cortical region or area, however, that has as its sole function the control of autonomic activity.

The parts of the cerebral hemispheres that figure most prominently in autonomic functions include much of the cortex of the frontal lobes, certain of the basal ganglia, in particular those called the amygdaloid bodies (see below), and what is generally called the limbic lobe or limbic system.

There is as yet no general agreement as to the precise nature or extent of the so-called limbic lobe or system, sometimes referred to as limbic areas. What does seem clear is that primitive vertebrates have simple cerebral hemispheres with a single-layered cortex and that this phylogenetically old cortex—*i.e.*, this cortex that developed early in the evolutionary process—is present in higher vertebrates, including man. Together with its connections and certain phylogenetically newer structures, it comprises the limbic system (the amygdaloid bodies, each of which consists of a number of different nuclei, are often included with it). This system is found at the base of the brain, where it forms part of the frontal and temporal lobes of the hemispheres, and it borders much of the lateral ventricle (*limbus*, from the Latin, meaning "border" or "edge"). In lower vertebrates, the primitive cortex is concerned chiefly with olfaction, the sense of smell. In higher vertebrates, the structures comprising the system are still associated with this sense but in addition are involved in many other activities, including autonomic functions and certain aspects of emotion and behaviour. Indeed, the limbic system has been called the visceral brain. The term system implies a unity of structure and function that does not exist, however, and it seems likely that a more rational nomenclature will arise as connections and functions continue to be clarified.

Functions of the limbic system

The limbic system receives nerve fibres from the olfactory tracts and their connections, from the hypothalamus, and from many parts of the neocortex (the phylogenetically newer, six-layered cerebral cortex). In turn, the limbic system, especially the primitive cortex and the amygdaloid bodies, sends nerve fibres to lower centres, in particular to the hypothalamus. It is the powerful connections between the limbic system and the hypothalamus that are chiefly responsible for making possible the autonomic functions of the cerebral hemispheres.

Diencephalon. *Hypothalamus.* The hypothalamus, which is part of the diencephalon, is located at the base of the brain, where it forms the lower part of the wall of the third ventricle. Although it comprises but a small part of the brain, the hypothalamus is structurally complicated, being made up of a number of different nuclei. Some of these nuclei, particularly two that are named the supraoptic and the paraventricular, are notable for their intimate relationship to the minute blood vessels—capillaries—that form a dense network around them, for their neurosecretory functions, and for the fact that they send projections to the neurohypophysis (see below). Still other hypothalamic nuclei have projections that are closely related to the blood vessels that supply the pituitary gland (see below). These relationships between the hypothalamus and the pituitary gland comprise the mechanism by which the nervous system, especially the portions concerned with autonomic functions, can influence the production of hormones.

In spite of the fact that most of the hypothalamic nuclei have been described and named, there is not yet complete agreement on their organization, and there is strikingly little specific information about the course and destination of many of the nerve fibres leaving the nuclei. As mentioned above, some fibres go to the pituitary, and many go to the limbic system. A powerful projection is sent to the thalamus, and many fibres descend to various parts of the brainstem.

Connections with the hypothalamus

With regard to incoming fibres, nerve impulses aroused by the sense of smell reach the hypothalamus (by way of the limbic system), and it seems likely that impulses from other special senses, such as vision, also reach the hypothalamus. The neocortex is connected with the hypothalamus, chiefly indirectly by way of the limbic system, and impulses reach the hypothalamus by way of sensory pathways from all parts of the body. All of these nervous connections bring information relating to environmental changes, both external and internal. Moreover, some of the nerve cells in the hypothalamus are directly sensitive to changes in the temperature and the osmotic pressure of the blood and to levels of circulating hormones.

Pituitary gland. The pituitary gland, also known as the hypophysis, is an important endocrine organ lying mostly in the hypophysial fossa (depression) of the sphenoid bone, immediately below the brain, to which it is connected by a thin stem or stalk. The gland has a double embryological origin, developing partly from the pharyngeal (throat) region of the embryo and partly from that portion of the neural tube that gives rise to the diencephalon.

The part of the pituitary arising from the pharynx is glandular in structure and is termed the anterior lobe or adenohypophysis. This is further subdivided into so-called distal, intermediate, and infundibular parts. The anterior lobe contains different kinds of epithelial cells, all closely related to a network of blood passages called sinusoidal capillaries. These cells produce various hormones, and it is likely that a specific cell type is related to each hormone.

The anterior lobe is also characterized by a special blood supply. Briefly stated, certain arteries enter the median eminence, a ridge on the hypothalamus to which the stalk of the pituitary is attached. There they form sinusoidal capillaries. The capillaries then collect into vessels known as portal vessels, which descend into the anterior lobe and break up into a second set of sinusoidal capillaries. The latter vessels are the direct blood supply of the adjacent epithelial cells and drain into adjacent veins.

What is of major importance is the fact that nerve fibres from certain hypothalamic nuclei end in close relation-

ship to the sinusoidal capillaries in the median eminence (no hypothalamic fibres go to the adenohypophysis itself). It seems likely that the nerve cells in these hypothalamic nuclei produce substances, sometimes termed releasing hormones, that are transported down the axons to the sinusoidal capillaries and from there by the portal vessels to the anterior lobe, where they activate the epithelial cells.

The neurohypophysis

The portion of the pituitary that develops from the diencephalon is called the posterior lobe or neurohypophysis. It consists of nervous rather than glandular elements and comprises (1) the median eminence, (2) a short stalk or stem, and (3) a larger, lower part that rests in the hypophysial fossa (the median eminence is often classified as a part of the hypothalamus).

Certain hypothalamic nuclei, chiefly and perhaps entirely the supra-optic and paraventricular, send their axons down the stalk into the main part of the neurohypophysis. This part of the pituitary is considered to be a storehouse for the secretions produced by the nerve cells of the hypothalamic nuclei. These neurosecretions are transported by the axons of the cells to the neurohypophysis. Here either they are stored or they enter the bloodstream by way of adjacent capillaries. At least two hormones—vasopressin and oxytocin—reach the body by this route.

Pineal body. The pineal body or gland, also known as the epiphysis, is located at the posterior end of the third ventricle. In lower vertebrates, such as the frog, it seems to be a photoreceptor (a receptor for light stimuli, like the retina of the eye). It remains sensitive to light in higher vertebrates. In mammals, the pineal is heavily supplied by postganglionic sympathetic fibres whose varicosities are closely related to the parenchymal (functioning) cells of the pineal. In all species, the active principle of the pineal appears to be melatonin, whose function, however, may differ according to species.

Brainstem. The brainstem is formed, from above downward, by the diencephalon, the midbrain, the pons, and the medulla oblongata. It is rather similar in structure and function in all vertebrates and contains, in addition to nerve fibre tracts that descend and ascend through it, collections of nerve cells that (1) form the nuclei of origin of most cranial nerves, (2) form vital "centres" concerned with the reflex control of respiration, circulation, and other visceral activities, (3) connect with the cerebellum, and (4) are functionally associated with motor and sensory systems and mechanisms of consciousness, awareness, and sleep. A special feature is the association with pathways for some of the special senses, such as hearing, balance, and taste.

Functions of the brainstem

The diencephalic part of the brainstem has already been discussed in connection with the hypothalamus and pituitary. Those vital brainstem centres concerned with lower level, more circumscribed autonomic functions are located in the midbrain, pons, and medulla oblongata, chiefly in the last two. The centres consist of nerve cells that, in spite of the term centre, usually do not form circumscribed groups but, instead, tend to be scattered throughout the parts of the brainstem, especially the medulla oblongata. The cells receive afferent nerve impulses from higher centres (especially the hypothalamus), from the spinal cord, and from brainstem reflex paths. In turn, the constituent cells send their axons to the appropriate cells comprising the parasympathetic and sympathetic outflows. The various brainstem centres integrate and regulate a number of autonomic functions, such as blood pressure and gastrointestinal activity, and often can do so in the absence of higher control.

Cerebellum. The cerebellum is a large fissured structure that is attached to the back of the brainstem by three peduncles, or stalks, on each side. It is grossly divisible into two hemispheres and a midline connecting portion. The cerebellum is heavily supplied with input from the spinal cord, from higher centres, and from brainstem centres. In turn, it sends fibres to these regions and structures. The cerebellum has important modulating effects upon ongoing activity. While it does not of itself contain autonomic centres, most autonomic activities come under its modulating influence.

Sympathetic and parasympathetic outflow and ganglia. The sympathetic and parasympathetic cells in the brainstem and spinal cord comprise still lower and more specific levels of autonomic function. The peripheral ganglion cells to which they send axons form the lowest and most specific level of autonomic function.

As mentioned previously, the sympathetic and parasympathetic outflows are characteristically two-cell pathways. The first nerve cell is in the central nervous system, and its axon is called preganglionic. The next cell body is in a peripheral ganglion, and its axon, called postganglionic, supplies an effector. The general arrangement of this outflow, its connections with higher centres, and the structures it supplies are shown in Figure 16.

From E. Gardner, *Fundamentals of Neurology*, 5th edition (1968); W.B. Saunders Company

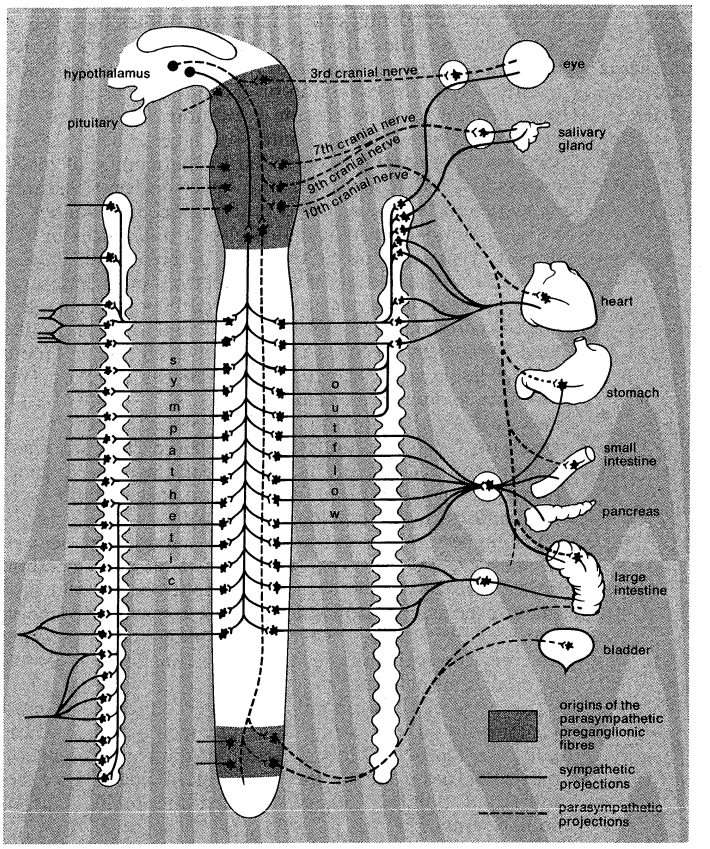


Figure 16: Schematic representation of the general arrangement of the autonomic nervous system. Autonomic fibres to the organs of the head and trunk are shown on the right side; the autonomic fibres on the left side represent the sympathetic outflow to blood vessels, sweat glands, and smooth muscle of hair follicles.

Outflow. As mentioned previously, the sympathetic outflow begins with cells that form a long column on each side of the thoracic and upper lumbar part of the spinal cord. Their axons, the sympathetic preganglionic fibres, leave by way of the corresponding ventral nerve roots, travel into the spinal nerves, and then leave the nerves by way of communicating branches, or rami communicantes. They synapse with ganglion cells in the sympathetic trunks, prevertebral plexus, or accessory ganglia or with cells of the adrenal medullae. The 12 thoracic and first two or three lumbar spinal nerves give off rami containing preganglionic fibres. All 31 spinal nerves receive rami containing postganglionic fibres from the sympathetic trunks. This means that each of the thoracic and upper lumbar spinal nerves has two or more rami communicantes. The nature of each ramus—that is, whether it is transmitting preganglionic or postganglionic fibres—is not readily ascertainable except by microscopic study. More often than not, however, the ramus that is attached to a spinal nerve more distally (further from the spinal cord) than other rami attached to the same spinal nerve

The rami communicantes

is preganglionic. This situation is of considerable practical importance to surgeons when it is necessary for them to cut the preganglionic fibres to a particular region or organ.

The columns of sympathetic cells in the spinal cord show regional differences in their distribution. For example, the cells in the first thoracic segment are the sympathetic outflow for the eye (Figure 17).

The parasympathetic outflow, as mentioned previously, has cranial and sacral parts. Both are more specifically organized than the sympathetic outflow. The parasympathetic cells of the cranial part are included in the nuclei of origin of the third, seventh, ninth, tenth, and 11th cranial nerves. The parasympathetic preganglionic fibres in the third nerve are destined for the eye; those in the seventh nerve, for the lacrimal gland, nasal glands, and salivary glands; those in the ninth nerve, for a salivary gland; and those in the tenth and 11th nerves, for viscera in the neck, thorax, and abdomen.

The preganglionic fibres in the third, seventh, and ninth cranial nerves end in certain well-defined ganglia called cranial or cephalic ganglia, from which postganglionic fibres are distributed to the various muscle cells and glands. The ganglia with which the preganglionic fibres in the tenth and 11th nerves synapse are scattered throughout the walls of the viscera. Examples of parasympathetic routes are given in Figures 17 and 18. The chief features of cranial distribution are summarized in Table 1.

Sacral portion of parasympathetic outflow

The sacral part of the parasympathetic outflow is derived from nerve cells that form a column in each half of the middle segments of the sacral part of the spinal cord (in man chiefly the third and fourth segments). The preganglionic fibres arising from these cells leave by way of the corresponding ventral roots, reach and pass through the sympathetic trunks, and emerge as the nerves known as the pelvic splanchnic nerves. These nerves enter prevertebral plexuses, chiefly the inferior hypogastric, from which fibres are distributed to the pelvic and genital organs and to the distal (most remote from the stomach) part of the large intestine (the proximal part is supplied by the vagus nerves). The ganglion cells with which the preganglionic fibres synapse are scattered in the walls of the viscera.

From E. Gardner, D.J. Gray, and R. O'Rahilly, *Anatomy*, 3rd edition (1969); W.B. Saunders Company

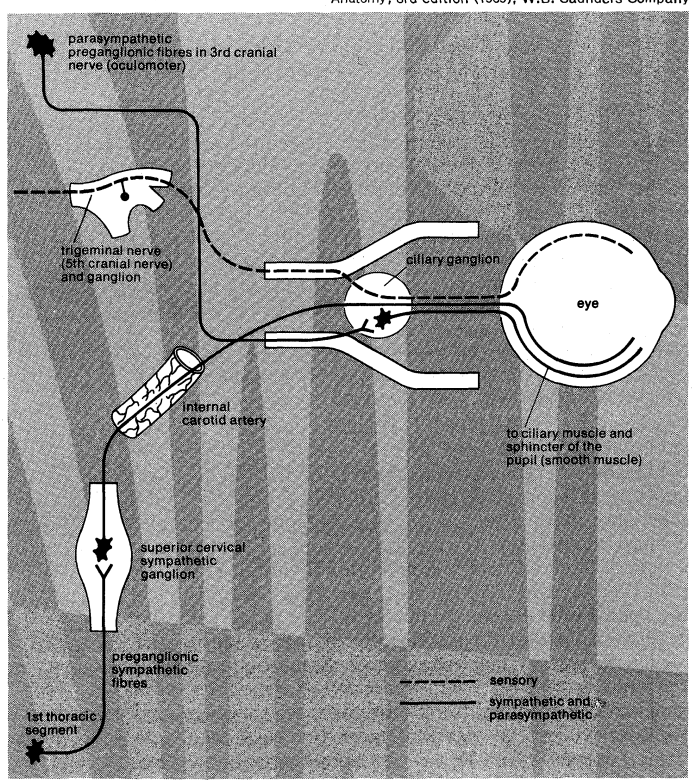


Figure 17: Schematic representation of the autonomic and sensory fibres to the eye.

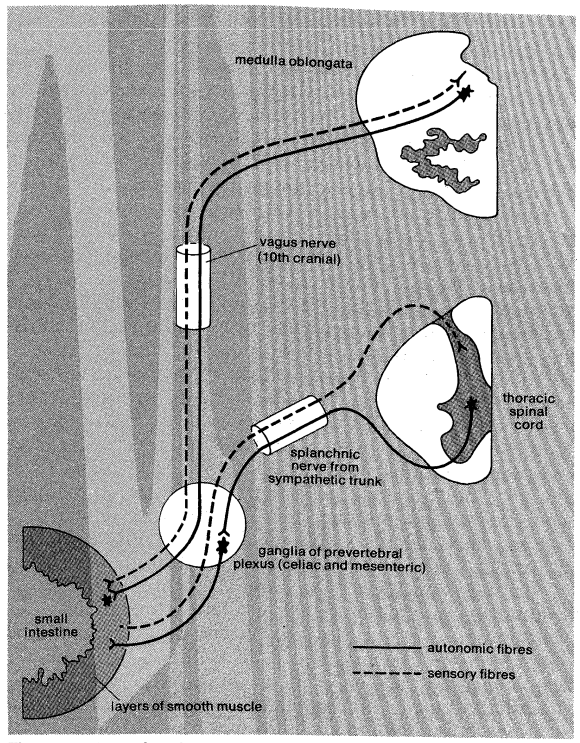


Figure 18: Functional components of the nerve supply of the small intestine, which has both a parasympathetic (from the brainstem by way of vagus nerves) and a sympathetic supply. Sensory fibres ascending in the splanchnic nerve to the spinal cord are chiefly pain fibres; sensory fibres ascending in the vagus nerve are for visceral reflexes.

From E. Gardner, D.J. Gray, and R. O'Rahilly, *Anatomy*, 3rd edition (1969); W.B. Saunders Company

Ganglia. The ganglia in which the preganglionic sympathetic fibres end are complicated in structure and are incompletely understood. The nerve cells that give rise to the postganglionic fibres are multipolar and are of varying sizes and staining characteristics. A single preganglionic fibre may synapse with many ganglion cells, perhaps as many as 30. It seems unlikely that these ganglia exist solely for relaying nerve impulses, and there is evidence suggesting complicated interactions between cells. Moreover, the ganglia may contain cells of the chromaffin type (see below), whose axons also end in relation to ganglion cells. As the lowest level of autonomic function, ganglia have the most restricted or specific functions. The axons leaving the uppermost ganglion of the sympathetic chain are entirely for the supply of the face, some of the neck, and most of the cranial cavity (Figure 17).

The sympathetic trunks

The sympathetic trunks, one on each side of the vertebral column, are each a ganglionated trunk extending from the base of the skull to the coccyx. The cervical (neck) part consists of three or four ganglia connected by an intervening cord (or cords), which consists mostly of preganglionic fibres. Rami communicantes, consisting chiefly of postganglionic fibres, are connected to each of the cervical spinal nerves, and the uppermost ganglion gives off a large branch or branches (postganglionic fibres) that accompany arteries to the face and cranial cavity. The thoracic parts of the trunks commonly have 11 or 12 separate ganglia each and are connected with each thoracic spinal nerve by two or more rami communicantes. In addition, branches run to adjacent viscera; the largest of such branches are known as splanchnic nerves. The lumbar parts of the trunks are seldom symmetrical, and the lumbar ganglia are irregular in size, position, and number. Each ganglion has two or more rami communicantes and is connected to the ventral rami of two or more spinal nerves. The pelvic parts of the sympathetic trunks consist largely of preganglionic fibres; each trunk usually has three or four ganglia and ends by forming an enlargement, the ganglion impar, with the other trunk in front of the coccyx. Each pelvic ganglion tends to be connected with but one spinal nerve by rami

communicantes that consist chiefly of postganglionic fibres.

As mentioned before, many of the preganglionic fibres pass through the sympathetic trunks and are carried by branches, such as the splanchnic nerves, to the prevertebral plexuses. These plexuses lie in front of the vertebral column and are named for adjacent viscera, for example, cardiac (heart), celiac (abdominal), and mesenteric (the mesenteries are folds of membrane continuous with the membrane—peritoneum—lining the abdomen and pelvis). Sympathetic ganglion cells, with which the above-mentioned fibres synapse, occur in these plexuses. They are numerous enough so that they form ganglia, some large enough to be named. The postganglionic fibres arising from the ganglion cells go directly to adjacent viscera.

The accessory ganglia

Some of the preganglionic fibres in spinal nerves synapse in accessory (intermediate) ganglia without reaching sympathetic trunks. These ganglia are collections of sympathetic ganglion cells scattered along spinal nerves and rami communicantes, especially in the cervical, lower thoracic, and upper lumbar regions. Most of these cells send their axons into the spinal nerves with which they are associated. These ganglion cells would not be affected by usual surgical procedures (sympathectomy) designed to cut autonomic fibres and are, therefore, a source of residual autonomic function after such an operation.

The reasons for the differing locations of sympathetic ganglion cells are considered later, in the section on early development.

Finally, it must be emphasized that sensory fibres arising from viscera are contained within the various nerves. Depending upon their functions, they may reach the brainstem by way of the vagus nerves, or traverse the sympathetic trunks, reach spinal nerves by way of rami communicantes, and then the spinal cord by way of dorsal roots.

The cranial parasympathetic ganglia likewise have a complicated structure and arrangement that is incompletely understood. In general, each cranial parasympathetic ganglion has postganglionic sympathetic fibres originating in the superior cervical ganglion and passing through the ganglion, sensory fibres passing through to a cranial nerve and thus to the brain, and parasympathetic preganglionic fibres from the appropriate cranial nerve and synapsing in the ganglion. Postganglionic parasympathetic fibres are distributed to adjacent structures, as indicated in Table 1. In general, each preganglionic fibre synapses with far fewer ganglion cells than is the case with sympathetic ganglia. In general, also, no parasympathetic fibres seem to go to blood vessels, except for a few arteries to certain viscera, or to the smooth muscle and glands of the limbs and body wall. The general distribution of autonomic fibres is summarized in Table 3.

The parasympathetic ganglion cells in the walls of the thoracic, abdominal, and pelvic organs rarely form recognizable ganglia but instead are usually scattered. In the alimentary canal, however, these cells occur between the muscle layers in arrangements complicated enough to warrant the term plexuses. Moreover, there is evidence of intrinsic connections and interactions independent of the external nerve supply; these may be a factor in intrinsic activity of the organ.

Adrenal (suprarenal) glands. The sympathetic system has a special situation in that some preganglionic fibres end in relation to, or "synapse" with, certain gland cells rather than ganglion cells. The gland cells are the medullary cells of the adrenal glands and the cells of the chromaffin system.

The adrenal glands are small, paired endocrine glands; in man, they lie above the kidneys. In most animals the glands are near the kidneys but not necessarily above them.

Adrenal structure and secretions

Each gland is made up of two endocrine glands, the cortex and the medulla. The cortex, or outer substance, produces steroid hormones that are essential to life. The medulla, or inner substance, the cells of which are supplied by preganglionic sympathetic fibres and are therefore analogous to sympathetic ganglion cells, produces epinephrine and norepinephrine, also called adrenaline and noradrenaline. The effects of epinephrine are generally similar to those resulting from stimulation of the sympathetic nervous system. The sympathetic system can therefore enhance its own action, because it can stimulate the adrenal medulla and cause the release of epinephrine and norepinephrine into the circulation.

Chromaffin system. With regard to the chromaffin system, cells similar to those of the adrenal medulla are commonly found in front of the vertebrae, especially along the aorta (the major artery that conveys blood to the body from the left ventricle of the heart). These paraganglia or para-aortic bodies (which may be fairly large and constant in position), together with the adrenal medullae, constitute the chromaffin system, so named because granules in the cells stain strongly with chromium salts.

Like the cells of the adrenal medulla, most of the paraganglia secrete epinephrine or norepinephrine in response to stimulation by preganglionic sympathetic fibres. Little is known, however, of the general purpose of the chromaffin system. It is of interest that certain cells that appear to be nerve cells and that occur in sympathetic ganglia exhibit the chromaffin reaction.

EARLY DEVELOPMENT

An understanding of the early formation and development of the nervous system is essential for an understanding of the anatomy and relationships of the autonomic nervous system. The following is a brief account of such development, with emphasis on the origin of the sympathetic and parasympathetic outflow.

The entire body develops from two highly specialized cells: one, the ovum, or egg, of the female; the other, the spermatozoon, or sperm, of the male. Development begins with the penetration of an ovum by a sperm cell, thus initiating a process known as fertilization. The human embryonic period covers about the first eight weeks after fertilization. At the end of this time, during which enormous cell proliferation and differentiation occur, the embryos are in many respects like miniature adults. Differentiation is nearly completed, and subsequent changes consist chiefly of growth and maturation of structures already formed. The human fetal period, which is characterized by such growth and maturation, is normally about seven months, from the end of the embryonic period until birth.

After fertilization, the proliferation of cells soon results in a three-layered arrangement. The cells on the back, or dorsal, part of the embryo form a layer called ectoderm. The ventral layer of cells is referred to as endoderm, and the more diffuse collection of cells between the ectoderm and the endoderm is termed mesoderm. Such structures as the outer skin and the nervous system develop from the ectoderm. The skeleton, muscles, connective tissue, heart, and blood vessels are derived from mesoderm. The digestive and respiratory tracts and their glandular derivatives arise from the endoderm. This early arrangement of embryonic cells into a pattern of three germ layers is characteristic of all vertebrates.

The cells in the dorsal layer, the ectoderm, proliferate rapidly and form a neural groove and then a neural tube, which separates from the overlying ectoderm. During this separation, the junctional ectodermal cells become detached and form cells called neural crest cells. The entire adult nervous system, except for its blood vessels and certain supporting cells, is formed from the neural tube and the neural crest cells. In spite of complex changes in structure, the central nervous system remains a tube throughout the lifetime of an individual. The nerve cells that develop in the ventral part (the basal plate) of the neural tube become motor cells. The nerve cells that develop in the dorsal part (the alar plate) become associated with sensory and reflex paths. Some of the cells in the neural crest give rise to the sensory cells of the spinal ganglia. Other neural crest cells migrate, however, and, together with cells migrating from the basal plate, develop into nerve cells of the autonomic ganglia and cells of the medullae of the adrenal glands, as well as chromaffin cells. Thus, from early stages, the adrenal medullae

The neural tube and neural crest cells

are closely associated with the autonomic nervous system.

Of the migrating cells that form ganglion cells, some form accessory ganglia. Others migrate still further to the sides of the vertebra and form the sympathetic trunks. Still others proceed still more peripherally and form the ganglia of the prevertebral plexuses, the medullae of the adrenal glands, and chromaffin cells. Those cells that migrate from the brainstem and the sacral part of the neural tube form the peripheral ganglia of the parasympathetic system.

It is a matter of considerable interest and potential importance that the development and growth of the sympathetic nervous system is affected by a growth factor and perhaps depends upon it. The growth factor is a protein that can be obtained from a number of sources, including developing nervous system and mouse salivary glands. When it is injected into young chickens or mice, sympathetic and sensory ganglia enlarge, and new ones may form. An antiserum to (a serum that contains antibodies against) this nerve growth factor can virtually obliterate the sympathetic system in young mice. The factor has been studied in only a few species, and its significance in man is still uncertain, although there is preliminary evidence suggesting that it has similar effects on sympathetic ganglion cells of human fetuses.

The part of the neural tube in the head region proliferates more rapidly and forms the brain, whereas that part within the vertebral canal forms the spinal cord. That part forming the brain soon exhibits three primary enlargements, the forebrain, midbrain, and hindbrain, with the forebrain and hindbrain rapidly showing secondary enlargements. The derivatives of these various enlargements, or vesicles, are summarized in Table 2. The original cavity also persists and forms the ventricles and central canal of the adult brain.

COMPARATIVE FEATURES

Although there has been no systematic or comprehensive study of the structure and function of the autonomic nervous system on a comparative basis, involving all phyla, there is, nevertheless, a substantial amount of information about a number of features. With regard to peripheral components, for example, the study of domestic and experimental animals has led the way to what is known about synaptic transmission, both chemical and electrical, centrally and peripherally. Comparative studies have also dealt with topographic details and variations in anatomical arrangements. These details and variations are important in surgical procedures and experimental studies and in the interpretation of results. For example, in experimental studies of the nerve supply of the heart, it is essential to know the detailed anatomy. Otherwise, unrecognized species differences may result in unwarranted interpretation of results. As another example, it is notable that the relationships of the adrenal cortex and medulla show major differences according to phyla. In fish, the cortex and medulla of the adrenal are separate. In amphibians the two tissues are united but do not intermingle. In birds, and especially in mammals, the gland consists of a chromaffin medulla surrounded by a yellowish lipoidal cortex.

The purpose of these few brief paragraphs is not to provide a detailed account of comparative features but rather to emphasize the following: in spite of species differences in topography, anatomy, and local functions, the fundamental characteristics of the autonomic nervous system are similar in most vertebrates. These characteristics can be outlined as follows: cardiac muscle, smooth muscle, and glands are supplied by motor pathways each formed by two successive nerve cells, the first located in the central nervous system and the second in a peripheral ganglion. These paths have features such that the motor outflow can be divided into sympathetic and parasympathetic parts. Moreover, the motor outflow is intimately associated in all of its functions with sensory pathways and mechanisms, and itself forms the lowest of a series of levels dealing with autonomic functions, all the others being located in the central nervous system.

Fundamental characteristics of the autonomic nervous system

Most of the parts of the central nervous system concerned with autonomic functions are remarkably similar in all vertebrates, especially as regards location and general organization. The differences that do exist relate to phylogenetic and comparative levels and to differences in habitat. There is, for example, no true cerebral cortex in fish but only a single-layered cortex; this, in higher vertebrates, is found in the so-called limbic system. In following the comparative scale upward, scientists see the emergence of higher, more complex, and, at the same time, more generalized functions as related to the increasing size and complexity of the cerebral hemispheres. With regard to autonomic functions, the hypothalamus is the dominant area in the lower vertebrates and retains a major function throughout the vertebrate scale. Like the brainstem and the spinal cord, the hypothalamus in all vertebrates is similar in structural organization and anatomical details, although specific functional features obviously vary according to species. Similarly, the portal system supplying the pituitary gland, or the equivalent of a portal system, has been found in all vertebrates studied.

In general, then, the differences that occur relate chiefly to anatomical details and to specific functions. The respiratory centre in fish obviously does not function like that in mammals. The levels of sympathetic outflow differ according to species. The fine microscopic structure of ganglia and the biochemistry of synaptic transmission in ganglia may show marked differences according to phyla. Nevertheless, as mentioned above, the fundamental characteristics of the autonomic nervous system are similar in most vertebrates.

AUTONOMIC TRANSMISSION

The anatomical classification that divides the autonomic nervous system into sympathetic and parasympathetic parts is unsatisfactory in certain respects. It implies a rigid or sharp distinction between these divisions that functionally does not always exist. Also, some organs and tissues are supplied by sympathetic fibres that function like parasympathetic fibres, and others may be supplied by but one set of fibres. There is in common use another classification that is based upon the types of chemical postganglionic transmitters and upon reactions to certain drugs. In it, postganglionic fibres are called cholinergic if they release the compound acetylcholine at their terminals and adrenergic if they release noradrenaline (norepinephrine). Although this classification is also unsatisfactory in certain respects, it nevertheless is an important one. Before it is taken up in more detail, it is necessary to review briefly how nerve impulses are transmitted from one nerve cell to another or from one nerve cell to an effector. It is important to realize that much of what is known has been derived from studies of autonomic ganglia, autonomic neuro-effector junctions, and motor endings in skeletal muscle.

Cholinergic and adrenergic fibres

When a nerve impulse (a wave of transient depolarization [reversal of positive and negative poles] of the surface membrane of an axon accompanied by a transient change in membrane potential) reaches a synaptic junction or neuro-effector junction, changes occur that involve what is termed synaptic transmission or endplate (motor or neuro-effector) transmission. The precise nature of these changes depends upon the origin of the axon, the distribution and functions of its cell of origin, the existing excitability of the succeeding nerve cell or effector, and a host of other factors.

Synaptic transmission can be either electrical or chemical. Electrical synapses are more common in invertebrates than in vertebrates but in all instances seem to be less common than chemical synapses. In vertebrates, they occur in association with certain giant nerve fibres in fish, they are present in certain parasympathetic ganglia, such as the ciliary ganglia of young birds, and they may be present in the developing nervous system. Synaptic endings in electrical synapses are usually large, and the synaptic region is a tight junction. Apparently, the current flow from the nerve impulse in the axon is sufficient to depolarize the membrane of the succeeding nerve cell and cause it to fire off.

Chemical transmission

Chemical transmission occurs somewhat as follows. A specific chemical compound called a transmitter is synthesized, mostly in the nerve cell body. From there it travels down the axon, to be stored in structures called synaptic vesicles. Some synthesis may also occur in the nerve terminals themselves. When a nerve impulse arrives at the ending, it liberates the transmitter in packets termed quanta. The released transmitter then diffuses across the synaptic cleft (the very narrow space between the two cells) and combines with a substance, called a receptor, in the membrane of the succeeding nerve cell or of the effector. The term receptor in synaptic and neuroeffector transmission usually means combinations of transmitter and cell or of drug and cell that initiate a sequence of specific effects. The first is a change in the permeability of the membrane to ions (substances in solution that bear an electrical charge), resulting in a specific kind of ionic flow through the membrane, accompanied by a change in electrical potential. These changes may be excitatory; that is, they may produce a local depolarization of the membrane. This local change, either by itself or in combination with other local changes nearby, may be of sufficient magnitude and involve enough area of the membrane so that the depolarization spreads in a continuing process. In the case of a nerve cell, the change spreads outward along the axon. In the case of an effector, the result is contraction or secretion. On the other hand, the transmitter may be inhibitory; that is, it may increase the polarization of the membrane and thereby prevent discharge or stop ongoing activity. Unused transmitters may be "destroyed" by enzymatic activity, and the products of such destruction may be used in the synthesis of transmitters. Some unused transmitter may diffuse back into the axon from which it was released.

Noradrenaline and acetylcholine are excitatory transmitters in the autonomic nervous system and in certain parts of the central nervous system. Dopamine may also be a transmitter in the central nervous system. Other neurologically active substances that might be transmitters include isoprenaline, serotonin, and perhaps histamine. Much less is known about the nature of inhibitory transmitters. There is substantial evidence that one of the amino acids, gamma aminobutyric acid, is an inhibitory transmitter in the central nervous system, and it is possible that its amino-acid precursor, glutamic acid, is an excitatory transmitter.

Functions of acetylcholine

Acetylcholine. The compound acetylcholine is the chief transmitter at all autonomic ganglion cells, both parasympathetic and sympathetic, it is the transmitter at all parasympathetic postganglionic terminals and at the motor nerve endings of skeletal muscle fibres, and it is the transmitter between certain nerve cells in the brain and spinal cord. Since parasympathetic postganglionic fibres release acetylcholine at their terminals, they are classified as cholinergic fibres. Some sympathetic postganglionic fibres (to sweat glands and smooth muscle of hair follicles) also release acetylcholine and are called cholinergic.

The synthesis of acetylcholine, which requires a continuing supply of glucose and of choline (a B-complex vitamin), is dependent upon a specific enzyme, choline acetyltransferase. Once synthesized, acetylcholine travels down the axon to the axon terminals (some may be synthesized in the terminals), where it is present in several different forms, or compartments. One of these forms, called depot acetylcholine, is stored in synaptic vesicles. When acetylcholine is released from these vesicles by the arrival of nerve impulses, it diffuses across the synaptic cleft and combines with a specific receptor substance in the membrane of the succeeding nerve cell. The result of this combination is the change in membrane permeability mentioned above. Although the exact nature of the receptor substance is unknown, it seems likely that there are several forms, so that the precise action of acetylcholine depends, among other things, on the nature of the receptor substance. Moreover, there is evidence that chromaffin cells in autonomic ganglia may release noradrenaline, which then can block the action of acetylcholine or inhibit the ganglion cell.

Some acetylcholine spontaneously diffuses into the synapse cleft, and some that is released by the nerve impulse remains as excess. In both cases the acetylcholine is split into choline and acetic acid by a specific enzyme, cholinesterase, also called acetylcholinesterase. The choline becomes available for synthesis of more acetylcholine. The action of acetylcholinesterase can be prevented by a compound called eserine (also termed physostigmine), which thereby accentuates the effects of acetylcholine.

Acetylcholine is synthesized in the postganglionic neuron as in the preganglionic and is similarly stored in synaptic vesicles. When released, it diffuses across the intercellular space to the membrane of the effector, much as noradrenaline does. Here it combines with a receptor, thereby initiating the specific activity of the effector. As in the case of synaptic transmission, the complexities of the receptor are not yet fully understood.

Noradrenaline. Noradrenaline (norepinephrine), one of a group of important biogenic amines, is released at the terminals of adrenergic fibres, which comprise most postganglionic sympathetic fibres. Together with dopamine and serotonin, it is also found in high concentrations in certain parts of the brain. It seems to be present in autonomic ganglia, where its release by chromaffin cells has been postulated, but its exact role and importance here are not generally agreed upon.

Noradrenaline is synthesized in the nerve cell body and probably in the axon terminals also. It is stored in synaptic vesicles and, when released, diffuses across the intercellular space (or synaptic cleft in central nervous system synapses) to combine with a receptor in the membrane of the effector. Most of the noradrenaline released is taken up again into the nerve terminals to be stored. Some, however, is metabolized outside the cell by the enzyme catechol-O-methyltransferase, and some is metabolized inside the cell by the enzyme monoamine oxidase. The sequence of events following the release of noradrenaline may be excitatory or inhibitory, depending upon the structure or organ and its state of ongoing activity or level of hormonal activation and depending upon the species. Some of these differences may be explained by the fact that two receptors seem to be involved, one termed the alpha receptor and the other called the beta receptor. The alpha receptors have been thought, in general, to be excitatory in nature; and the beta receptors, inhibitory. There are important exceptions to this rule, however, and many aspects remain to be clarified. Noradrenaline acts chiefly on the alpha receptor; hence certain effects that it has on sympathetic activity—for example, dilatation of the coronary arteries supplying the heart—are difficult to explain. It is of considerable importance that the synthetic drug isoprenaline, which is structurally similar to noradrenaline, is the most potent mimicker of beta-receptor activity. Whether isoprenaline occurs naturally and serves as a transmitter is quite uncertain. Beta-receptor activities include dilatation of arteries supplying the heart and skeletal muscles (but not in all species), acceleration of the heart, and relaxation of bronchioles. Alpha-receptor activity is characterized by constriction of the arteries supplying the skin, abdominal organs, and heart, by dilatation of the pupil, and by relaxation of the intestine. The situation is further complicated by the fact that certain mechanisms, such as dilatation of a blood vessel, might be explained as resulting from a diminution of alpha-receptor activity as well as from an increase in beta-receptor activity. Moreover, adrenaline, which is released into the bloodstream by cells of the adrenal medulla after stimulation by the acetylcholine of preganglionic sympathetic fibres, has both alpha and beta actions. Hence, it can enhance the general actions of the sympathetic system.

Alpha and beta receptors

What emerges from the present, rather conflicting state of knowledge is that many important variables, including species differences, must be kept in mind during consideration of autonomic functions. Such functions in man, insofar as they are known, are summarized in Table 3.

As indicated by Table 3, most organs are supplied by both systems. One of the most notable exceptions has to do with blood vessels. Nearly all peripheral blood vessels

are supplied by the sympathetic system (a few vessels in the head and in the pelvis are supplied by cholinergic, vasodilator fibres, which may be parasympathetic in origin). With these few exceptions, the sympathetic system mediates both vasoconstriction and vasodilatation. One of the best examples is in connection with the coronary arteries, as mentioned previously. Similarly, the changes in blood flow associated with constriction and dilatation are important components of temperature regulation, mentioned earlier in this article. It is difficult to explain dilatation solely on the basis of a relaxation of vessel walls after a decrease in sympathetic activity. Constriction seems much more related to alpha-receptor activity, and dilatation to beta-receptor activity.

Neuropharmacology. Much of the present knowledge of the autonomic nervous system has been obtained through the experimental use of a variety of important drugs, many of which bring about profound physiological, behavioral, and emotional changes because of their effects on metabolism and synaptic transmission. A number of these drugs can be classified according to the kinds of autonomic effects they have. These effects can be considered according to whether they are chiefly parasympathetic or sympathetic in nature or can involve ganglionic transmission.

Muscarinic and nicotinic effects

A number of drugs act on the parasympathetic postganglionic terminals or the neuro-effector junction or the effector, in a manner resembling that of acetylcholine. One such drug is muscarine, which acts on the smooth muscle of viscera much as acetylcholine does; the actions of acetylcholine at postganglionic terminals are often called its muscarinic action. This action is blocked by atropine, which can be a severe poison. The drug nicotine stimulates autonomic ganglion cells when applied to the cells in low concentrations, but in high concentrations it depresses or blocks transmission. Acetylcholine acts similarly, and its properties as a ganglionic neurotransmitter are known as its nicotinic effects. Muscarine is termed a parasympathomimetic drug—that is, one that, when administered, stimulates or inhibits autonomic effectors as in parasympathetic postganglionic stimulation. In general these drugs include acetylcholine and some of its derivatives, alkaloids such as pilocarpine and muscarine, and a variety of anticholinesterases such as eserine. Many are produced synthetically, and many may have central actions also.

Sympathomimetic drugs stimulate or inhibit effectors as in sympathetic postganglionic stimulation. These drugs often have profound effects on the central nervous system also. Some of them are the biogenic amines that are also normal autonomic components, namely noradrenaline, adrenaline, isoprenaline, and dopamine. When administered, they act directly on alpha or beta receptors or on both. A number of other drugs that are also sympathomimetic, including amphetamine, cocaine, and ephedrine, act by bringing about the rapid release of stored noradrenaline in a physiologically active form. Because this release occurs in the central nervous system also, the accentuation of noradrenaline functions brings about profound behavioral and mood changes. Amphetamine is a notable example of a drug that has serious central nervous system effects brought about by the mechanism just outlined. It is of interest that another drug, reserpine, likewise brings about the release of noradrenaline, but much more slowly, so that it is deaminated (is caused to lose its amino fraction) by monoamine oxidase within the nerve cell and is released in a physiologically inactive form. The result is an induced behavioral depression and is the basis for the use of reserpine as a tranquillizer.

Monoamine oxidase inhibitors

Also sympathomimetic are the drugs that inhibit monoamine oxidase. As mentioned previously, this widely distributed enzyme is responsible for the intracellular degradation of noradrenaline. It likewise degrades adrenaline and dopamine, and by deamination it inactivates 5-hydroxytryptamine (serotonin). The inhibitors are a heterogeneous group of drugs that by blocking the actions of monoamine oxidase cause enhancement of sympathetic and serotonin activity.

Finally, there are drugs that produce synaptic and effector blockade. They may block transmission in autonomic ganglia; examples include certain quaternary ammonium compounds, as well as anesthetics such as ether, and compounds such as nicotine in high concentration. Peripherally, some drugs compete with acetylcholine at the receptor site; that is, they have an antimuscarine effect. Examples are atropine and scopolamine. Some drugs prevent the release of transmitter; an example is guanethidine, a powerful drug for combatting high blood pressure. Others block the effects of noradrenaline directly at either the alpha- or the beta-receptor sites. Examples of an alpha-receptor blocking drug are the ergot alkaloids (which have other actions also) and thymoxamine. Beta-receptor blocking agents include the compound propranolol.

GENERAL AND INTEGRATIVE FUNCTIONS OF THE AUTONOMIC NERVOUS SYSTEM LEVELS

The specific motor effects of the autonomic nervous system upon various organs are summarized in Table 3.

Table 3: Chief Functions of the Autonomic Nervous System in Man

organ	parasympathetic symptom	sympathetic symptom
Eye	stimulates ciliary muscle and sphincter pupillae—pupillary constriction to light and during accommodation	stimulates dilator pupillae and possibly radial fibres of ciliary muscle—pupillary dilatation
Lacrimal, nasal, palatine, and salivary glands	stimulates secretory cells—serous and mucous secretions	no important effect
Heart	inhibits intrinsic cardiac activity—heart rate slows	stimulates intrinsic cardiac activity—heart rate increases
Tracheobronchial tree	stimulates secretory cells and smooth muscle—serous and mucous secretions, narrowing of bronchioles	inhibits smooth muscle—relaxation of bronchioles
Alimentary canal	stimulates secretory cells and smooth muscle—digestive secretions, peristalsis, evacuation	inhibits smooth muscle—decreases peristalsis
Liver and pancreas	stimulates pancreatic cells (effect on liver unknown)	no important effect
Adrenal medulla	no effect	stimulates secretory cells—secretion of epinephrine
Urinary bladder	stimulates smooth muscle (detrusor)—emptying of bladder	no effect on emptying; may activate internal sphincter during ejaculation
Genital organs	uncertain and variable; may stimulate smooth muscle and glands; vascular dilatation	uncertain and variable; may stimulate vasoconstriction
Blood vessels	no important effect on most blood vessels	constriction and dilatation
Sweat glands	no effect	stimulation of secretory cells, mostly by cholinergic fibres—sweating
Arrectores pilorum	no effect	stimulation of smooth muscle, mostly by cholinergic fibres—erection of hair

These actions are important components of the general, more complex mechanisms of homeostasis, as well as of the outward expressions of emotions and behaviours. As mentioned previously, homeostasis is the maintenance of the internal environment; for example, body temperature, fluid balance, and ionic composition of the blood. In helping to carry out these functions, the two divisions of the motor outflow, sympathetic and parasympathetic, exhibit distinctive functional attributes. These are outlined briefly before the functions of the various levels are described.

The sympathetic system tends to respond as a whole, especially during emergencies or sudden environmental changes, and is therefore an important part of the mechanisms by which one responds to stress. Among the situations that cause stress are pain, rage, fright, exercise, cold, many drugs, and asphyxia. All of these may evoke similar patterns of response. For example, a stressing situation that results in rage or fear may also result in a rapid increase in blood pressure, pulse rate, heart output, sweating, and blood sugar; all are changes designed for "fight or flight." These acute responses to stress result

Responses to stress

from sympathetic discharge initiated by cortical and hypothalamic activity. They are generalized and widespread because of the organization of the system (for example, one preganglionic fibre synapses, or interconnects, with many ganglion cells) and also because the medullary cells of the adrenal glands are stimulated, causing epinephrine to be released into the bloodstream. Epinephrine may maintain or enhance many sympathetic activities.

In addition to the acute responses to stress, there may be more slowly developing changes in metabolic activities and defense mechanisms. These are brought about by the activation of the hypothalamic–pituitary system, leading to an increased secretion by the pituitary of adrenocorticotropic hormone. This acts on the cortex of the adrenal glands, leading to a greater output of the cortical steroid hormones, with changes in metabolic activities and cellular composition of the blood.

Characteristic functions of the parasympathetic system

The parasympathetic system is concerned with the initiation and maintenance of a number of specific functions, such as digestion, intermediate metabolism of foods, and excretion. These actions are usually initiated in response to specific stimuli; hence the necessity for specific anatomical arrangements and reflex pathways. Widespread connections such as are characteristic of the sympathetic system would interfere with, or make impossible, the specific parasympathetic functions. Widespread, generalized parasympathetic actions can occur, but they are often abnormal responses to unusual situations, such as bladder and bowel evacuation during unusual stress.

Cerebral hemispheres. The highest mental and behavioral activities characteristic of man take place in the cerebral hemispheres, particularly the cerebral cortex. Some cortical areas control muscular activity; others are concerned with sensory perception, and still others with speech mechanisms and association and integration. No region of the cortex is specifically and solely involved in autonomic activity, however. Rather, when the cortex controls, regulates, or activates a visceral function, it does so during the course of some other activity. For example, cortical activity leading to muscular contraction—a voluntary activity—also brings about an increase in blood flow in the contracting muscles—an instance of autonomic activity. Many other examples of participation in local autonomic mechanisms could be provided, all of them mediated by cortical projections to the diencephalon and brainstem. As concerns the present article, what is of particular importance is the integrated role of cortico-diencephalic mechanisms in emotional behaviour.

Emotions and behaviour. There is as yet no completely satisfactory theory or experimental formulation of emotional behaviour. It is known that the cortico-diencephalic mechanisms, including the limbic system, control or activate a wide variety of somatic and autonomic phenomena closely related to the broad range of behavioral activities customarily associated with emotional expression. Within these complex, overlapping, highly integrated mechanisms, the cortex can be viewed as the "high-level mediator," probably suppressing or inhibiting many outward expressions of behaviour (sham rage can be elicited in animals with much of the cortex removed). The diencephalon, in particular the hypothalamus, activates, controls, and integrates the peripheral mechanisms that bring about the bodily changes associated with emotional behaviour, namely, autonomic, somatic, and endocrine activities.

Effects of emotional upsets

The cortico-diencephalic mechanisms have important implications for disorders of behaviour. There is no autonomic activity that cannot be modified, or even completely disorganized, by emotional upsets. Beginning with what is usually considered relatively "normal," one can cite blushing. This is a sudden dilatation of blood vessels, usually but not always limited to those of the face, in response to an emotional situation, or the memory of one. Fright, anger, or apprehension may completely upset digestive processes and, at the same time, may cause blood pressure to rise and heart rate to increase. Students are well acquainted with such phenomena before examinations; speakers may know them as stage fright.

Dramatic behavioral changes may follow removal of various parts of the cortico-diencephalic mechanisms. Loss of cortex from certain parts of the frontal lobes of the brain may be followed by marked alterations of affective behaviour often known as "changes in personality." Temporal-lobe damage may be followed by docility and compulsive behaviour. Lesions of the amygdaloid bodies have likewise been associated with docility (taming of wild animals) and also with hypersexuality. Other symptoms include hallucinations, disordered recognition, disturbances of reality, sensory fits, and psychomotor epilepsy. In nearly all of these situations or conditions, a significant component is disordered autonomic function.

It is of interest that there seems to be a fairly constant relationship between the biogenic amines, which are also transmitters or related to transmitters, and the affective or behavioral states. Norepinephrine is found in high concentrations in certain parts of the brain, such as the hypothalamus. Drugs that depress or inactivate norepinephrine centrally cause sedation. The converse is generally associated with stimulation or excitement. For example, the amphetamines are used as psychic stimulants. They release physiologically active norepinephrine from nerve cells (and may also block its inactivation). Similarly, and as mentioned previously, reserpine interferes with the intraneuronal binding of catecholamines and serotonin, causing them to be released slowly and to be inactivated by monoamine oxidase. The result is a depression; hence the use of reserpine as a tranquillizer. It is clear, therefore, that many drugs with profound psychic and behavioral effects bring about these effects by acting upon central and peripheral autonomic structures and mechanisms, especially the transmitting systems.

Conditioned responses. The modification of behaviour on the basis of past experience is thought to be primarily a function of the cerebral cortex, although obviously requiring the participation of many other neural structures and mechanisms. It is now known that, contrary to past, indeed current, beliefs, many visceral responses can be learned. This is important for an understanding of the functions of the autonomic nervous system and normal homeostasis and for the cause and cure of abnormal psychosomatic symptoms. It also has major implications for theories of learning. How the learning of visceral responses is studied requires a brief description of conditioning.

Unconditioned and conditioned reflexes

The sight or smell of food may be as potent a stimulus for salivation as is taste. The objective response—salivation—which is an autonomic response activated by parasympathetic fibres to salivary glands, is present in newborn infants, but only later do sight and smell induce salivation. Salivation in response to taste is an unconditioned reflex, brought about by the stimulation by taste fibres of the cells of origin of the parasympathetic fibres in the third cranial nerve.

Many unconditioned reflexes may be modified by, or adapted to, new stimuli. The sound of a bell is not ordinarily followed by salivation; yet it is well-known that, if an animal is given food, or if a mild acid is placed in its mouth so that salivation occurs, and if this is accompanied or shortly preceded by another stimulus such as the sound of a bell, after repeated trials the sound alone may induce salivation (an example of the type of conditioning that is called Pavlovian). This is a conditioned response—an increase in salivary-gland activity in reaction to an auditory stimulus.

Conditioned responses generally have to be reinforced. In classical, or Pavlovian, conditioning, which is generally thought to be involuntary, the reinforcement must be by an unconditioned stimulus that already elicits the specific response to be learned. In the conditioning called operant conditioning or trial-and-error or instrumental learning, which is believed to be responsible for voluntary behaviour, the reinforcement is a reward that has no natural relationship to the response. This reward can strengthen any immediately preceding response.

Rigorously controlled studies have shown that visceral responses can be learned through both classical and operant conditioning. An example is the psychogalvanic re-

Psychogalvanic reflex

flex. This is an increase in sweating (mediated through sympathetic fibres to sweat glands) during emotional states. Where it occurs, as in the hands, resistance to the flow of an electrical current decreases, roughly in proportion to the amount of sweating. Thus, the response can be measured in a way that tends to furnish objective evidence of emotional reaction. In classical conditioning, this reflex can be tied to a stimulus that does not normally elicit it. In operant conditioning, this reflex and many other autonomic activities (including heart rate and salivation) can be learned; that is, brought about by the subject in order to obtain a reward or some desired change. It must be emphasized, however, that the mechanisms involved in conditioning, whether classical or operant, somatic or visceral, are still obscure.

Diencephalon. *Hypothalamus.* As has been mentioned previously, the hypothalamus is that part of the corticodiencephalic mechanisms that activates, controls, and integrates the peripheral autonomic mechanisms, endocrine activity, and many somatic (bodily) functions. For example, it is known to regulate water balance, body temperature, the development of secondary sex characteristics, sleep, and food intake. These functions will be illustrated by brief discussions of a major function carried out by hypothalamic projections to the brainstem and of functions mediated by hypothalamic control of the pituitary gland.

In the regulation of body temperature, described earlier in this article, changes in external temperature stimulate nerve receptors in the skin. The resulting nerve impulses reach the spinal cord and brainstem, where they initiate local reflexes. They also ascend and reach the hypothalamus, where they activate nuclei that initiate a range of activities designed to conserve heat. In addition, certain cells in the hypothalamus are directly sensitive to changes in temperature of the blood that flows through the area in which the cells are located. The overall result is that, if the external temperature drops, some hypothalamic nuclei initiate or regulate the sympathetic activities necessary to prevent heat loss. The converse occurs if the external temperature rises. Under more extreme conditions of cold, metabolic processes would be initiated to increase the production of heat (hypothalamic activation of the pituitary cells that produce the thyroid-activating hormone). The thyroid gland regulates the general metabolic level and, when rapid increase in heat production is necessary, can bring about the characteristic contractions of skeletal muscle called shivering. In other words, the hypothalamus acts as a thermostat. Under normal conditions, these various mechanisms, under the control of the hypothalamus, brainstem, and spinal cord, maintain body temperature within a degree or less of 98.6° F (37° C).

The stages of fever

The balanced maintenance of body temperature may be altered in a variety of ways, as when one has a fever, a common sign of a bacterial infection. Bacterial toxins alter the sensitivity of the hypothalamic cells so that they operate as if the thermostat were set at a higher level. For the body temperature to rise to this new level, the processes of heat retention are started. Cutaneous vessels constrict, and shunt vessels open, thus reducing the loss of heat by radiation. This may be accompanied by a subjective sense of coldness. Shivering may occur, thereby increasing heat production. The chills that may signify the onset of fever consist of shivering, plus the sensation of cold. Body temperature rises because more heat is produced than is lost. When the new level is reached, compensatory processes occur. More heat is lost, as evidenced by flushing (dilatation of cutaneous vessels). The fever does not drop, however, until the infection is over and hypothalamic cells begin to return to normal.

That the phenomena associated with fever are the result of altered hypothalamic sensitivity is shown by the fact that surgical operations in or near the hypothalamus may be followed by marked changes in body temperature, probably as a result of direct damage to hypothalamic cells. The postoperative rises or falls of temperature may be so extreme that the patient may die.

Neuroendocrinology. The hypothalamic–pituitary mechanisms comprise most of what is termed neuroendocrinology. It is through these relationships that the central nervous system can influence or regulate all endocrine activities. Among the important ones are intermediate metabolism, growth, electrolyte balance, reproduction, and the development of secondary sex characteristics. The activity of the hypothalamus in these mechanisms can be initiated by nerve impulses from many sources (for example, the special senses—sight, hearing, and so on) and by a variety of conditions in the blood, including temperature, osmotic pressure, and level of circulating hormones.

The hypothalamic nuclei that control the anterior lobe of the pituitary—the adenohypophysis—do so by synthesizing substances that enter the portal system and activate the glandular cells of the pituitary. These cells produce at least seven hormones: growth hormone, thyroid-stimulating hormone, adrenocorticotropin, prolactin, follicle-stimulating hormone, interstitial-cell-stimulating hormone, and melanocyte-stimulating hormone. These hormones, when released into the bloodstream, can activate their target organs or cells. The substances produced by the target organs, when released into the bloodstream, can, in addition to their primary actions, inhibit the pituitary. This is a feedback mechanism that provides regulating control. Moreover, some of the hypothalamic cells are also directly sensitive to, and regulated by, circulating substances from target organs, in particular, ovarian hormones.

Destruction of the hypothalamus, like destruction of the pituitary, produces profound endocrinological changes; these include changes in the skeleton, the sex glands, and metabolism.

Antidiuretic hormone and oxytocin

The two hypothalamic nuclei from which axons extend to the neurohypophysis produce the antidiuretic hormone and oxytocin, the antidiuretic chiefly from the supra-optic nuclei and oxytocin chiefly from the paraventricular.

The antidiuretic hormone is also known as vasopressin ("vessel-constricting substance"), although it produces a rise in blood pressure through activation of vascular smooth muscle only when given in high doses. The chief function of this hormone is to act on the tubules of the kidney in such a way as to decrease production of urine. It thus constitutes an important means of conserving body water. It is of interest that the cells of the supra-optic nucleus are directly sensitive to the ionic concentration and osmotic pressure of the blood. Changes in these features undoubtedly alter the activity of the nucleus. If the nucleus or its projections are destroyed, a disorder known as diabetes insipidus results. Because of the loss of antidiuretic hormone, large quantities of urine, amounting to many litres, are eliminated by the kidney each day. In order to maintain balance, large amounts of water must be drunk.

In mammals, including man, oxytocin acts chiefly on the smooth muscle of the uterus and also on certain smooth muscle cells of the mammary gland known as myo-epithelial cells. In the case of the uterus, the hormone is particularly effective during the later stages of pregnancy, and especially during parturition. It is often used as an aid in labour. In the mammary gland, it activates the special smooth muscle cells that cause the ejection of milk (the formation of milk is regulated by the hormone prolactin from the adenohypophysis). The release of oxytocin from the neurohypophysis is triggered by the act of suckling; the nerve impulses aroused in sensory endings in the nipple reach the spinal cord and ascend in the brainstem to the paraventricular nuclei of the hypothalamus.

Pineal body. Because of vast species differences, there is as yet no complete agreement on the functions of the pineal body, and many questions remain to be answered. Nevertheless, certain features have been clarified. In mammals, the gland is rich in noradrenaline, serotonin, and histamine. Activity of sympathetic postganglionic fibres, brought about by a decrease in the level of light, results in the release of melatonin. It is still uncertain whether this hormone inhibits or promotes the action of the melanin-stimulating hormone of the pituitary. Beyond this, however, melatonin seems to have an effect

upon the sex glands and, under certain conditions of overproduction may cause them to atrophy.

Autonomic functions of brainstem cells

Brainstem. The brainstem is concerned with the regulation of many specific visceral functions. For example, groups of nerve cells that are scattered throughout the pons and medulla oblongata form respiratory, vasomotor (blood-vessel-contracting or blood-vessel-dilating), and cardiac "centres." Still other groups of cells are concerned with intestinal movements, control of the pupillary diameter, swallowing, salivation, and so on. These centres are functionally restricted in the sense that the respiratory centres have no specific vasomotor or cardiac functions, although they may affect their activity. But the hypothalamus, which sends fibres to both, may control or activate them simultaneously, as when respiration and blood pressure increase during emotional states. The various centres also function on an integrated reflex level, as in the control of blood pressure (see below). Moreover, as in the case of respiration, which involves skeletal muscles, the centres are often coordinated with somatic mechanisms.

An idea of the importance of the cerebral cortex and hypothalamus as higher levels and of how the brainstem functions can be gained by studying a decerebrate animal—*i.e.*, one in which all of the brain above the level of the midbrain is destroyed or removed or isolated by a cut through the midbrain. Such an animal, even a human being, may still live. Breathing is carried on, and blood pressure is maintained reflexly at a fairly normal level. The animal, however, is unconscious, cannot eat, maintains body temperature poorly, if at all, and shows abnormal skeletal and autonomic reflexes.

Control of blood pressure. In the walls of the bifurcation (branching) of the common carotid arteries in the neck, and also in the wall of the arch of the aorta in the chest, there are nerve endings that are sensitive to changes in blood pressure. Those in the carotid arteries comprise what is termed the carotid sinus. These endings are all sensitive to the mechanical stimulation of distension caused by an increase in blood pressure. The nerve impulses resulting from such an increase ascend in visceral afferent (sensory) fibres in the vagus and glossopharyngeal nerves to the medulla oblongata and reach the cardiac and vasomotor centres. Connections of these centres to the vagus nerves cause motor impulses (parasympathetic) to descend in these nerves (via preganglionic fibres) to the heart, which they inhibit, thereby slowing its rate of contraction. At the same time, other nerve impulses from the centres descend to the spinal cord. Here they act upon cells whose preganglionic fibres in the sympathetic system control the diameter of the peripheral blood vessels, causing them to dilate. The resistance of these dilated blood vessels to the flow of blood decreases, and blood pressure drops. This may be likened to widening the nozzle of a garden hose, as a result of which the force of the stream of water decreases. With the decrease in blood pressure, fewer nerve impulses reach the brainstem centres, and the process lessens or reverses. The blood pressure then rises as peripheral vessels narrow and the heart rate increases. Under resting conditions, then, there is continually a slight shifting back and forth between rising and falling pressures. The net result is a finely balanced and integrated activity whereby blood pressure is reflexly maintained at an average level.

Slowing of the heart and dilatation of vessels

It is to be noted that, in a spinal animal (an animal in which the junction of the spinal cord and medulla oblongata is severed), neither blood pressure nor respiration can be reflexly maintained. Artificial respiration is necessary to maintain life.

Control of respiration. Respiration is also reflexly controlled. Although the muscles of respiration are skeletal in type, they are not completely under voluntary control; no one can voluntarily hold his breath to the point of asphyxiation. The control of respiration has been studied in experimental animals, but it is likely that the processes are generally similar in all mammals, including man.

Nerve cells scattered throughout the pons and medulla oblongata comprise the so-called respiratory centres. The axons of these cells project to the upper (cervical) part of the spinal cord, where they synapse with the nerve cells whose axons supply the other muscles of respiration.

Many of the above cells in the brainstem respiratory centres are intrinsically active; that is, they continuously maintain a repetitive discharge to the motor cells of the spinal cord. This intrinsic activity, however, is modified by sensory impulses and other factors so as to produce a rhythmic discharge, leading to a rhythmic pattern of breathing. The modulation is brought about as follows. When the lungs expand during inspiration, nerve endings in the lungs are stretched and stimulated. The nerve impulses that result ascend in the vagus nerves to the active cells, which they inhibit. In addition, a small mass of tissue called the carotid body, at the bifurcation of the common carotid arteries in the neck, contains nerve endings that are sensitive to changes in gaseous tensions of the blood flowing through the carotid bodies. The resulting nerve impulses ascend along sensory fibres in the glossopharyngeal nerve and reach the respiratory centres, where they have reflex connections and effects. For example, an increase in carbon dioxide tension in the blood reflexly results in an increase in the rate and depth of respiration, thereby blowing off the excess carbon dioxide. Finally, the cells in the respiratory centres themselves are directly sensitive to the concentration of carbon dioxide in the blood.

Action of the carotid body

What is of particular interest is that the reflex control of respiration is coordinated with the reflex control of blood pressure. For example, inspiration, which is characterized by an increase in chest volume and a decrease in pressure within the chest, is accompanied by an increased return of blood to the heart, necessitating an increase in cardiac output. The vascular reflexes initiated by the increased return of blood to the heart are coordinated with the respiratory reflexes to bring about the appropriate changes in the activity of the heart. This is an excellent example of the coordination of somatic, sympathetic, and parasympathetic functions to bring about an integrated reflex control of a vital function.

Control of movements of the alimentary canal. When food enters the mouth and is tasted and smelled, saliva begins to flow into the mouth. Its contained enzymes begin the digestion of certain constituents of the food. The salivation is a reflex act, the motor side of which is parasympathetic. When the food is swallowed, other reflexes are initiated. The impact of the food upon the walls of the pharynx stimulates nerve endings in the walls. The resulting nerve impulses reach the medulla oblongata over sensory fibres in the vagus and glossopharyngeal nerves. In the medulla oblongata, the impulses activate nerve cells comprising special centres, and these in turn activate the motor nuclei of the glossopharyngeal, vagus, and accessory nerves, as well as other brainstem centres. There is thus set into play a complex series of mostly involuntary actions, as follows.

The palate and the upper part of the pharynx constrict so as to prevent food from entering the nasal cavity. The upper part of the larynx, or voice box, constricts so as to prevent food from entering the passage to the lungs. The muscles just around the mass of chewed food contract, whereas those just ahead of it (below it) relax to receive it. This wave of contraction, preceded by a wave of relaxation, proceeds from the pharynx into the esophagus and thence into the stomach. The timing of this mechanism is intricate, especially in view of the fact that initially the contraction involves striated muscle fibres, but, from the middle of the esophagus on, smooth muscle is involved. Hence, the impulses descending in the vagus nerves to the smooth muscle in the esophagus now travel along parasympathetic preganglionic fibres. Also, impulses descend to the spinal cord to sympathetic fibres and thence to the esophagus, where they inhibit smooth muscle. This inhibition is evidenced as the wave of relaxation.

A wave of relaxation preceding a wave of contraction together comprise what is known as peristalsis. One of the fundamental types of movements of the alimentary canal, it propels contents, whether solid, semisolid, or liquid.

Definition of peristalsis

Additional complexities in the coordinated mechanisms

described briefly above are indicated by the fact that secretory phenomena involving digestive enzymes are usually taking place at the same time. Many of these are initiated or affected by parasympathetic activity.

Movements of the alimentary canal may be in the reverse direction, as in vomiting. The act of vomiting is brought about chiefly by the simultaneous contraction of the external abdominal muscles and the diaphragm, thereby increasing the intra-abdominal pressure and forcing movement of intestinal contents in the direction of the mouth. It is not known whether a special brainstem centre controls vomiting. Vomiting may be reflexly induced, however, by nerve impulses that result from irritation of the mucous membrane of the stomach, by putrid odours, and so forth. Also, there is a drug, apomorphine, that, when given, apparently acts directly on brainstem cells and induces vomiting. Finally, brainstem damage, for example, from an increase in intracranial pressure, may cause vomiting.

Cerebellum. Little can be added to what was said previously about the cerebellum. This important organ regulates and coordinates all muscular activity, and its influence extends to virtually all autonomic functions. Presumably, the modulation is carried out by way of cerebellar nerve fibres that extend to the diencephalon, brainstem, and spinal cord.

Sympathetic and parasympathetic outflow. The organization, arrangement, and connections of this autonomic outflow have already been described, and some of the specific functions are listed in Table 3. What remains is to illustrate how local and restricted functions can occur, as exemplified by certain reflexes and by a consideration of the autonomic functions of an isolated spinal cord. A common brainstem reflex is parasympathetic-mediated salivation in response to the taste or smell of food. Another example relates to alteration of the diameter of the pupil. If a light is flashed in the eye, the pupils of both eyes narrow, thereby restricting the amount of light that can enter. The nerve impulses that are initiated in the retina by the light travel along fibres of the optic nerve. Some of these fibres enter the brainstem; there, after several synaptic connections, impulses reach the cells from which parasympathetic preganglionic fibres in the third cranial (oculomotor) nerves originate. Impulses over these fibres then reach and activate the constrictor muscle of the iris (the second, or postganglionic, neuron is in the ciliary ganglion). The converse, a widening of the pupils, may result from inhibition of this parasympathetic pathway but may also be due to an activation of the dilator muscle of the iris by the sympathetic system. The cells from which the preganglionic sympathetic fibres originate are in the upper part of the thoracic cord; the postganglionic fibres originate in the superior cervical sympathetic ganglion and ascend along the carotid arteries and their branches to the eye.

Response of pupils to light

Local autonomic reflexes in the spinal cord are clearly demonstrable when the spinal cord is severed (the transection must be at the lower cervical segments, or lower, if respiration is to continue and life be maintained). A study of such injuries in man as well as in lower animals has given some insight into the autonomic functions of the spinal cord. As has been mentioned before, they are characterized by their restricted nature.

Transection of the spinal cord is immediately followed by spinal shock during which all functions, including the autonomic, are depressed. As the spinal shock begins to disappear, evidences of sympathetic activity begin to reappear. Cutaneous blood vessels begin to show reflex reactions, and sweating may be reflexly induced. After full recovery from spinal shock, stimulation of the skin may be followed by mass reflexes accompanied by widespread reflex sweating. Blood pressure is unstable and is poorly controlled reflexly, and body temperature in the affected parts is also poorly maintained.

Of particular importance are autonomic activities related to the bowels and the urinary bladder. The rectum and large intestine may gradually develop reflex emptying mechanisms that, in part, involve parasympathetic fibres from the sacral portion of the spinal cord. The sensory stimulus for evacuation seems to be distension of the rectum. The resulting nerve impulses reach motor cells and parasympathetic cells in the sacral portion of the spinal cord. The result is a contraction of the sigmoid colon and the rectum and relaxation of the sphincters of the anal canal.

The urinary bladder

The urinary bladder may function almost automatically and become what is called a "cord bladder." It tends to fill and empty spontaneously. An increase in pressure in the bladder as urine flows into it from the kidneys stimulates nerve endings in the bladder wall. The resulting nerve impulses reach parasympathetic cells in the sacral portion of the spinal cord. The nerve impulses leave over preganglionic fibres and reach the bladder wall, where the ganglion cells are located. The postganglionic fibres activate the smooth muscle of the bladder. At the same time the urethras contract, as do the external abdominal muscles; these increase the intra-abdominal pressure, resulting in emptying of the bladder.

In a man whose spinal cord has been severed at its junction with the medulla oblongata, sexual functions are also interfered with. Psychic stimuli are unable to influence sexual activity. There appear to be mechanisms in the lumbosacral part of the cord, however, that, especially in males, can produce integrated sexual activity, not only after direct stimulation of the external genitalia but frequently after any type of cutaneous stimulation.

The peripherally located autonomic ganglia are still more restricted in outflow and function. Moreover, so far as is known, no reflex connections of any physiological significance are established within them, although their complex structure certainly indicates that they have more than a simple relay function (see previous discussion), both in terms of preganglionic–postganglionic ratios and distribution and in terms of transmitter mechanisms.

If the preganglionic fibres to a specific ganglion are cut, the structures supplied by that ganglion are effectively denervated. However, these structures may not show any altered sensitivity. If a ganglion is removed, or its rami (branches) to peripheral nerves cut, the denervation is likewise complete (postganglionic sympathectomy). In this instance, however, the structures supplied may show altered sensitivity to circulating substances.

If a peripheral nerve is cut, postganglionic fibres in it are severed, just as in the case of removal of a ganglion or its rami. Depending upon the type of nerve, there will be sensory deficits, paralyses or weaknesses, and autonomic losses. The effects of loss of autonomic function include vasodilatation (evidenced by a warm skin) and loss of sweating (evidenced by a dry skin) in the area of distribution of the peripheral nerve.

Finally, it must be emphasized that, with an isolated spinal cord, the sympathetic supply to the adrenal glands is intact and that widespread reflex autonomic responses may in part bring about the release of adrenaline and in part be accentuated by such release.

PATHOPHYSIOLOGY

These few paragraphs are intended merely to provide a few examples of some pathological changes, autonomic disorders, and clinical applications. It has already been pointed out that the autonomic nervous system is involved in many complex, integrated physiological and behavioral activities. Consequently, its malfunctioning is a factor in many psychosomatic diseases and disorders. Of particular importance are those involving the hypothalamus, pituitary, and cerebral cortex. Some of the major implications have already been discussed.

The autonomic nervous system is frequently involved in surgical procedures. With the exception of vagotomy (involving severing of parasympathetic fibres) in the treatment of gastric ulcer, most surgery of the autonomic nervous system involves the sympathetic outflow. Modern surgical procedures include those aimed at sympathetic denervation of a limb with a vascular disease, in order to improve the circulation of that limb. They include procedures for treating certain kinds of hypertensions by denervating abdominal viscera, especially the

Surgery to improve circulation

kidney. Other procedures include denervation of the carotid sinus for the relief of fainting and other symptoms of a suddenly falling blood pressure due to a sensitive sinus. Finally, there are many procedures aimed at relieving pain; for example, sympathetic denervation of a limb for phantom limb pain and other painful disorders, sympathectomy for the relief of heart pain, splanchnic-nerve section for the relief of the pain of acute pancreatitis, and section of sympathetic plexuses for the relief of dysmenorrhea (painful menstruation). The rationale of treating pain by surgery on the autonomic nervous system is often not clear, nor are the procedures always successful. When they are, the relief may be temporary. Sometimes a procedure for relieving pain is effective because it cuts sensory fibres that happen to be travelling in or through an autonomic nerve or ganglion. This does not explain the good results sometimes obtained where sensory fibres may not be involved, however. For example, sympathetic block or denervation may relieve phantom limb pain. Patients who have had a limb amputated usually feel subsequently as if that limb were still present. Apparently, the cut ends of the nerves in the stump continue to be stimulated. The resulting nerve impulses arrive at the brain, where they are interpreted as if they were coming from the limb that the nerves previously supplied. Sometimes these sensations are painful. In such instances, the nerve impulses may reflexly activate sympathetic cells in the spinal cord. The autonomic impulses arriving at the vessels in the stump may produce vascular changes that further activate the cut ends of the nerves, or these autonomic impulses may directly stimulate sensory fibres adjacent to them in the stump. Whatever the explanation (and similar explanations have been offered for other painful conditions), stopping the autonomic impulses by sympathetic denervation may relieve or eliminate the pain.

Degeneration and regeneration of nerve fibres

When peripheral autonomic fibres, whether preganglionic or postganglionic, are cut, they degenerate. Provided that their cells of origin remain intact, regrowth—that is, regeneration—usually takes place, beginning at the cut ends. As a result, function may be restored, often completely in the case of postganglionic-fibre section. Moreover, it is difficult to remove all of the postganglionic fibres to an organ or a region, and, in addition, extraordinary precautions may have to be taken to prevent regeneration of postganglionic fibres. In attempting to sever sympathetic nerves to a region, such as a limb, surgeons generally use a combination of preganglionic-fibre section (cutting) and ganglionectomy—that is, removal of appropriate autonomic ganglia. Even then, denervation may not be complete because impulses may still reach the region through accessory ganglia unaffected by the surgical procedure.

One of the significant effects of denervation exhibited by certain effectors, especially vascular smooth muscle, is denervation sensitivity: the effector becomes sensitive to circulating or directly applied transmitters and drugs. For example, certain vascular disorders in which vasoconstriction causes pale, cold, aching fingers and hands may be temporarily relieved by sympathectomy. Unfortunately, these vessels may become so sensitive to circulating adrenaline that they again constrict, sometimes to a greater degree than before.

It is of interest that transplanted hearts are completely denervated. The conducting system of the transplanted heart, however, maintains relatively normal rhythmic activity. This points up what has been mentioned before, that many organs possess intrinsic activity that can carry on without an autonomic supply. There is some evidence that, in experimental animals, transplanted hearts may be re-innervated, but as yet there is not good evidence that this occurs in man.

In general, it may be said that the autonomic system is involved in a wide variety of disorders, including those of emotion and behaviour, and that understanding of its role, limited as that may be, depends upon understanding of the normal structure and function of the autonomic nervous system, an understanding whose enhancement depends upon research at all levels. (E.D.G.)

BIBLIOGRAPHY. CHARLES R. NOBACK, *The Human Nervous System* (1967), a general text with an account of the development of the nervous system and sensory receptors; RAYMOND C. TRUEX and MALCOLM B. CARPENTER, *Human Neuroanatomy*, 6th ed. (1969), includes a classic account of the development of the nervous system and sensory receptors; JAN LANGMAN, *Medical Embryology*, 2nd ed. (1969), for a summary of nervous system development; WILFRED M. COPENHAVER, RICHARD P. BUNGE, and MARY B. BUNGE, *Bailey's Textbook of Histology*, 16th ed. (1971); and WILLIAM BLOOM and DON W. FAWCETT, *A Textbook of Histology*, 9th ed. (1968), both contain summaries of sensory receptors; ARTHUR C. GUYTON, *Structure and Function of the Nervous System* (1972), an account of functional aspects of sensory receptors.

Central nervous system: LEONARD E. BAYLISS, *Principles of General Physiology*, 5th ed., vol. 2, pp. 492–533 (1960), a clear description of the development and functioning of the central nervous system, particularly concerned with the properties of the neurone, spinal reflexes, nerve excitation and inhibition, and reciprocal innervation; JOSE M.R. DELGADO, *Physical Control of the Mind: Toward a Psychocivilized Society* (1969), studies by electrical stimulations and recordings of behaviour from electrodes implanted at many levels within the brains of chimpanzees; JOHN F. FULTON, *Physiology of the Nervous System*, rev. ed. (1939), a detailed text on the functions of the nervous system of particular aid to those interested in the study of clinical medicine; HENRY GRAY, *Anatomy of the Human Body*, 28th ed. by CHARLES MAYO GOSS (1966), contains a superb description of the brain and spinal cord; IVAN P. PAVLOV, *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (1927, reprinted 1960), the classic experiments and studies of cerebral function in response to signals and reflex behaviour as carried out in dogs and their application to man.

Peripheral nervous system: ALF BRODAL, *Nevro-anatomi* (1943; Eng. trans., *Neurological Anatomy in Relation to Clinical Medicine*, 2nd ed., 1969), an authoritative systematic approach to understanding the form and function of the nervous system; A. DURWARD. "The Peripheral Nervous System," in *Cunningham's Textbook of Anatomy*, 10th ed. by G.J. ROMANES (1964), a comprehensive systematic text of human anatomy; WILFRED HARRIS, *The Morphology of the Brachial Plexus* (1939), a comparative study of the brachial plexus in vertebrates; WEBB HAYMAKER and BARNES WOODHALL, *Peripheral Nerve Injuries: Principles of Diagnosis*, 2nd ed. (1953), a text explaining the basic considerations associated with peripheral nerve injuries; EARL L. HOUSE and BEN PANKSY, *A Functional Approach to Neuroanatomy*, 2nd ed. (1967), a simplified approach to neuroanatomy; FREDERICK A. METTLER, *Neuroanatomy*, 2nd ed. (1948), a succinct summary of the cranial and spinal nerves; STEPHEN W. RANSON and SAM L. CLARK, *The Anatomy of the Nervous System*, 10th ed. (1959), a systematic consideration of the functional components of the cranial and spinal nerves; see also CHARLES R. NOBACK, and RAYMOND C. TRUEX and MALCOLM B. CARPENTER (both *op. cit.*).

Autonomic nervous system: Detailed discussion of the autonomic nervous system, as well as recent and historical reviews, include G.A.G. MITCHELL, *Anatomy of the Autonomic Nervous System* (1953); ERNEST D. GARDNER, DONALD J. GRAY, and RONAN O'RAHILLY, *Anatomy*, 3rd ed. (1969); JOSEPH PICK, *The Autonomic Nervous System* (1970); and OTTO APPENZELLER, *The Autonomic Nervous System* (1970). General physiological features are presented in ARTHUR C. GUYTON, *Textbook of Medical Physiology*, 4th ed. (1971); pharmacological features in LOUIS S. GOODMAN and ALFRED Z. GILMAN (eds.), *Pharmacological Basis of Therapeutics*, 4th ed. (1970); certain comparative aspects in C.U. ARIENS KAPPERS, G. CARL HUBER, and ELIZABETH C. CROSBY, *The Comparative Anatomy of the Nervous System of Vertebrates, Including Man*, 2 vol. (1936, reprinted in 3 vol., 1960); MALCOLM E. MILLER, GEORGE C. CHRISTENSEN, and HOWARD E. EVANS, *Anatomy of the Dog* (1964); MELVIN J. SWENSON (ed.), *Dukes' Physiology of Domestic Animals*, 8th ed. (1970); and clinical features and applications in JAMES C. WHITE, REGINALD H. SMITHWICK, and FIORINDO A. SIMEONE, *The Autonomic Nervous System*, 3rd ed. (1952); P.A.G. MONRO, *Sympathectomy* (1959); E. STEPHENS GURDJIAN and L. MURRAY THOMAS, *Operative Neurosurgery*, 3rd ed. (1970); JOSEPH PICK, *The Autonomic Nervous System* (1970); an overview of properties is provided by ERNEST D. GARDNER, *Fundamentals of Neurology*, 5th ed. (1968); and more detailed neuroanatomical features by RAYMOND C. TRUEX and MALCOLM B. CARPENTER, *Strong and Elwyn's Human Neuroanatomy*, 5th ed. (1964); J.P. SCHADE and DONALD H. FORD, *Basic Neurology* (1965); and ALF BRODAL (*op. cit.*).

(C.R.N./M.B.C./E.D.G./Ed.)

Nervous System Diseases

Diseases of the human nervous system underlie a remarkably broad array of symptoms of disorders, ranging from the subtlest disturbances of personality through such major signs as gross crippling, blindness, and violent and suspicious behaviour; all too often the result is the premature death of the sufferer (*e.g.*, through heart failure of neural origin). The discussion to follow assumes some understanding of the basic arrangement and function of such structures as the brain and spinal cord as given in NERVOUS SYSTEM, HUMAN; see also NERVES AND NERVOUS SYSTEM. In medical practice the diagnosis and treatment of nervous system diseases hinge on a carefully conducted, systematic neurological examination, which itself rests on a thorough and detailed study of the individual's medical history.

This article is divided into the following sections:

The neurological examination
- Tests of the cranial nerves
- Muscle examination
- Testing reflexes
- Other sensory tests
- Laboratory and other special tests

Neurological manifestations secondary to other diseases
- Endocrine glands
- Other organ systems

Neurochemical disorders
- Proteins
- Fats and other lipids
- Carbohydrates

Developmental defects
Diseases of peripheral nerves
- Radiculitis
- Neuritis
- Neuralgia

Diseases of the spinal cord
- Vascular lesions
- Tumours
- Syringomyelia

Diseases of the autonomic nervous system
Diseases of the brainstem and cranial nerves
- Brainstem
- Cranial nerves

Motor system disorders
- Chorea
- Paralysis agitans (Parkinson's syndrome)
- Dystonia musculorum deformans
- Spasmodic torticollis (wryneck)

Diseases of the cerebellum
Intracranial tumours
Diseases of the cerebrum
- Cerebral vascular accident
- Intracranial aneurysms and primary subarachnoid hemorrhage
- Presenile dementia (Alzheimer's and Pick's diseases)
- Degenerative diseases
- Down's syndrome (mongolian idiocy)
- Cerebral palsy (Little's disease)

Infections
- Bacterial infections
- Virus infections
- Other infections

Demyelinating disorders
- Myelinoclastic (sheath-destructive) diseases
- Dysmyelination disease

Craniocerebral trauma
- Complications of head injuries
- Sequelae (residual symptoms) of head injuries

Epilepsy (recurrent convulsive seizures)
Toxins
- Bacterial toxins
- Metallic poisoning

Neuromuscular disorders
- Amyotrophic lateral sclerosis
- Progressive muscular dystrophy
- Myotonia congenita
- Myotonic muscular dystrophy
- Myasthenia gravis
- Familial periodic paralysis

THE NEUROLOGICAL EXAMINATION

Examination of the nervous system is routinely preceded by a thoroughgoing general physical examination that includes measures of vital signs (*e.g.*, temperature, pulse rate, and blood pressure) and study of the heart, lungs, abdomen, skeletal structure, skin, and other organs. During the course of history taking and general physical examination, valuable information will be elicited with regard to sensory and muscular function and psychiatric condition. The status of the muscular system, for example, may be observed as the person walks into the physician's office or by noting spontaneous movements of his face, trunk, and extremities as he lies in bed.

The general appearance and cooperation of the individual (*e.g.*, neatness of clothing, evidence of personal hygiene, facial expressions, tone of voice, posture, presence of tics or other abnormal movements) provide evidence of contact with the environment and of possible neural disease. Signs of irritability, suspiciousness, hostility, suicidal depression, inattentiveness, or indifference are especially significant. Poverty of speech may signal depression; a continuous flow of words with punning or flight of ideas can mean manic excitement; senility is associated with repetitive and garrulous conversation, while speech is incoherent and rambling in delirium; strange new words (neologisms) or high-sounding, apparently meaningless sentences often are observable in schizophrenia. There also may be gross evidence of such psychiatric symptoms as hallucinations, delusions, obsessive ideas, and compulsions (see PSYCHONEUROSES; PSYCHOSES).

Psychiatric symptoms

Remote memory is readily assessed by the person's ability to give an account of his life (as known from other sources, such as relatives) and of his recorded illnesses. Recent memory is tested by asking him about events of the last few days, and immediate recall is gauged by reading a short sentence and asking him to repeat it. Somewhat longer retention is indicated if he can successfully recall after five minutes a list of a few common objects, such as chair or knife, and a fictitious address (see also MEMORY, ABNORMALITIES OF).

Difficulties in speech and other language skills often can signal disease of the nervous system. Dysarthria (dysphonia), for example, in which there is impaired control of the muscles of phonation, can arise from damage (lesions) in several parts of the brain and associated nerves (see SPEECH, PHYSIOLOGY OF). Questions and commands designed to assess the person's ability to understand spoken words should be fairly complicated, because even a person with extensive brain lesions may still be able to respond well to simple commands. People who are unable to understand spoken words (auditory agnosia) yet still have the capacity to talk may be quite voluble but unable to monitor what they are saying. They are likely to become angry when they cannot make themselves understood. Lesions involving a very limited portion of the brain cortex (Wernicke's area) are associated with trouble in understanding spoken words, while in other parts of the cortex (angular gyrus of the inferior parietal lobe) the lesions may cause disturbances in understanding written words.

Aphasia and apraxia

In what is called motor aphasia, the patient is hard put to express himself and may respond to questions only by repetition of a few of the words he hears. This is usually accompanied by loss of ability to read (alexia) and is associated with cortical lesions in the inferior frontal gyrus of the dominant cerebral hemisphere (in more than 95 percent of people, on the left half of the brain). A less severe disturbance in motor speech is anomia, difficulty in recalling nouns, particularly the names of objects. Anomia characterizes the disorder called amnesic aphasia and arises from lesions in the temporal lobe (on the side of the head) in the dominant hemisphere. Amnesic aphasia should be differentiated from agnosia; in visual agnosia, for example, the person cannot recognize familiar objects or words that he sees. The latter disturbance may accompany lesions of the occipital lobe at the back of the brain.

Apraxia describes the inability to perform voluntary, purposeful movement of parts of the body that are not paralyzed; automatic or associated movements may still be observed. Apraxia of the tongue, for example, may be manifested by an inability to protrude it on command. The sufferer may, however, freely lick fluids or food placed on the lips. Apraxia is usually associated with le-

sions of a limited portion (the supramarginal gyrus) of the parietal lobe (toward the top of the head) in the dominant hemisphere and is not infrequently accompanied by aphasia.

Inspection of the head

The head should be inspected carefully for any skin lesions and palpated for tender spots, swellings, or local overgrowth of bones. Infants with increased fluid pressure within the brain (hydrocephaly) may show bulging soft spots (fontanels) and enlargement of the skull. The blood vessels in the neck are felt and a stethoscope applied to listen for unusual noises there and in other parts of the head. Voluntary and examiner-manipulated (passive) movements of the neck in several directions are made to reveal any weakness or limitation of motion.

Tests of the cranial nerves. The cranial nerves are the 12 pairs of nerves connected directly with the brainstem; most nerves enter and leave by the spinal cord.

Olfactory. In examining olfactory (smell) function, the person is asked to sniff strongly alternately through each nostril, since there are two separate olfactory nerves. He is asked to identify by smell such things as cigarettes, coffee, vanilla, peppermint, and menthol.

Optic. Clues to optic nerve function derive from examination of the eye, including determination of the pressure in the eyeballs, testing of visual acuity and visual fields, and inspection of the retina of the eye by a device called an ophthalmoscope. Of special neurological interest are drooping lids, protrusion of either eyeball (exophthalmos), and inequalities in the size and shape of the pupils (see EYE AND VISION, HUMAN).

Oculomotor, trochlear, and abducens. These three cranial nerves work together in the control of eye movements. The individual is asked to follow the movement of a small light or the examiner's finger upward, downward,

Eyeball movements

and to either side and to report any experience of double vision (diplopia). The position of the eyeballs at rest, any defect in movement, and their position when diplopia occurs can indicate which of these nerves may be injured.

Trigeminal. The functions of the trigeminal nerve are tested by stimulation (*e.g.*, with a pin) on the upper and lower jaw and around the eye. Normal trigeminal function serves the corneal reflex; *i.e.*, contraction of the lids elicited by touching the front of the eyeball (*e.g.*, with a wisp of cotton) as the person looks upward. The state of the chewing muscles is tested by palpation and by observing jaw movements.

Facial. Both taste and muscle function are served by the facial nerve. Asymmetry or lack of mobility of facial expression are signs of disorder. Voluntary movements of the muscles of the face are tested by asking the person to wrinkle his forehead and to forcibly close his eyes and by having him show his teeth, pucker his lips, whistle, or smile. Taste (gustatory) function is tested usually with sweet, sour, and salty solutions to selected parts of the tongue. Lesions of the facial nerve are accompanied by impairment of taste on the front two-thirds of the tongue and lesions of the glossopharyngeal nerve (see below *Glossopharyngeal*) by impairment on the posterior one-third.

Vestibulocochlear. This cranial nerve serves both hearing and equilibrium. Hearing is tested with normal and whispered voice or with a ticking watch held close to one ear while the other is blocked. Further testing can be made with a special device (audiometer). Neural disorder is indicated if the person has trouble hearing even when such structures as the eardrum are bypassed (*e.g.*, by placing the stem of a vibrating tuning fork on the mastoid bone behind the ear).

Unilateral (one-sided) disturbance of the vestibular branch of the nerve is manifested by such signs as dizziness, falling to one side, or even complete inability to walk. Less severe vestibular disturbances are indicated by such behaviour as difficulty in attempting to raise one arm overhead and accurately return it to the same spot on a horizontal rod. Oscillatory eye movements (nystagmus) are observed in vestibular disturbances because of an intimate connection with parts of the brainstem concerned with eye movements. The vestibular nerve can be further tested by syringing the ears with hot and cold water or air (caloric test) or by spinning the person in a rotating chair with his head held in various positions to selectively stimulate different semicircular canals in the inner ear (Bárány test; see also EAR AND HEARING, HUMAN).

Glossopharyngeal. Functions of the glossopharyngeal nerve are tested by observing the movement and sensibility of the pharynx during examination of the mouth and throat. The taste function of the nerve is assessed during the examination for the facial nerve (see above *Facial*).

Vagus. Some impairments of the vagus nerve are evidenced by difficulties in talking and swallowing. Paralysis on one side is indicated by movement of the palate to the opposite side. Abnormal movements of the vocal cords, also signs of vagus difficulties, are observed with a special laryngeal mirror.

Spinal accessory. The function of the spinal-accessory nerve is tested by noting the state of nutrition of specific neck muscles (sternomastoid and trapezius muscles) and by testing their strength in rotation of the head and in shrugging the shoulders.

Hypoglossal. The hypoglossal nerve is tested by noting the state of the muscles of the tongue, especially being alert for trembling (fibrillations) or other unusual movements. A tendency to protrude the tongue to one side indicates weakness of muscles in that half of the tongue.

Muscle examination. Careful observation of the person as he walks may reveal abnormalities of posture and gait that reflect nervous system disease.

Muscular atrophy (shrinking) and hypertrophy (overdevelopment) may be determined by simple palpation and by measurement. The circumferences of the same portions of the limbs on opposite sides of the body are compared. The presence or absence of fibrillary twitchings are sought, especially when there is generalized or localized muscular atrophy. Limbs are flexed and extended at the joints by the examiner to see if the degree of resting muscular contraction is abnormally reduced (hypotonus) or increased (spasticity or rigidity).

Coordination is tested by asking the person to extend his arms to the side and then to bring the tips of the index fingers together or to place them one at a time on the nose, to repeatedly flex and extend the fingers, or to rotate the forearm. He may be instructed to touch the examiner's finger with a big toe or to place one heel on the opposite knee and run the heel down the leg (heel–shin test).

Finger-to-nose test

Testing reflexes. A number of reflexes have been found to be of value in determining the site of lesions in the nervous system. Tendon reflexes in the extremities are elicited by placing a finger on the tendon of the muscle (*e.g.*, just below the kneecap) and tapping the tendon briskly with a soft hammer. When reflexes are exaggerated, a single stimulus may set off a continuous repetition (clonus) of the reflex. Clonus is most common in muscles of the wrist, knee, and ankle.

Abdominal reflexes are cutaneous (elicited by stimulating the skin). The person lies flat on the back with muscles relaxed; a brisk stroke with a pencil or dulled pin is made in each quadrant of the abdomen, starting from the edge and directing the stroke toward the umbilicus. The normal response is contraction of the abdominal muscles and a movement of the umbilicus toward the stroking instrument. The sole (plantar part) of the foot is also stroked, starting from the heel and across to the base of the big toe. The normal response consists of flexion of all of the toes, accompanied by a slight flexion and inversion of the tip of the foot. In all but the youngest infants, an extensor plantar response (Babinski reflex) is abnormal; it consists of extension of the big toe accompanied by flexion and fanning out of the other toes.

Other sensory tests. Sensitivity to light touch is usually tested by the stroke of a fine, soft brush or a wisp of cotton. All areas of the skin of the trunk, extremities, and face are tested, and comparison is made of the intensity of the stimuli needed to elicit a response in various areas.

Pain sensitivity of the skin is tested by light pricks of a needle or pin. All areas of skin are tested, and comparisons of areas are made as with light touch. Ability to

localize the point stimulated is determined by asking the person to indicate (without having looked) the exact place just pricked.

Temperature sensation may be tested by the application of test tubes containing hot and cold water. Comparison of the ability to recognize the difference between the hot and cold stimuli is made in all parts of the body.

Two-point cutaneous discrimination is tested by determining the shortest distance between the two points of an artist's compass at which the points can definitely be distinguished as separate by touch alone.

Proprioception, the sensing of changes in the muscles and tendons, is evinced by ability to recognize passive movements of the joints and to detect vibrations produced by a tuning fork placed on the bony prominences and on the skin.

Laboratory and other special tests. Major portions of the nervous system can be observed or assessed only by the use of such methods as chemical analysis, X-ray examination, and electroencephalography. Chemical examination of the blood, for example, to seek the presence of such drugs as bromides, barbiturates, and alcohol, is made in undiagnosed cases of coma. Similar studies are made of the urine and other body fluids, including cerebrospinal fluid (ordinarily tapped from the spine with a long hypodermic needle). Thus, the cerebrospinal fluid may reveal bacteria that produce brain inflammation (meningitis or encephalitis).

Roentgenography (X-ray examination). Ordinary X-ray films of the skull and spine are of value in revealing abnormal conditions such as tumours. To enhance the contrast in the films it is often helpful to introduce some foreign material (a contrast medium) into the tissues before they are X-rayed. Thus, the cavities (ventricles) of the brain are easier to visualize in roentgenograms after the injection of air as a contrast medium. The introduction of air into the head, however, is accompanied by definite risk to life. X-ray visualization of blood vessels (angiography) by the injection of dyes as contrast media (*e.g.*, into the carotid artery in the neck, which delivers blood to the brain) has the advantage over air injection in that it is a less dangerous procedure with a low mortality rate, but transient untoward symptoms occur in about 10 percent of the cases. Angiography is of great value in the localization of tumours of the brain and in revealing abnormalities of the blood vessels. Similarly, the introduction of contrast media (oils or air) is of value in localizing lesions (*e.g.*, ruptured disks) that compress the spinal cord or nerve roots.

Injection of air into the brain

Echoencephalography. In echoencephalography, high-frequency vibrations (ultrasound) are passed through the head, the returning echoes being recorded. Echoencephalography is an easy, rapid, and harmless procedure with high accuracy in the diagnosis of a brain tumour, blood clot, or other space-occupying lesion in the skull.

Radioactive scanning. Localization of lesions in the brain by means of their differential tendencies to incorporate radioactive substances is a valuable diagnostic aid. Radioactive arsenic, mercury, iodine, or other substances are injected intravenously, and scanning of the skull is accomplished with devices related to Geiger counters. The nature of any brain lesion present can be inferred from the rapidity of appearance and disappearance of abnormal concentrations of radioactivity. This relatively harmless procedure commonly can be used in lieu of the more troublesome angiography or air studies.

Rheoencephalography. Rheoencephalography is the study of variations in the electrical conductivity of cranial tissues to high-frequency alternating currents. Such electrical changes are found extremely useful in the detection of disturbances in the circulation of blood within the skull.

Electroencephalography. Recording of the electrical activity of the brain can be performed with safety and comfort. The electroencephalogram (EEG) is of special diagnostic value in cases of head injury, brain tumours, cerebral infections, cerebral-blood-vessel lesions, degenerative diseases of the nervous system, and convulsive disorders (*e.g.*, epilepsy). Latent brain-wave abnormalities can be activated by deep breathing (hyperventilation), stimulation with flickering light, or with small doses of stimulant drugs.

Neuromuscular electrodiagnosis. Nerve-conduction velocity is indicated by how long it takes between stimulation of a nerve and the appearance of action potentials (electrical activity) in the muscles supplied by the nerve. Slowing of the conduction velocity below normal indicates damage to the nerve and suggests that the spinal cord is not responsible for any muscular paralysis present. Comparison of the conduction velocity of the nerve at various points may indicate the exact site of a lesion.

The electrical activity of muscles can be determined by the insertion of needle electrodes. There are characteristic changes in the activity when the nerves or muscles are damaged. With complete interruption of a nerve, there is no action potential from the paralyzed muscle. Subsequent return of activity indicates regeneration of the injured nerve. When the spinal cord is damaged (as in some forms of poliomyelitis), increased muscle activity is seen on insertion of the needle, and there are distinctive electrical discharges when the muscles are at rest.

Electrical activity of muscles

Biopsy. Removal of living tissue (biopsy) from nerves is of value in determining the presence and nature of neural lesions. Degeneration of the nerve's insulating material (myelin sheath) is found in most of the common varieties of nerve inflammation (polyneuritis). Inflammatory reaction around the nerves is seen in a number of disorders, including Hansen's disease (leprosy). For ethical reasons, biopsy of brain tissues is rarely made, except for tumours of the brain, in which case it is valuable in determining whether an attempt should be made to remove a tumour or to treat with radiation.

NEUROLOGICAL MANIFESTATIONS SECONDARY TO OTHER DISEASES

The nervous system is involved secondarily in diseases of other organ systems, particularly those of the endocrine glands, blood, liver, the connective tissues, and bone.

Endocrine glands. The functioning of the nervous system is affected to some degree by all of the endocrine (ductless) glands. Disturbances of consciousness and thinking as well as various sensory and motor symptoms may occur when these glands are overactive (hyperfunction) or underactive (hypofunction). As an example, diabetes insipidus (characterized by excessive amounts of urine and insatiable thirst) usually arises from an injury to the fibre tracts connecting the base of the brain (the hypothalmus) with the pituitary gland. A disorder called Cushing's syndrome is associated with oversecretion of a special portion (the basophilic cells) of the pituitary. The fully developed syndrome, most common in females, includes swollen face, obesity, overgrowth of facial and body hair, purple abdominal streaks (striae), high blood pressure, increase in the number of red cells in the blood, absence of menstruation, elevated blood sugar, muscular weakness and wasting, and softening of the bones of the skeleton. Mental symptoms are common. Most of the features of Cushing's syndrome also can be produced by the administration of pituitary or adrenal hormones.

Pituitary gland

Oversecretion of the portion of the pituitary that secretes the growth hormone (acidophilia cells) causes gigantism if the activity starts before puberty, and acromegaly when it occurs after this age.

Deficiency of the secretion of the thyroid gland (hypothyroidism) in infancy produces cretinism, in which there is retardation of physical and mental development, both of which may be severe. Hypothyroidism developing in late childhood or in adult life results in myxedema, characterized by lethargy, weakness, slowness of speech, swelling (edema) of the tissues, coarse, pale skin, dry, brittle hair, and excessive sensitivity to cold. Neurological complications include headache, cranial-nerve palsies (*e.g.*, bodily tremors), peripheral neuritis, psychotic episodes, and convulsions. Excessive secretion of the thyroid (hyperthyroidism) is associated with such neurological symptoms as tremor of the extended hand, infrequency of blinking, weakness of muscles, emotional instability, and even psychotic episodes.

Deficiency in the secretion of the calcium-regulating parathyroid hormone produces parathyroid tetany, characterized by numbness and cramps in the extremities, spasm of the fingers or toes, convulsive seizures, and psychotic episodes. Increased intracranial pressure with associated visual effects is one of the rarer neurological manifestations of the disease. The abnormal laboratory findings include decrease in blood and urinary calcium and various abnormalities (*e.g.*, slow waves) in the electroencephalogram. When there is an excess of the parathyroid hormone, thinning of the bones, formation of stones in the kidneys, and muscular weakness may occur.

Insulin, the hormone secreted by the pancreas, regulates sugar (glucose) metabolism. The nervous system depends on glucose for its energy, and loss of consciousness develops rapidly when the amount of that sugar in the blood falls below critical levels (hypoglycemia). An excess of glucose in the blood associated with a deficiency in the secretion of insulin (diabetes mellitus) may cause a variety of disturbances of the nervous system. Excessively high levels of blood glucose (hyperglycemia) usually are accompanied by confusion and coma. Involvement of the cerebral blood vessels may be followed by blood clots (thrombosis) with paralysis (*e.g.*, hemiplegia) or other signs appropriate to the location of the clot within the brain. Neuritis with or without involvement of the cranial nerves may also occur, probably related to a generalized diabetic disturbance of the nerves of as yet unknown nature.

Oversecretion of insulin (hyperinsulinism) is rare, although hypoglycemia may occur after an injection of too much insulin. Occasionally, it is found in people with tumours of the pancreas. The symptoms of severe hyperinsulinism are paroxysmal (*e.g.*, convulsions) and tend to occur at times when the blood sugar is expected to be low; *i.e.*, early in the morning or after fasting or exercise. Other symptoms attributable to hypoglycemic disturbance of the nervous system include lightheadedness, tremors, sweating, pallor, palpitations, headache, abdominal pain, nausea, vomiting, or ravenous hunger. There also may be tingling in the extremities, diplopia, or blurred vision, generalized weakness, transient hemiplegia, aphasia, confusion, abnormal behaviour, and coma (the latter usually preceded by convulsions).

Low blood sugar

The adrenal gland is composed of two portions, an outer cortex and an inner medulla: both elaborate a number of specific hormones that can affect the nervous system. Hypofunction of the adrenal cortex (Addison's disease) may be associated with such symptoms as weakness, fatigue, low blood pressure, low blood sugar, irritability, and anxiety. Hyperfunction of the adrenal cortex can produce Cushing's syndrome (see above) and, depending on which hormones are oversecreted, can lead to such sexual disturbances as premature puberty, enlargement of the genitals or feminization in the male, and virilism (*e.g.*, hairiness) in females. Oversecretion of the adrenal medulla can produce high blood pressure.

Neurological complications of the administration of contraceptive hormones that affect the ovaries are uncommon considering the number of women taking them. Symptoms of involvement of the nervous system are headaches of the migraine type, increased intracranial pressure, optic neuritis, diplopia, and cerebral vascular lesions; in most cases the symptoms subside with discontinuance of the use of the hormones.

Other organ systems. *Bones.* Osteitis deformans (Paget's disease) is characterized by bowing and irregular flattening of the bones. The skull and spine may show signs of overgrowth and softening, and involvement of the nervous system arises from compression of nervous tissue. Headaches, convulsions, and compression of cranial nerves are the most common symptoms. Paralysis (paraplegia) may occur as a result of compression of the spinal cord. Degenerative and arthritic changes (*e.g.*, cervical spondylosis) are common in the spine in middle-aged or elderly people. Symptoms that result from compression of the spinal nerves of the cord itself include pains and paresthesias (*e.g.*, pricking and tingling) in the neck and back extending to the extremities, stiffness of the spine, weakness of muscles, impairment of sensation, and loss of the reflexes innervated by the affected nerves.

Liver and blood. Incoordinated gait, tremor, dementia, paraplegia, and coma may occur with disease of the liver, and the ammonia content of the blood tends to be increased.

Indeed, damage to the nervous system may be a feature of various diseases affecting the blood. Among these are degeneration of the spinal cord associated with anemia and other neurological symptoms arising from thrombosis, as when the number of red cells is grossly elevated (*e.g.*, polycythemia vera). The nervous system may be the site of collections of tumour cells in leukemia, with resulting cranial nerve palsies, polyneuritis, hemiplegia, and other localized neurological symptoms. Degeneration of the spinal cord associated with primary anemia was formerly common but became quite rare after discovery of an effective treatment (which includes vitamin B_{12}) for the anemia.

Cancer of other organs. The nervous system may be affected by malignant tumours in the lungs or other organs without direct involvement of the nervous tissue. In addition, infection of the brain membranes (meningitis) and abscess of the brain with yeast and other fungi is likely to occur in cancer sufferers whose immune reactions are suppressed by the administration of anticancer drugs. Indirect effects of malignant tumours in the nervous system include damage in the brain and spinal cord (multifocal leukoencephalopathy), probably from invasion by viruses. Polyneuritis may occur as a result of the spreading of cancers (metastatic invasion) to the nerves or by an unknown toxic process.

NEUROCHEMICAL DISORDERS

A number of degenerative diseases of the nervous system have been shown to result from inborn metabolic defects. Many of these hereditary diseases are associated with inborn defects in the metabolism of proteins, fats, and other substances.

Proteins. A number of diseases are recognized in which there is a deficiency of a specific enzyme concerned with the metabolism of one or more amino acids (protein constituents). Injury to the nervous system results from the accumulation of abnormal metabolic products in the blood. In general, it can be said that these diseases are rare, the most common being phenylketonuria (see the Table), for which the estimated incidence in the U.S. is one in every 20,000 births. Onset of symptoms is usually in infancy, and fatalities or serious disabling symptoms are usually related to involvement of the nervous system. Mental retardation and convulsive seizures are the most common symptoms. Many of the affected individuals die in infancy, although this may be preventable through early treatment by feeding a diet low in the specific amino acid for which metabolism is defective.

Some Defects of Protein Metabolism

metabolic disease	amino acids involved	typical neurological symptoms
Phenylketonuria	phenylalanine	convulsions; mental retardation
Maple-syrup urine disease	valine; leucine; isoleucine	convulsions; muscular rigidity; rapidly fatal
Homocystinuria	methionine	mental retardation; convulsions; cerebral thromboses
Hartnup's disease	many excreted in excess in urine	periods of ataxia; psychotic episodes; mental deterioration
Oculo-cerebral-renal syndrome of Lowe	many excreted in excess in urine	convulsions; mental retardation; loss of reflexes; visual defects

Fats and other lipids. There are a number of diseases in which there is a disturbance of the metabolism of fatlike substances (lipids). In some, the nervous system is the primary or chief site of the disturbance (*e.g.*, Tay-Sach's disease), and in others it is affected as part of a general involvement of other organs; *e.g.*, the abdominal viscera (Gaucher's disease and Niemann-Pick disease) or the bones (Hand–Schüller–Christian disease).

Cerebromacular degeneration (Tay-Sachs disease). Also called amaurotic familial idiocy, this is a relatively

rare hereditary disorder characterized by loss of vision, convulsive seizures, paralysis, and dementia. There is degeneration of the nerve cells (neurons) in the central nervous system (brain and spinal cord) and in the retina, with deposition of lipoid (fatlike) pigments in the degenerated cells. The onset of symptoms is common in the first few months of life (infantile form), but it may be delayed to the third to fifth year of life (late infantile type) or to early or late childhood (juvenile form). Symptoms are similar in all three forms, but there is considerable variation in their severity. The infantile form is characterized by its occurrence in Jewish children, the presence of a cherry-red spot at the point of direct vision (the macula) in the retina and a rapidly downhill course. The late infantile and juvenile forms usually occur in non-Jewish families; there is pigmentary degeneration of the retina rather than a cherry-red spot in the macula. Severity is related to the age of onset; when symptoms first appear after the age of eight years, the degree of cerebral degeneration is less, and the length of life is greater. Nervous system symptoms similar to those of cerebromacular degeneration may occur in Niemann–Pick or Gaucher's disease.

Cranial xanthomatosis (Hand–Schüller–Christian disease). This arises from a disturbance of metabolism of a fatlike substance (cholesterol) with deposition of lipoid-containing cells in membranous bones, particularly the skull. The characteristic symptoms include defects in the bones (easily demonstrated by roentgenography), diabetes insipidus, and bulging eyes (exophthalmos). Rubbery masses of lipoid-containing cells in the dura (one of the membranes) under the frontal lobes and in the hypothalamic regions of the brain are responsible for the exophthalmos and the diabetes insipidus, as well as the retardation of growth that is present in about one-third of the cases. The disease is rare; males are affected twice as often as females. The onset of symptoms is usually in the first two decades of life, and the course, if unsuccessfully treated, is slow, progressive deterioration.

Related disorders include hereditary ataxic polyneuritis (Refsum's disease), a rare form of polyneuritis resulting from an inborn error of lipid metabolism. It is also characterized neurologically by pigmented, degenerated retina and by deafness. These and other symptoms have their onset in the first or second decades of life with slow progression and death before the age of 50 years, although special dietary restrictions (of foods containing phytol and phytanic acid) may have a beneficial effect.

Another example is the Bassen–Kornzweig syndrome, a rare lipid disease affecting chiefly the cerebellum of the brain; the inherited metabolic defect is associated with ataxia (incoordinated gait). Such neurological symptoms start in childhood and progress slowly to adult life, when they tend to become static.

Carbohydrates. *Galactosemia.* A rare inherited defect in the metabolism of the sugar galactose (related to lactose, or milk sugar) results in an intolerance to milk. The symptoms usually appear in the first week of life and include vomiting, diarrhea, jaundice, enlargement of the liver and spleen, accumulation of fluids (ascites) in the body cavities, and cataracts. When treatment is not instituted because of failure in recognizing the disease, there is psychomotor and mental retardation. There is an increase in the excretion of galactose, protein, and amino acids in the urine. All of these are reversible if milk and other sources of lactose and galactose are rigidly excluded from the diet; mental development then will be normal unless, of course, the brain already has been too seriously damaged.

Gargoylism (Hurler's disease). This is a rare metabolic disease in which there is a peculiar malformation of bones associated with mental deficiency. The symptoms are caused by deposition of carbohydrates called mucopolysaccharides (*e.g.*, chondroitin sulfate and heparitin sulfate) in the central nervous system, the retina, and other organs. The disease is manifest in the early months of life, with the child taking the appearance of a short-limbed dwarf with coarse facial features. While mental and physical development are retarded, the degree of mental deficiency is variable, and the disease is not incompatible with many years of life. Other disorders of mucopolysaccharide metabolism with variations of the clinical picture of gargoylism have been described.

Glycogen-storage disorders. Carbohydrate is normally stored in the liver in the form of a chemical called glycogen. Abnormal deposition of glycogen may occur in a number of congenital disorders, each probably representing a specific enzyme deficiency. The most common is von Gierke's disease (hepatorenal glycogenosis), in which the liver and kidneys are chiefly involved secondary to a deficiency of the enzyme glucose 6-phosphatase. In generalized glycogenosis (glycogen heart disease, or Pompe's disease), glycogen deposition is most marked in the heart and skeletal muscles. Nerve cells in the brain and autonomic nervous system may also be involved. Onset is in early infancy, and symptoms of skeletal muscle weakness and hypotonia (flabbiness) may suggest other muscle diseases. Death from heart failure usually occurs within the first two years.

DEVELOPMENTAL DEFECTS

Many defects of the nervous system may be congenital (present at birth). The damage may be limited to the brain and spinal cord, or the skull and spinal bones may be affected. Congenital defects in other organs may also be present (see BIRTH DEFECTS AND CONGENITAL DISORDERS).

Defects in the development of nervous tissue may be extensive, as in absence of the cerebral hemispheres (anencephaly), or may be limited to defects in the development of one or more brain structures, including the cerebellum, brainstem, and the spinal cord. Other defects include microcephaly ("pinhead"), porencephaly (abnormal cysts or cavities in the brain), and hydrocephalus ("water on the brain"). Defects in the development of the bony structures may occur with or without concomitant damage to nervous tissue.

Common forms of developmental defects of the central nervous system and its bony covering include spina bifida, in which there is failure of closure of the bony spine, and cranium bifidum, in which there is failure of closure of the skull. The Arnold–Chiari malformation is characterized by downward elongation of the brainstem and cerebellum into the spinal canal of the neck. Ocular hypertelorism, characterized by a facial defect in which there is an excessive distance between the eyes, is associated with mental retardation.

Causes of developmental defects

Not all of the causes of these and other related malformations are known. Genetic factors play a role; many of these developmental defects are associated with chromosomal abnormalities and have a tendency to run in families, but other factors are probably of greater importance. The nervous system of the developing fetus is especially vulnerable to lack of oxygen (anoxia), radiation, infectious and metabolic diseases in the mother, and her ingestion of drugs during critical stages of development of her unborn child. Serious defects in the development of the nervous system usually result in death within the uterus or shortly after birth. Minor defects (*e.g.*, in some lesser forms of spina bifida) may not be accompanied by any neurological symptoms, and moderately severe defects tend to produce a degree of disability related to the extent and site of the damage.

DISEASES OF PERIPHERAL NERVES

Peripheral nerves are those that extend beyond the central nervous system (*i.e.*, beyond the brain and spinal cord).

Radiculitis. This inflammation of the roots of the spinal nerves (*i.e.*, at their point of entry into the spinal cord) causes paralysis of the muscles and loss of sensation in the skin innervated by the root involved. Severe burning pains are present when injury is due to either compression or inflammation.

The nerve roots and plexuses (mingled fibres of two or more nerves) may be damaged at birth or by trauma (injury) after birth. Individual nerve roots may be injured by tumours, ruptured intervertebral discs, and ar-

thritic changes in the vertebrae. The posterior root and its relay centre (ganglion) are inflamed (herpes zoster) in specific virus infections.

Neuritis. Neuritis is the term generally used to denote inflammation of nerves from any cause. In the majority of the cases this is not caused by an infectious process, and the term neuropathy is preferred by some authors. Injury (trauma) is the most common cause of localized inflammation of a single nerve (mononeuritis), whereas toxic agents and metabolic disturbances usually cause diffuse involvement of many nerves (polyneuritis).

The symptoms that result from injury to a single nerve are related to the nature of the nerve affected. Involvement of motor nerves (those that carry impulses away from the central nervous system) causes paralysis, atrophy, and loss of reflexes in the muscles supplied by the given nerve. Loss of sensation in an area of tissue slightly less than that of the anatomical distribution of the nerve occurs with lesions of sensory nerves. Vasomotor disturbances (*e.g.*, circulatory difficulties) are common when a sensory or mixed (sensory and motor) type of nerve is damaged. Peripheral nerve injuries are common in war from gunshot and shrapnel wounds, and trauma (as in athletics and vehicular collisions) is also the most common cause of nerve injuries in civil life. Other causes include compression by tumour and trauma associated with the injection of drugs.

When the injury to a peripheral nerve is of minor nature, recovery usually takes place in the course of a few days or weeks. If the injury is severe or if there has been degeneration of the fibres (axons) of the nerve cells, recovery is delayed. Severed nerve ends can be brought together (anastomosed) by surgical suturing.

Polyneuritis

Polyneuritis (multiple peripheral neuritis) can be caused by a great number of toxic and metabolic disturbances, including vitamin deficiency, lead poisoning, leprosy, thyroid-gland disorders, and cancers. The actual cause of the degenerative changes in the peripheral nerves is unknown, but it is probable that both toxic and metabolic disturbances may interfere with the enzymes concerned with the nutrition of the nerves. The clinical symptoms are similar regardless of the cause of the polyneuritis, but there are variations in the distribution of paralysis and in the degree of the motor and sensory defect. Sufferers of mild polyneuritis complain of pains and paresthesias, but no weakness, sensory loss, or reflex disturbances occur; the disorder is termed moderate when the pains and paresthesias are accompanied by slight weakness of the upper or lower extremities and loss of deep reflexes; it is severe when there are complete paralysis and loss of sensation in the distal parts (ends) of all extremities and loss of all deep reflexes. The weakness usually is confined to the extremities, but, in severe cases, it may extend to involve the muscles of the trunk and the muscles supplied by the cranial nerves.

Paralysis of the facial muscles and of the pharyngeal muscles is common in infectious and diphtheritic polyneuritis. Sensory loss, when it occurs, usually is confined to a drop in or loss of sensitivity (anesthesia) of the distal parts of the extremities over the areas of skin commonly covered by a glove or stocking. Proprioceptive (muscle and joint) sensibility usually is more impaired than is cutaneous (skin) sensibility. Confusion, disorientation, amnesia, and other memory defects are occasionally seen in any form of polyneuritis but are most common in the alcoholic-vitamin-deficiency form (Korsakoff's psychosis) or that associated with pregnancy.

Complete recovery is the rule in most of the cases of polyneuritis, but it is slow when damage to the nerves is extensive, and there may be residual symptoms in the form of muscular wasting and sensory loss. Treatment of polyneuritis is accomplished by the removal of any toxic agent and the alleviation of any metabolic or deficiency state. Physiotherapy is most helpful in the convalescent stage.

Neuralgia. Pain of neural origin (neuralgia) is a common symptom in patients with neuritis and severe neuralgic symptoms in the form of lightning-like pains and burning paresthesias are likely to occur with injury to certain nerves. Burning pains (causalgia) often accompanied by trophic (nutritional) changes in the skin may occur with incomplete lesions of sensory or mixed nerves.

Nerve compression

Certain nerves are especially subject to compression by fibrous tissues (fascial bands), resulting in pains, paresthesias, and weakness in the distribution of the affected nerves (entrapment syndrome). The median nerve, for example, may be compressed by the transverse carpal ligament at the wrist (carpal tunnel syndrome); the lateral femoral cutaneous nerve of the thigh often may be compressed by tight belts or corsets, with resulting pains and paresthesia in the lateral surface of the thigh (meralgia paresthetica); the posterior tibial nerve may be compressed at the anklebone (medial malleolus), with pains and paresthesias in the feet similar to those in the hand resulting from compression of the median nerve; and the genitofemoral nerve under compression of the fibres in the abdominal wall produces painful hyperesthesia (enhanced sensitivity) of the root of the thigh and scrotum. The treatment of neuralgic disturbances associated with entrapment is relief of the constriction by section (cutting) of the fascial band or by surgical relocation of the nerve.

The fifth (trigeminal) and ninth (glossopharyngeal) cranial nerves are subject to neuralgic disturbance of unknown origin; for example, there are no significant lesions observable in the nerves. Trigeminal neuralgia is the most common form of neuralgia and is characterized by the paroxysmal occurrence of attacks of burning, stabbing pains in the face along the course of one or more branches of the nerve. Attacks are usually brief, but they may occur quite frequently and are often precipitated by a touch on a so-called trigger point on the face or by movements of the facial muscles. The disease commonly has its onset in middle or later life, the second and third division (branch) of the nerve being most frequently affected. There may be periods of remission (freedom from symptoms) lasting several months or years, but recurrence is the rule. Treatment is by the administration of drugs (carbamazepine alone or in combination with diphenyl hydantoin). Surgical section of the nerve root is indicated for the cases in which these drugs are not effective.

The pains in glossopharyngeal neuralgia are similar to those of trigeminal neuralgia except that they are localized in the region of the tonsils, posterior pharynx, back of the tongue, and the middle ear. They may occur spontaneously but are often precipitated by stimulation of these areas by the swallowing of the food. The pains may be relieved by anesthetizing the sensitive areas, and the paroxysms may be prevented by section of the ninth nerve. The medical treatment is the same as that for trigeminal neuralgia.

Other neuralgias signalled by sharp shooting pains in various portions of the body may occur as a sign of herpes zoster (virus infection) and are not uncommon in tabes dorsalis (a form of syphilis). In addition, there are a number of patients with persistent or unremitting pains in the face, head, and neck that are not confined to the distribution of the trigeminal or glossopharyngeal nerve (atypical facial pain). The pains are not relieved by the accepted treatment for trigeminal neuralgia. Relief of symptoms may occur, but by no means commonly, after operations on the nerve supply for the head and neck. It is possible that the pain in some of the sufferers is a neurotic manifestation (conversion hysteria).

Pains in the face, head, and neck

DISEASES OF THE SPINAL CORD

The spinal cord may be involved directly by infections, tumours, trauma, developmental defects, vascular lesions, degeneration, and demyelinization (loss of myelin, which insulates neurons) and secondarily by metabolic diseases and blood disorders. The lesions may be localized in the motor or sensory cells, in the horn-shaped gray matter within the spinal cord, or in the fibres (white matter) of the tracts that carry nerve impulses along the cord.

Destruction of the motor cells in the anterior horn produces a flaccid (flabby) paralysis and atrophy (shrivel-

ling) of the muscles innervated by the segments of the spinal cord involved. The classical examples of these are acute anterior poliomyelitis (infantile paralysis) and chronic anterior poliomyelitis (progressive muscular atrophy or amyotrophic lateral sclerosis). Damage, usually from the virus of herpes zoster, to the sensory cells in the posterior root ganglion (posterior poliomyelitis) causes pain and an eruption in those parts of the skin served by the affected segments.

Damage to the descending, or motor, tracts (particularly the corticospinal) produces a spastic paresis (stiff paralysis) of the muscles below the point of the lesion in the cord, with an increase in the tendon reflexes and an extensor plantar reflex response of the foot (Babinski sign).

Injury to ascending, or sensory, tracts in the spinal cord produces a loss of sensation below the level of the lesion, the loss usually being most pronounced for pain, temperature, and proprioceptive sensation. Severe damage to the fibres of some of the principle divisions of white matter (posterior funiculi) at the appropriate level may result in loss of the tendon reflexes.

Spinal cord transection

The spinal cord may be structurally (anatomically) or functionally (physiologically) transected (transverse myelitis) by trauma, tumours, and infectious or demyelinating processes. Below the level of the lesions there is complete paralysis of the muscles, loss of all forms of sensation, and paralysis of the muscles that control the bladder and rectum. The reflexes are lost in the acute stages of severe lesions (spinal shock), but with partial recovery the reflexes return and may become hyperactive. There is also some degree of return of bladder and rectum control. The isolated segment of the spinal cord may become hyperirritable to produce flexor or extensor spasms of the legs when the skin there is stimulated.

Section of only one half side of the spinal cord results in the Brown–Sequard syndrome, characterized by symptoms below the level of lesions. On the same side as the lesion there is paralysis of muscles, an increase in tendon reflexes, a Babinski response, and impairment of proprioceptive sensation. On the side opposite, impairment of pain and temperature sensibility may be observed. The Brown–Sequard syndrome may be caused by tumours, trauma, and by vascular and demyelinating diseases.

Vascular lesions. Hemorrhage into the substance of the cord (hematomyelia) is usually the result of trauma, occurring only rarely in diseases of the spinal blood vessels. Thromboses (clots) of the arteries of the spinal cord are uncommon in comparison to the number of those in the brain. Clots may occur in association with inflamed arteries (arteritis) in syphilis or in other infections of the membranes (meninges), but they are uncommon in sufferers of hardening of the arteries (arteriosclerosis). Acute softening of the spinal cord may follow compression of the arteries or veins by tumours, as in Hodgkin's disease, in which there are abnormal growths of tissue (granulomas). In these cases, the signs of cord involvement are usually of sudden onset and take the form of a transverse myelitis (see above).

Tumours. Spinal-cord tumours are relatively common. They may be located outside of the dural membrane that envelopes the cord (extradural), between the dura and the cord (intradural), or intramedullary (within the substance of the cord). In contrast to tumours of the brain, the extramedullary tumours are more common than those within the substance of the spinal cord; there is also a higher percentage of relatively benign tumours.

Extramedullary tumours are accompanied by nerve-root pains and signs of compression of the cord in the nature of complete or incomplete transverse myelitis or hemisection of the cord (Brown–Sequard syndrome). Intramedullary tumours are accompanied by pains and signs similar to those of the extramedullary type.

Spinal-cord tumours are localized by roentgenograms of the spine and contrast-injected cord (myelography). Treatment is by surgical removal of the extramedullary tumour and radiation therapy of the intramedullary growths. The prognosis (probable outcome) is good in the cases of extramedullary tumours in which the development of symptoms has been slow.

Syringomyelia. Syringomyelia is a disease characterized by growths of special cells called neuroglia (gliosis) and the formation of cavities in the spinal cord and lower brain (medulla oblongata) and by the presence of muscular weakness and wasting. There are various types of sensory disturbances, often of a segmental nature, involving mainly pain and temperature, with signs of injury to the long motor and sensory tracts and the development of disturbances in the skin and bones. Cavitation of the medulla oblongata (syringobulbia) is manifested by difficulty in talking and swallowing, weakness of the muscles of the neck, loss of pain and temperature sensitivity in the face, and drooping eyelids (as in Horner's syndrome). Onset of symptoms is most common in the third or fourth decades of life, but it may occur at any time. Deterioration in syringomyelia is slowly progressive, and the course may extend over many years before the sufferer is incapacitated. There is no known treatment, although radiation therapy is of value in cases with an associated tumour of the spinal cord.

DISEASES OF THE AUTONOMIC NERVOUS SYSTEM

The autonomic nervous system is concerned with the regulation of activities of parts of the body that are not ordinarily subject to voluntary control: the heart, lungs, intestines, genitourinary tract, the pupils of the eyes, blood vessels, salivary glands, lachrymal (tear) glands, and the glands of the skin. It is also concerned with the regulation of the activity of the endocrines and in turn is greatly affected by the secretions of the glands, particularly those of the thyroid, adrenal, and pancreas.

Customarily divided into two portions (sympathetic and parasympathetic), the autonomic nervous system is composed only of peripheral neural structures (*i.e.*, those outside the brain and spinal cord). Nevertheless, it is subject to regulation by centres in the brain's cortex (limbic lobes), hypothalamus, brainstem, and spinal cord.

There are only a few diseases in which the autonomic nervous system is primarily affected. Among them are Raynaud's disease, with such symptoms as pain and cyanosis (blueness) of the fingers and toes resulting from constriction of the blood vessels. In another there is a heightened tendency toward a feeling of giddiness or actual fainting on changes of posture from sitting or supine posture to the erect position; this orthostatic hypotension arises from failure of compensatory constriction by blood vessels in the extremities and torso.

Disturbances in the hypothalamus, the main regulator of the autonomic system in the brainstem, are accompanied by a variety of symptoms, among which are diabetes insipidus (see above), abnormal body temperature, excessive sleep (hypersomnia), obesity, emaciation, stomach ulcerations, and psychiatric disturbances.

Dysautonomia

A rare disease known as dysautonomia (Riley–Day syndrome) is accompanied by widespread disturbance of the autonomic nervous system. In this disorder, which occurs predominantly in Jewish children, there is emotional instability, decreased secretion of tears, increased sweating, transient blotching of the skin, attacks of high blood pressure, episodes of fever and vomiting, convulsive seizures, difficulty in swallowing, absent tendon reflexes, relative insensitivity to pain, impairment of taste discrimination, and absence of the fungiform and circumvallate papillae of the tongue. Degenerative changes have been found in the peripheral sympathetic ganglia and in the lateral columns and various tracts of the spinal cord. In some cases with symptoms of dysautonomia, no pathological changes of this sort have been observed. It is probable that the symptoms are then secondary to a defect in metabolism because a chemical called homovanillic acid is excreted in excessive amounts in the urine. There is no known effective treatment for the Riley–Day syndrome, and it is usually progressive, with death before the age of 20 years.

DISEASES OF THE BRAINSTEM AND CRANIAL NERVES

Brainstem. The brainstem functions in a manner similar to that of the spinal cord in the control of movements of the muscles of the eyes, face, jaw, palate, and pharynx and in receiving touch, pain, and proprioceptive impulses

from these parts. It has special groups of cells (motor nuclei) that serve in regulating the action of the viscera, in controlling the sphincter muscles of the iris of the eye, and the activity of the lachrymal (tear) and salivary glands. There are also sensory nuclei that receive impulses for hearing and equilibrium (vestibular) from the ear and for taste from the tongue. The brainstem serves as a conduit through which impulses are transmitted up to the brain's thalamus and down to the spinal cord. It also has a number of special nuclei that serve to coordinate the activity of other parts of the brain (*e.g.*, the cerebellum and cerebral cortex).

The brainstem is involved in many of the same diseases of infectious, demyelinating, or degenerative nature that affect other portions of the nervous system. This section thus deals mainly with brainstem difficulties produced by vascular (blood-vessel) lesions and tumours.

Vascular lesions. The brainstem is supplied by arteries that are not infrequently the site of sclerosis (hardening). In general, structures near the midline (paramedian area) of the brainstem are supplied by short blood vessels that enter the backside (dorsal) surface. Structures in the lateral portion are nourished by long circumferential vessels that enter the lateral surface and also send branches back to supply portions of the brain's cerebellum. Occlusions (*e.g.*, thromboses) of these arteries or their major branches produce characteristic syndromes (groups of symptoms).

Occlusion of the short vessels to the paramedian area usually involves one or more of the motor nuclei and the corticospinal tract to produce such symptoms as paralysis of eye muscles or of the tongue in one side of the body and the arm and leg on the opposite side (alternating hemiplegia). Lesions of the vessels supplying the lateral area of the brainstem are characterized by impairment of the pain sense in the face on the side of the lesion and in the trunk and extremities on the opposite side (alternating hemianesthesia). In addition, there can be evidence of injury to such structures as the visceral motor nuclei and their tracts.

Tumours. Intramedullary tumours of the brainstem produce symptoms by invading and destroying nuclei and fibre tracts. Both sides are usually involved, and there is enlargement of the brainstem in the region affected. The signs and symptoms are appropriate to those of the nuclei and tracts affected. Signs of increased intracranial pressure are usually absent. Destruction of both abducens nuclei, which serve the sixth (abducens) pair of cranial nerves, commonly leads to paralysis of the muscles that permit one to gaze to either side.

Vestibulocochlear neuromas

Neuromas (nerve tumours) arise in the vestibular portion of the vestibulocochlear nerve and produce symptoms by compression of neighbouring cranial nerves, brainstem, and cerebellum. Ringing in the ear, loss of hearing, jerky eye movements (nystagmus), facial palsy (paralysis), impairment of sensation in the face, and ataxia are common symptoms. The treatment of vestibulocochlear neuroma is by surgical removal.

Cranial nerves. *Olfactory (first cranial nerve).* The sense of smell can be impaired on one or both sides in cases of common cold or of atrophy of the mucous membrane of the nose. The sense of smell may be lost when the nerve filaments of the nose have been torn in trauma to the head or when the nerve has been damaged by tumours in the skull. Hallucination of smell, accompanied by the déjà vu sensation of strange familiarity, transient confusion, or convulsions (hippocampal fits), may occur in association with tumours or other lesions in the olfactory centres of the temporal (side) lobe of the brain.

Optic (second). The fibres of the optic nerve partially cross over (decussate) where they enter the base of the brain (at the optic chiasm). The fibres from the temporal (side) fields of vision cross in the central (medial) portions of the chiasm, while those visual fields closest to the nose stay on the same side of the brain. Lesions in the optic nerve produce blindness in the eye on the same side. Lesions in the medial portion of the optic chiasm cause a loss of vision in both temporal fields (bitemporal hemianopia). Damage on one side of the brain between the chiasm and the visual area of the occipital cortex at the back of the brain produces blindness in the opposite temporal field of vision and in the nasal field on the same side (homonymous hemianopia).

Inequalities in the size of the pupils may be related to disease of the iris or of the optic or oculomotor nerves. As a rule, the pupil of a blind eye is larger than that of a normal opposite eye. Inequalities in the size of the pupils and impairment of their reaction to light may be present with damage to reflex centres in the brainstem by tumours or other diseases. Almost distinctive of syphilis of the nervous system is the Argyll–Robertson pupil, a small pupil that does not react to light or to noxious stimuli (pinching the skin of the neck) but that reacts promptly to voluntary changes in near and far vision (accommodation). A pupillary disturbance sometimes confused with the Argyll–Robertson pupil is seen in Adie's syndrome; in this benign condition the pupils may react so slowly to light that they appear on superficial examination to be nonreactive. The accommodation reflex may also be similarly affected, and the tendon reflexes in the legs are absent.

Oculomotor (third), trochlear (fourth), and abducens (sixth). Injury to these nerves produces a paralysis of the muscles they innervate. Their brainstem nuclei, which affect the movements of the eyes, also may be damaged by tumours, vascular lesions, and demyelinating and inflammatory processes. The nerves may be compressed by extramedullary tumours of aneurysms (abnormal saclike bulges or "blowouts" in blood vessels). The third nerve may be compressed by herniation (rupture) of the brain with sudden increase in intracranial pressure. The sixth cranial nerve, because of its long course at the base of the brain, may be damaged in persons with excessive intracranial pressure. In such cases, however, the paralysis is not specifically useful in localizing the lesion. Paralyses of the ocular muscles are common in a generalized muscular disorder called myasthenia gravis and occasionally in exophthalmos (bulging eyes) associated with hyperthyroidism (oversecretion of the thyroid gland).

Muscular paralysis

Trigeminal (fifth). This is a mixed (sensory and motor) nerve, transmitting sensory impulses from the face and controlling the muscles of chewing.

Lesions in the pons (a bulge of white matter in the brainstem) usually involve the motor and main sensory functions, while lesions in the medulla oblongata affect only the motor portion of the fifth nerve.

The fifth nerve itself may be injured by trauma, tumours, aneurysms, or infections. Occasionally, it may be involved in poliomyelitis and generalized polyneuritis. The sensory and motor brainstem nuclei in the pons and medulla may be destroyed by intramedullary tumours or vascular lesions. In addition, the motor function is frequently impaired in syringomyelia and multiple sclerosis (patches of hardening in the central nervous system). The fifth nerve is also the site of a neuralgic disturbance, trigeminal neuralgia (see above).

Facial (seventh). Injuries to the facial nerve produce a paralysis of the facial muscles with or without loss of taste sensation on the anterior (front) two-thirds of the tongue and a disturbance in the secretion of the lachrymal and salivary glands, depending upon the portion of the nerve that is involved.

Lesions of the facial nerve's nucleus in the brainstem cause a paralysis of the muscles that change the expression of the face. Lesions of the facial-area motor cortex of the brain or in the connections between the cortex and the facial nucleus are accompanied by a partial paralysis of the muscles, which is usually most severe in the muscles of the lower half of the face (supranuclear palsy). The motor nucleus of the seventh nerve in the medulla oblongata may be damaged by tumours, inflammatory lesions, vascular lesions, demyelinizing processes, and sclerosis.

The facial nerve itself may be damaged by extramedullary tumours that are usually vestibulocochlear neuromas (see above). The peripheral portion of the nerve may be damaged by trauma and in such disorders as the polyneuritis associated with leprosy. Not infrequently, there is a

poorly understood paralysis of muscles served by the seventh nerve (Bell's palsy) that often follows exposure to cold (as in riding in an open car). Bell's palsy may occur at any age but is somewhat more common from the 30s to the 50s.

Vestibulocochlear (eighth). Symptoms of involvement of the cochlear (acoustic) branch of the nerve are impairment of hearing and ringing in the ear; those involving the vestibular portion include dizziness, disturbance of postural equilibrium, and nystagmus (involuntary oscillation of the eyeballs).

The nuclei of the eighth nerve may be affected by inflammation, tumours, vascular lesions, and demyelinization. Lesions of the pathways from the nerve's nuclei into the brain on one side of the body generally produce vestibular symptoms only; *e.g.*, nystagmus.

The peripheral portion of the nerve may be damaged by inflammatory processes, especially meningitis. The total and permanent hearing loss in many such cases suggests that the nerve is impaired through thrombosis of the internal auditory artery. The eighth nerve may be compressed by tumours, affected by inflammatory, degenerative, or toxic processes in the middle or inner ear, and by drugs (quinine, streptomycin). A disturbance in the function of the eighth nerve, characterized by ringing in the ear (tinnitus), impairment of hearing, and paroxysmal sensations of the environment revolving (vertigo), is known as Menière's syndrome.

Menière's syndrome

Glossopharyngeal (ninth). The ninth cranial nerve has motor fibres that serve the constrictions of the pharynx and has other fibres that transmit general sensory impulses from the upper part of the pharynx and from the taste receptors in the posterior third of the tongue. Isolated lesions of the ninth nerve are rare, but the nerve may be affected by any of the processes that involve the brainstem. The glossopharyngeal nerve is occasionally the site of a neuralgic disturbance, glossopharyngeal neuralgia (see above).

Vagus (tenth). Lesions of one of the vagal nuclei (the nucleus ambiguus in the medulla) cause difficulty according to their site. Lesions confined to the lower portion of that nucleus impair swallowing, while lesions of its upper portion produce difficulty in talking. Damage to only one side of another (the dorsal motor) nucleus produces no symptoms of autonomic disturbance, but damage to both sides is fatal. Lesions of the nuclei of the vagus nerve may result from inflammatory disorder (particularly anterior poliomyelitis), syringomyelia, intramedullary tumours, vascular damage, demyelinizing disorders, and degenerative diseases.

The peripheral portion of the tenth nerve or its branches may be involved in polyneuritis or can be compressed by tumours or aneurysms. Injury to the pharyngeal branches results in difficulty in swallowing. Lesions of the superior laryngeal nerve produce anesthesia of the upper part of the larynx and paralysis of the cricothyroid muscle in the Adam's apple. Impairment of the recurrent laryngeal nerve, frequent in aneurysms of the aorta (the great artery of the heart), causes hoarseness and difficulty in speaking (dysphonia) as a result of unilateral paralysis of the abductor muscle of the vocal cord. Bilateral abductor paralysis, which may occur rarely as a result of pressure on both recurrent laryngeal nerves (or when both nuclei ambiguus are involved by tumours or syringobulbia), causes breathing difficulty without any direct impairment of voice production (phonation).

Spinal accessory (11th). The nucleus of the 11th nerve may be destroyed by tumours and inflammatory, degenerative (syringobulbia, progressive muscular atrophy), and demyelinizing processes in the medulla. The extramedullary and peripheral portions of the nerve may be injured in meningitis, tumours, and destructive processes in the occipital bone at the back of the skull.

Hypoglossal (12th). Injury to the 12th nucleus or to the nerve itself causes atrophy of the muscles of the tongue. In bilateral damage, the tongue cannot be moved from the floor of the mouth. In unilateral lesions, half of the tongue is atrophied and deviates to the side of the lesion on protrusion. Movements toward the opposite side are impaired. The 12th nucleus may be involved in acute anterior poliomyelitis, syringobulbia, intramedullary tumours, progressive bulbar palsy (in which cases the atrophy will be accompanied by fibrillation of the tongue), or by vascular lesions. Lesions of the peripheral portion of the nerve are rare.

MOTOR SYSTEM DISORDERS

Movements of muscles are under the direct control of motor neurons (cells) in the brainstem and spinal cord. Voluntary contraction also depends on the motor cortex of the brain. The motor cells in the spinal cord and brainstem are affected by impulses from many sources, including the muscles; there is also considerable interaction (feedback) with other parts of the nervous system.

Control of muscle movements

Since diseases of the spinal cord, brainstem, cerebellum, and cerebral cortex are considered elsewhere in this article, this section deals with injuries to the basal ganglia, a collection of nuclei at the base of the brain. Specialized structures included among the basal ganglia are called the red nucleus, substantia nigra, subthalamic nucleus, the globus pallidus, the putamen, and the caudate nucleus. All of these nuclei receive incoming (afferent) impulses and send fibres to the brain's thalamus, which relays them to the motor cortex and other areas of the brain's cerebrum. There are feedback mechanisms from the cortex to the basal ganglia, as well as to the cerebellum and the brainstem.

In general, disease of the basal ganglia results in disturbance in the tone of the muscles (so-called cogwheel rigidity) and produces tremor and other involuntary movements (described by such terms as chorea, athetosis, dystonia, and hemiballism). These disorders are often accompanied by symptoms of autonomic dysfunction, because neighbouring hypothalamic nuclei also may be injured. The tremor and other involuntary movements are influenced by the emotional state of the sufferer, increasing when he is excited or tense and disappearing when he is asleep.

The basal nuclei may be individually injured by the usual infections, trauma, and tumours, but classical symptoms of basal ganglia disease involve the cells of these nuclei in a diffuse manner. The symptoms (*e.g.*, rigidity and tremor) are absent when specific motor tracts in the internal capsule (higher in the brain) are also injured, and they also tend to yield to symptoms of paralysis (hemiplegia) from a cerebral vascular lesion.

Chorea. *Sydenham's chorea.* The basal ganglia may be involved in rheumatic fever, with the occurrence of choreic (jerky) movements of the extremities, face, and trunk. Also called St. Vitus' dance, Sydenham's (or minor) chorea is most common in children but may occur in young pregnant women (chorea of pregnancy). The chorea usually abates by itself in a few weeks; it is managed with rest in bed in a quiet room and by the administration of sedatives or tranquillizers.

St. Vitus' dance

Chronic progressive chorea (Huntington's chorea). This hereditary disorder of the basal ganglia also involves shrinkage of other brain parts, such as the cerebral cortex. It is characterized by choreic movements of the trunk and extremities, psychotic behaviour, and gradually worsening dementia. It is relatively rare except in communities where a number of families with the genetic tendency are located. The symptoms may become manifest at any age but commonly appear between the ages of 30 and 50. The choreic symptoms may precede or follow the development of psychotic symptoms or mental deterioration. This adult chorea is chronically progressive, requiring institutional care and leading to death about 15 years after its onset. There is no treatment known, although the violence of the choreic movements is influenced to some extent by tranquillizers.

Paralysis agitans (Parkinson's syndrome). The symptoms of Parkinsonism have been traced to widespread, diffuse lesions (usually of unknown origin) in the basal ganglia. Chief among the signs are a masklike face, speech difficulties (dysarthria), alternating tremor, stooped posture, abnormalities of gait, slowness and poverty of movement (akinesia), loss of associated move-

ments, and symptoms of dysfunction of the autonomic motor system.

Symptoms resembling those of Parkinsonism may result from carbon monoxide or manganese poisoning. The syndrome was a common sequel of a worldwide epidemic of encephalitis lethargica (sleeping sickness) that abated in the mid-1920s. Patients taking certain drugs (*e.g.*, reserpine or phenothiazines) may develop the typical symptoms of Parkinsonism, but these disappear when the medications are withdrawn.

Although the degenerative changes are widespread, the basal nucleus called the substantia nigra is severely affected with almost complete loss of its melanin- (pigment-) containing cells. A decrease in the concentration of a substance called dopamine in the substantia nigra and in other basal ganglia is found. It is thought that the disturbance in the function of these cells and their eventual destruction result from the deficiency. Dopamine is essential to the transmission of nerve impulses in the basal ganglia.

Typically a disorder of old age, paralysis agitans is one of the most common diseases of the nervous system. It has been estimated that one out of every 40 individuals develops the disease. The onset of the symptom is insidious, usually in the sixth to eighth decades of life. In spite of its somewhat dramatic nature, the tremor itself is less disabling to the patient than are the accompanying rigidity and akinesia.

Other symptoms that almost inevitably develop are impairment of handwriting (micrographia), increased sweating, undue sensitivity to heat, excessive salivation, nocturnal incontinence (*e.g.*, bedwetting), and mental symptoms. The latter are often related to medication (particularly with such drugs as atropine). The course of the disease is usually chronically progressive, extending over one or more decades. Death is usually the result of factors (*e.g.*, infections) other than the disease itself.

Levodopa

Treatment includes the administration of a substance called levodopa, a precursor of dopamine. A bioamine (related to the amino acids of proteins), levodopa has a remarkable beneficial effect in relieving the rigidity and akinesia but is less effective for the tremor. It remains to be shown whether levodopa will halt the progress of the disease; other drugs with atropine-like action are also of value.

Surgical destruction of various portions of the basal ganglia has been used in the treatment of Parkinson's syndrome. The most effective method is destruction by freezing (cryotherapy) of part of the thalamus (the ventrolateral nucleus). This usually results in relief of tremor and rigidity, but the akinesia and the progress of the disease are not as significantly affected.

Dystonia musculorum deformans. Of unknown cause, this rare disease of the basal ganglia is characterized by the development of slow, forceful, and sustained twisting and writhing movements of the muscles of the trunk or extremities. Perhaps there is a hereditary predisposition, since it is most common in families of Russian-Jewish extraction. The onset is commonly between the ages of five and 15, often showing as an abnormality of the gait with bizarre stepping or bowing movements. The disease tends to be progressively incapacitating, although mental faculties are preserved. In some cases, the disease ceases to progress, and the sufferers are able to carry on useful lives compatible with their muscular disability. Good results have been reported from surgical destruction of the same portion of the thalamus as in paralysis agitans, and a limited percentage of cases have been treated with levodopa with beneficial results.

Spasmodic torticollis (wryneck). This typically intermittent disorder may be a feature of postencephalitic Parkinsonism or of Wilson's disease (a hereditary defect in copper metabolism). Wryneck may also occur without any known cause, although such influences as psychoneurosis (hysteria) have been suggested. The symptoms are usually limited to twisting movement of the head caused by intermittent or sustained contraction of the muscles of the neck. The course is usually characterized by an increase in the severity of the movements, but partial or complete remissions (improvements) may occur. The symptoms have been relieved in some patients with thalamic surgery.

DISEASES OF THE CEREBELLUM

A two-lobed, wrinkled structure at the rear underside of the brain, the cerebellum may be damaged by tumours, trauma, vascular lesions, infections, intoxications, and metabolic diseases. In addition, the cerebellum is involved in a number of hereditary degenerative diseases, the most common of which is Friedreich's ataxia.

This disorder is characterized by the development of ataxia of the trunk and extremities, absence of deep reflexes, loss of proprioceptive sensations in the extremities, extensor plantar (Babinski) responses, clubfeet, and curvature of the spine (scoliosis) in the first or second decade of life. Dysarthria, muscle atrophies, degeneration of the optic nerve, and involvement of the heart muscles may occur in the late stages. The disease may occur in abortive forms (*formes frustes*) that may be compatible with a relatively normal life-span, but in the usual case the course is slowly progressive, with complete incapacity by the age of 20. Death is usually caused by intercurrent infections or involvement of the heart muscles.

Hereditary ataxia

Other heredodegenerative diseases of the cerebellum that have been considered to be variations of Friedreich's ataxia are (1) hereditary ataxia with muscular atrophy (Lévy–Roussy syndrome), which differs from Friedreich's ataxia in that there is more severe wasting of the muscles of the legs, a lesser degree of ataxia, and a less severe course; (2) hereditary cerebellar ataxia with spasticity (Marie's or Sanger-Brown's ataxia), which differs from Friedreich's ataxia in showing late onset of symptoms, increased response of the tendon reflexes, and frequent occurrence of optic atrophy and oculomotor palsies; and (3) olivocerebellar and olivopontocerebellar atrophy, a rare syndrome characterized pathologically by degenerative changes in the cerebellum, olivary nuclei in the medulla oblongata, and pons. Symptoms include the development in middle life of ataxia of trunk and extremities and scanning speech (regularly occurring pauses in speaking); there are signs of involvement of the basal ganglia in a small percentage of cases.

INTRACRANIAL TUMOURS

Tumours are found in approximately 2 percent of all human skulls opened after death. They may arise from the skull itself, in the meninges, or in any of the elements of the nervous system, or they may be metastases that spread from tumors in other organs. Cancers (gliomas) of specialized cells (called glia or neuroglia and with which the neurons are closely associated) lead the list of intracranial tumors, accounting for about 45 percent.

The general symptoms include headache, nausea and vomiting, diplopia, lethargy, somnolence, mental changes, and convulsive seizures. They are related to such factors as increased intracranial pressure and edema (swelling) of the brain. Localized (focal) bodily symptoms of brain tumour that are the result of destruction of brain tissue reflect the specific part of the brain that is involved.

The onset of symptoms is insidious, and the progress extends over many months, except in cases of rapidly growing glial tumours (glioblastomas) and of metastatic tumours, with which death may ensue in a few months.

An intracranial tumour is suspected when severe headaches develop in a person previously free of them, when convulsive seizures have their onset in adult life, or when there is behavioral or clinical evidence of slowly progressive damage to one portion of the brain. The findings on neurological examination serve to localize the areas of the brain damage in most of the sufferers. In others, special diagnostic tests may be required. The examinations that are of value in establishing the diagnosis of a brain tumour and in determining its location and type include plain X-rays of the skull, echoencephalography, electroencephalography, examination of the cerebrospinal fluid, ventriculography, angiography, and radioactive scanning (see above *The neurological examination*).

The clinical course of the sufferer is related to the type

of brain tumour and to its location. In general, the signs and symptoms are of slow onset and progressive in most instances until they cause death; remissions are rare.

Meningiomas (tumours of the membranes around the brain) and acoustic (eighth-cranial-nerve) neuromas are usually slow-growing, well-encapsulated hard tumours located on the outer surface of the brain. This makes them relatively accessible to the neurosurgeon, and usually they can be removed partially or *in toto*. Metastases and soft glial tumours (particularly the glioblastomas that infiltrate or invade both cerebral hemispheres or brainstem) typically have such a rapid course that they are not amenable to surgical therapy. The treatment of brain tumours by surgical removal or by radiation therapy is supported by the use of anticancer drugs. Chemotherapy of malignant brain tumours, however, is still in its experimental stages.

The choice between surgery and radiation therapy is often difficult. Surgery is most often used for such growths as meningiomas, cystic (saclike) tumours of the brain's third or fourth ventricle, and most tumours growing from congenital rests (fragments of embryonic tissue trapped inside the body). Radiation therapy (alone or combined with partial surgical removal) is favoured for malignant gliomas and metastatic tumours. Tumours of the pituitary gland are likely to be treated by radiation therapy unless surgery becomes urgent because of rapid failure of vision or when it is suspected that there has been cyst formation in the tumour.

DISEASES OF THE CEREBRUM

The cerebrum, which comprises the upper bulk of the brain, may be involved in malformations and has the typical susceptibility of neural tissue to trauma, tumours, bacterial and other infections (encephalitis and abscess formation), demyelinating diseases, neurochemical disorders, and toxins. This section, however, focusses on injury to the cerebrum by vascular and degenerative diseases.

Cerebral vascular accident. Sudden cerebral damage (accident) is a leading cause of death throughout the world. In addition, large or small vascular lesions are discovered in about 25 percent of routine autopsies. Cerebral arteriosclerosis, usually a specific feature of generalized hardening of the arteries, is the most common cause of cerebral vascular disease.

Strokes

Perhaps two-thirds of all cerebral vascular accidents (sometimes called strokes) are the result of thrombosis: blood clots that plug (occlude) the vessels of the brain. Cerebral hemorrhage, although less common (about 20 percent) than cerebral thrombosis, is the most serious form of cerebral vascular accident. The hemorrhage is usually from a vessel at the base of the brain. Blood is discharged (extravasated) into the brain substance itself, often resulting in apoplexy (coma and paralysis). Death is almost inevitable if the hemorrhage is large.

Occlusion of one of the branches of the main arteries in the brain by arteriosclerosis or embolism (any obstruction, including air bubbles, fat globules, as well as clots) usually results in the death of tissues (infarct formation) served by the vessel. Occlusion of the large vessels in the neck may be accompanied only by transient symptoms (cerebral vascular insufficiency) or by none of consequence. If, however, collateral circulation (through smaller, unaffected vessels) is inadequate, infarction still may ensue.

Obstruction of the anterior cerebral artery on one side of the brain usually causes paralysis and sensory loss in the leg on the opposite side. If the artery of the dominant (almost always left) hemisphere is affected, mental confusion and aphasia (*i.e.*, speech impairment) may also occur.

Plugging of the main branch of the middle cerebral artery on one side produces paralysis and sensory loss on the opposite side, mainly affecting the face and the arm. Aphasia is also present when the lesion is in the dominant hemisphere. Occlusion of smaller vessels of the middle cerebral artery produces hemiplegia, hemianopia (one-sided blindness), inability to read (alexia) or recognize sensory stimuli (agnosia), or to perform skilled movements (apraxia) in combinations or as isolated symptoms.

Occlusion of the main trunk of the posterior cerebral artery on one side causes damage to the thalamus and the visual (occipital) cortex, with hemianopia, fleeting hemiparesis (muscular weakness), and sensory loss accompanied by burning pains, choreiform movements, and ataxia (thalamic syndrome of Déjerine–Roussy).

Blocking of the basilar artery on the underside of the brain is serious, often resulting in paralysis and sensory loss in all extremities and in the muscles supplied by the brainstem nuclei. Occlusion of the vertebral artery that runs down toward the spinal cord or of minor branches of the basilar artery is usually not fatal, although the symptoms can be disabling.

Chances of survival are poor for the victim of a cerebral hemorrhage of any considerable size. The majority of those with a cerebral thrombosis are likely to recover from the initial damage, although a number of them die in the first few weeks after the cerebral accident from pneumonia or from heart or kidney complications. Survivors tend to show gradual improvement in the neurological difficulties over the ensuing months.

General care in the initial stages is the chief form of therapy for cerebrovascular accidents. Physiotherapy (*e.g.*, special exercise) and speech therapy are important in the convalescent stages of people with hemiplegia or aphasia. Attempts have been made to remove surgically the vessels and debris of intracranial hemorrhage, but the poor condition of the sufferer militates against success. Long-continued administration of drugs to prevent clotting (anticoagulants) has been recommended for persons with intermittent symptoms of cerebral thrombosis (or any embolus). Reconstructive surgery (*e.g.*, reaming out) of the vessels in the neck for patients with arteriosclerotic narrowing of those vessels has been tried as a preventive (prophylactic) measure with a moderate degree of success.

Intracranial aneurysms and primary subarachnoid hemorrhage. Intracranial aneurysm usually arises from congenital weaknesses in the walls of affected vessels, which bulge locally in a manner resembling that of a defective inner tube in a pneumatic tire. Found in approximately 2 percent of routine autopsies, these aneurysms may rupture, being responsible for about 10 percent of all cerebral vascular lesions. About 85 to 90 percent of intracranial lesions are found in five arteries: the middle cerebral, internal carotid, basilar (and its branches), anterior communicating, and anterior cerebral. Rupture may occur at any age but is most common in middle life. In addition, the presence of an intracranial aneurysm may not even be suspected until it ruptures. Occasionally, it may be so located that it compresses one or more of the cranial nerves, thus alerting the physician to the peril of future hemorrhage.

Blood-vessel rupture

Symptoms associated with the rupture of an intracranial aneurysm are usually those resulting from the presence of blood in the subarachnoid spaces between the brain and its membranes. Headache and stiffness of the neck are common, and coma may occur if the aneurysm ruptures in the cerebrum as well as into the subarachnoid space. After such hemorrhages the cerebrospinal fluid becomes grossly bloody, and intracranial pressure increases. About 35 percent of sufferers of aneurysm rupture die as a result; those that survive are in constant danger of subsequent intracranial bleeding. The presence and location of an aneurysm that caused a subarachnoid hemorrhage can be established by angiography.

There is controversy as to the best method of treatment. Some advocate conservative treatment in the form of bed rest and therapy that avoids radical intervention in the lesion. Other physicians, because of the ever-present risk of rupture, advise surgical treatment of the aneurysm by tying off (ligating) its neck or by ligation of the common or internal carotid artery in the neck.

Presenile dementia (Alzheimer's and Pick's diseases). Generalized sclerosis (hardening) of the intracranial arteries may result in occlusion of small cortical vessels with

gradual decrease in mental faculties (arteriosclerotic dementia). In addition, the cerebral cortex is subject to degenerative changes associated with diseases other than arteriosclerosis, the most common of these being Alzheimer's and Pick's diseases. The cardinal symptoms of these two diseases have their onset between ages 40 and 65 and include progressive dementia and disturbances in speech (aphasia or dysphasia). The degenerative changes involve the nerve cells in parts (association areas) of the frontal and temporal lobes of the brain. Although there are differences in the nature of lesions found in Alzheimer's and Pick's forms of the disease, the clinical symptoms are so similar that differentiation between them is rarely feasible while the victim is alive. The course is progressive, terminating in incapacity and death within four to ten years. Paralysis is not a feature of the disease until the terminal stages.

Degenerative diseases. *Primary degeneration of the corpus callosum (Marchiafava–Bignami disease).* This rare degenerative disease involves the tracts of white matter (the corpus callosum, the anterior commissure, and the middle cerebellar peduncles) that connect the two halves of the brain. The disorder usually but not exclusively affects Italian males who are addicted to the use of red wine. The onset is insidious, with psychotic episodes and dementia, the progressive course leading to death in from three to six years.

Central pontine myelinosis. This rare disease is characterized by degeneration of the central portion of the pons (which bridges the medulla with the cerebellum). Occurring in young or middle-aged adults, central pontine myelinosis is commonly associated with chronic alcoholism or other conditions associated with chronic malnutrition. Confusion and other mental symptoms, quadriplegia (paralysis of all four limbs), pseudobulbar palsy, and cutaneous sensory loss are common symptoms. Death usually occurs within a few weeks or months after the onset of symptoms.

Alcoholism and malnutrition

Down's syndrome (mongolian idiocy). In this congenital disorder associated with chromosomal abnormality, the mentally defective sufferers exhibit facial and bodily features that have a superficial resemblance to those normally present among such Asians as the peoples of Mongolia (see GENETICS, HUMAN; BIRTH DEFECTS AND CONGENITAL DISORDERS).

Physical features of the disease include shortening of the skull, flattening of the face, narrowing and slanting of the gap between the eyelids, shortening of the nose, broadening of the tongue, shortness of stature, widening of the hand with simian (apelike) crease on the palm, spina bifida, webbing of fingers or toes (syndactylism), and maldevelopment of internal organs, particularly the heart.

Many of the affected individuals die in infancy, but a number reach adulthood. The chief causes of death are the associated congenital heart disease and infections.

Cerebral palsy (Little's disease). Cerebral palsy describes a heterogeneous group of symptoms that arise from damage to the nervous system before birth, at birth, or early in life. The syndrome that results is related to the portion of the nervous system affected. The groups of symptoms can be roughly identified as spastic, athetotic, or ataxic, depending upon whether the cerebral cortex, basal ganglia, or the cerebellum is the most severely affected (see above).

Cerebral palsies can derive from chromosome abnormalities, inherited metabolic defects, prenatal injury to the fetus and factors that lead to premature birth, damage to the infant because of oxygen deficiency or mechanical injury at the time of delivery, and jaundice, hypoglycemia, or infections of the nervous system in the first few months after birth. Cerebral palsy is observed slightly less often than once in every 1,000 births in the U. S.

Severe damage to the nervous system may be evident at the time of birth, but it is more usual for the symptoms to be manifested during the second to fourth year of life through failure of the child to attain normal physical and mental development.

The course in cerebral palsy is quite variable; if damage to the nervous system is severe, the infant may die in the first few months of life. Children with a lesser degree of damage may develop to normal adult life with only mild spasticity, athetosis, or ataxia. In most cases, however, there are serious handicaps requiring special care throughout life (*e.g.*, physiotherapy and special education).

INFECTIONS

The nervous system may be invaded by a wide variety of disease-producing agents (*e.g.*, bacteria, viruses). The symptoms that result are related to the site of the infection and the type of invading agent. These infections can be roughly divided into large groups (meningitis and encephalitis), depending on whether the coverings (*e.g.*, meninges) or the functioning elements (parenchyma), such as the neurons of the brain, are mainly affected.

Bacterial infections. Infections of leptomeninges (the arachnoid and pial membranes) are divided into two groups, acute (brief and severe) and subacute (more prolonged and milder), according to the nature of the disease process and the mode of onset of symptoms. They can be further subdivided as purulent (associated with pus formation) and nonpurulent, the former representing the greater degree of inflammatory reaction.

Acute purulent meningitis. This syndrome is usually the result of infection with disease-producing (pathogenic) bacteria, commonly meningococcus, *Hemophilus influenzae,* pneumococcus, streptococcus, or staphylococcus. With the exception of infections with meningococcus, which tend to appear in epidemics, the occurrence of the disease is sporadic. People of any age may be affected, but the incidence is high in infancy and early childhood.

Early symptoms include fever, headache, nausea and vomiting, stiffness of the neck, pains in the back and clouding of consciousness. Convulsions may occur in infants and children. The diagnosis is made from the fluid tapped from the spine (lumbar puncture). The fluid is under increased pressure and is purulent, containing several thousand pus (white-blood) cells per cubic millimetre, and the glucose content is low. The causative organism can be cultured (grown from the spinal fluid).

While the mortality rate was once quite high, most victims of acute purulent meningitis can be cured with drugs (*e.g.*, penicillin). The choice of the drug to be used is determined by testing the sensitivity of the cultured organisms. Occasionally, especially in sufferers who have already received antibiotics, organisms fail to be cultured from the cerebrospinal fluid. Under these circumstances, treatment with a combination of drugs (*e.g.*, penicillin, sulfadiazine, and chloromycetin) is commonly undertaken in the expectation that at least one of them will be effective.

Drug treatment

Subacute meningitis. This form of meningitis is usually the result of infection with tubercle bacilli or mycotic organisms (fungus). The syndrome shows a less acute onset than does acute purulent meningitis; the inflammatory reaction also is more chronic, the course is more prolonged, and relapses are likely to occur. The cerebrospinal findings also differ in that the fluid is usually cloudy ("ground-glass") in appearance rather than purulent. The number of pus cells is lower, falling in the range of 50 to several hundred per cubic millimetre, and the glucose content is not so sharply reduced.

Tuberculous meningitis. The occurrence of tuberculous meningitis greatly decreased after the introduction of effective therapy for tuberculosis of the lungs and other organs; nevertheless, it remains the most common cause of death from tuberculosis in children. The mortality rate in tuberculous meningitis has been reduced from almost 100 percent to less than 20 percent by combining such drugs as streptomycin, isoniazid, and para-aminosalicylic acid administered over a period of eight to 12 months.

Fungal meningitis. A great increase in the number of cases of meningitis arising from infections with fungi (most commonly cryptococcus) followed the general use of drugs that kill bacteria normally present in the body. Apparently, such bacteria prevent the growth of fungi; in addition, some drugs (*e.g.*, corticosteroids) suppress nor-

mal immune reactions to increase susceptibility to invasion by fungi. Low-grade symptoms similar to those of tuberculous meningitis are the most common manifestation of mycotic (fungal) infections of the membranes, but such problems as abscesses may also arise.

Infections with the fungi are usually fatal after an acute or prolonged course. Treatment remains unsatisfactory, although sulfonamides, tetracycline, or penicillin may be helpful in some forms. Prolongation of course and presumptive cures in some cases have been reported with the use of other drugs (*e.g.*, hydroxystilbamidine and amphotericin B).

Cerebral epidural and subdural meningitis. The cerebral spaces on either side of the outermost of the meninges (the dura) may be invaded by pathogenic bacteria by extension of infections from the scalp, bones of the skull, ear or nasal sinuses. The symptoms, associated with epidural or subdural pus, superficially resemble those of subacute meningitis, but convulsions, hemiplegia, aphasia, or other focal signs may occur when there is thrombosis of the vessels in the meninges. The cerebrospinal-fluid findings are those that occur with infections near the meninges that are not accompanied by a bacterial invasion of the pia and arachnoid (the leptomeninges). There is a moderate increase in the number of cells in the fluid, but the glucose content is normal, and no organisms can be cultured from the fluid. Treatment is by surgical evacuation of the pus and the administration of drugs to combat the infection.

Spinal epidural abscess. The fatty tissue in the spinal epidural space may be invaded by micro-organisms that first infect the bones of the spine or from purulent infections of the skin. The symptoms of this spinal form of meningitis are pain in the back and stiffness of the neck. As the abscess grows, the spinal cord may become compressed. Thrombosis of the spinal vessels as a result of the infection produces an acute softening of the cord with signs of transverse myelitis (see above). Surgical evacuation of the pus and the administration of antimicrobial agents comprise the treatment. Complete recovery is the rule if the operation is performed before the onset of paralysis.

Brain abscess. Invasion of the brain itself by pathogenic organisms results in the formation of pus that eventually becomes encapsulated (walled in). Brain abscess usually follows infections of the mastoid bone (behind the ear), of the nasal sinuses, or of gaping (compound) fractures of the skull, but it may also be the result of the spreading of micro-organisms that initially attacked the lungs or heart. The incidence of brain abscess greatly decreased after improvement in methods for treating infections of the mastoid bone and nasal sinus.

The symptoms and signs of brain abscess are similar to those of brain tumours or other expanding intracranial lesions. An abscess may be suspected when signs of an expanding intracranial lesion develop in someone known to have an infection in the skull or elsewhere in the body. This can be confirmed, and the site of the abscess may be located with angiography or pneumoencephalography (see above). Treatment calls for the surgical evacuation of the pus or removal of the whole encapsulated abscess followed by the administration of antimicrobial agents.

Virus infections. The central nervous system of man may also become infected with viruses (see VIRUS), including those that produce mumps, so-called cold sores (herpes), and rabies. Such virus infections in man usually occur in epidemics, but sporadic cases are not uncommon. Many mammals serve as reservoirs of viruses (*e.g.*, horses in equine encephalomyelitis), and a number of arthropods (*e.g.*, ticks and lice) have been identified as transmitting viral infection. The virus of infantile paralysis can produce disabling motor symptoms, and that of herpes zoster (shingles) may lead to severe sensory disturbance (*e.g.*, excruciating pain), but in most viral disorders the symptoms are those of acute meningitis or encephalitis of mild or moderate degree. In the latter cases, the disease is usually self-limited, recovery even without treatment ensuing with few or no aftereffects (sequelae). Fatalities or severe sequelae are uncommon except in rabies, equine encephalitis, herpes invasion of the brain, and acute anterior poliomyelitis (see below). In such viral infections, severe destruction of the brain or spinal cord may result in extensive paralyses and impairment of mental function.

Cold sores, rabies, and shingles

Diagnosis is established by the recovery of the virus from such sources as the blood, cerebrospinal fluid, saliva, nasal washings, and brain tissues (usually at autopsy) and from special serological (blood) tests similar to those for some venereal diseases (*e.g.*, syphilis). No adequate treatments for the majority of virus infections of man are known. Immunization procedures are available for some of them (*e.g.*, rabies, acute anterior poliomyelitis).

Acute anterior poliomyelitis. Although the introduction of effective immunization (Salk and Sabin vaccines) almost completely eradicated it in most places, this disease (known also as infantile paralysis and Heine-Medin disease) was a scourge of children before the late 1950s. The widespread distribution of the disease, its devastating symptoms, and difficulties in treating the acute and convalescent stages all highlight the dramatic benefits obtained by vaccination. The disease was at one time worldwide in distribution, affecting millions of victims, commonly children between the ages of one and ten. Particularly severe epidemics occurred in the early 1950s in Belgium, Denmark, Germany, and the U.S. Inadequate immunization programs apparently led to outbreaks in the late 1960s in Spain, Turkey, Africa, and Latin America. The symptoms at the onset are those of any acute infection (*e.g.*, fever), followed in a few days by headache, stiffness of the neck, and flaccid paralysis of muscles.

The extent of the paralysis varies from weakness of a few muscles in one extremity to complete paralysis of all the muscles of the trunk and extremities. When death occurs, it is usually from paralysis of the respiratory muscles. Partial paralysis of the affected muscles is a common residual symptom, and many survivors of infantile paralysis live into old age with atrophy and weakness of one or more extremities.

Other infections. *Rickettsial diseases.* A mild or moderately severe degree of encephalitis may accompany infections with micro-organisms known as rickettsiae, which are intermediate in structural complexity between bacteria and viruses. Rickettsial infections are transmitted to man by such animals as the flea, louse, or tick. These human diseases include typhus, spotted fever, and scrub typhus (tsutsugamushi disease), all of which can impair the blood vessels of the brain. The diagnosis is made by recovery of the organism from laboratory animals injected with blood from infected persons and through serological tests (Weil–Felix reaction). Treatment is with such drugs as aureomycin, chloramphenicol, or terramycin.

Typhus and scrub typhus

Neurosyphilis. Syphilis of the nervous system became less common after the introduction of penicillin and other drugs for treating early manifestations of syphilis (see REPRODUCTIVE SYSTEM DISEASES). The nervous system seems affected in approximately one-third of all cases of syphilis that remain long untreated. Although evidence of mild subacute meningitis has occurred occasionally in the primary or secondary stages of the disease, symptoms of involvement of the nervous system are usually delayed for some time (15 to 25 years) after the initial infection. Any part of the nervous system can be affected, but the most likely involvements are of the spinal cord (tabes dorsalis) or of the brain (dementia paralytica).

Tabes dorsalis is characterized by degenerative changes in the sensory (posterior) roots of spinal nerves and in the spinal cord itself and by sharp, cutting pains, progressive ataxia of the limbs, loss of reflexes, varying degrees of sensory loss, and disorders of the functions of the bladder, rectum, and sexual organs.

Dementia paralytica (general paresis of the insane) results from a chronic syphilitic meningoencephalitis that destroys the cerebral cortex, leading to general dissolution and death within three to ten years. The disease at one time accounted for 10 percent of all admissions to mental hospitals.

Syphilitic invasion of the nervous system can be pre-

vented by drug treatment in the earlier (primary or secondary) stages of syphilis, and once in the nervous system it can be cured or arrested the same way.

DEMYELINATING DISORDERS

Loss of myelin, the fatlike sheath around many nerve fibres, may be a feature of vascular lesions, infectious processes, nutritional deficiencies, and other diseases. Demyelinization in these cases is only one feature of the pathological process; there are other destructive neuronal changes as well. The term demyelinating diseases, however, is used specifically for significant myelin loss of unknown cause. When there is destruction of previously normal myelin, the disease is called myelinoclastic. By contrast, in what are termed leukodystrophies, there is a defective development (dysmyelination) rather than destruction of the myelin sheath. Dysmyelination is often accompanied by the deposition of pigment in the degenerated parts and occasionally also in the peripheral nerves and other organs.

Myelinoclastic (sheath-destructive) diseases. *Multiple sclerosis.* Multiple sclerosis is a chronic disease characterized by the presence of many separate patches of demyelination (lesions) in the brain and spinal cord and by a wide variety of neurological signs and symptoms that at first have a tendency to come and go. While there is variation in the number of symptoms from time to time, evidence of multiple lesions is usually present at any one time. The cause of this myelinoclastic disorder is poorly understood; some theorize that it is attributable to an autoimmune reaction (*i.e.*, immunity against one's own cells) or to infection by slow-acting viral invaders.

Symptoms of multiple sclerosis

Onset of symptoms may be at any age but is most commonly observed in the late teens or in the third or fourth decade of life. Any part of the nervous system may be affected, but common early symptoms are those of involvement of the optic nerve (optic neuritis) or of the spinal cord with paresthesias (*e.g.*, prickling or burning) in one or more extremities and weakness or ataxia in the arms or legs. The initial symptoms usually disappear almost completely, but, in the course of a few months or years, new symptoms are most likely to appear. The course of the disease is quite variable; some victims may live for many years without suffering severe disability, but more often each new attack is followed by more severe sequelae. The disease is seldom rapidly fatal, and the average duration of life after onset is about 25 years, or about five years after relatively complete incapacitation. In the terminal stages, practically every portion of the nervous system is affected. There are variations from the typical form of multiple sclerosis: neuromyelitis optica, in which the optic nerves and the spinal cord (transverse myelitis) are mainly affected; and acute encephalomyelitis, in which there is an acute onset of widespread evidence of damage to the nervous system. Many of these atypical cases follow a course similar to that of multiple sclerosis. No treatment is known to favourably influence the symptoms of multiple sclerosis at any stage of the disorder.

Schilder's disease (encephalitis periaxialis diffusa). This is a rapidly progressive sheath-destructive disease in which there is widespread demyelination of the neurons in the cerebral hemispheres. Children are more commonly affected than adults, the disease rapidly progressing to death within a few years of onset. The common symptoms are convulsive seizures, repeated myoclonic jerks (as in the occasional sudden, jerky awakening experienced when one is drifting into sleep), optic atrophy, hemianopia, hemiplegia, blindness from cortical damage, aphasia, and mental symptoms (*e.g.*, dementia).

Dysmyelination disease. The classic example of this form of myelin disease, metachromatic leukodystrophy (sulfatide lipoidosis), exhibits diffuse impairment in the development of myelin in the brain associated with a firm toughness of the white matter of the cerebrum and cerebellum. The disorder shows a predilection for the fibre tracts that normally should myelinate after birth and tends to spare the arching (arcuate) fibres in the cerebral cortex. The affected tissues contain granular material that stains (with toluene blue) various shades of red in the early lesions and golden brown in the older lesions. Accumulation of this pigmented material results from a defect in the metabolism of fatty substances called sulfatides, appearing in the nerves, pituitary gland, liver, testes, and kidneys and excreted in the urine. Signs of involvement of the nervous system may be manifest in infancy, childhood, or early adult life in the form of convulsive seizures, optic atrophy, paralyses, and such mental symptoms as dementia and emotional abnormalities. The disease rapidly progresses to death within a few years of onset. Satisfactory treatment remains to be developed.

CRANIOCEREBRAL TRAUMA

Injury to the head is one of the common causes of death in young adults. Severity of symptoms depends on the extent of the damage to the brain rather than on the presence or absence of bone fracture. Survivors of head injury may show convulsive seizures, psychosis or other mental disturbances, and a constellation of other symptoms (posttraumatic syndrome).

The diagnosis of a head injury is not difficult except when the victim is found in unexplained coma. The head is carefully examined for lacerations (torn wounds) or contusions (bruises); the ears and nose are inspected for bleeding. X-rays are taken to reveal possible fractures, and an echoencephalogram may be employed to show any displacement of structures within the skull. Electroencephalograms are taken routinely at many hospitals in cases of coma. Lumbar puncture is also of value in assessing the severity of any injury; *e.g.*, through examination of the spinal fluid for the presence of blood. Pneumoencephalography or angiography may be needed to establish the diagnosis of extradural, subdural, or large intracerebral hemorrhages.

Concussion

The mortality rate is almost nil in sufferers of simple concussion (blows to the head) and less than 2 percent if the injury produces only a minor degree of cerebral edema or congestion. The mortality rate rises when the brain cortex itself is contused (5 percent) and increases sharply when the brain is lacerated (40 percent). Death may occur in the period immediately after the injury, or it may be delayed for several weeks.

The course of progress in victims of head injury is related to the severity of damage to the brain and to the presence of complications. In general, there is a rapid return of consciousness in those who have incurred minor injury (simple concussion). Usually there is impairment of memory for the immediate period prior to the injury (pretraumatic amnesia) and the period directly following recovery of consciousness (posttraumatic amnesia). Headaches, dizziness, or vertigo are usually present after consciousness is regained. These symptoms usually disappear in a few weeks, but not infrequently they may endure as long as several months.

With severe injuries the period of coma may be prolonged, and there is a period of confusion and mental clouding before full consciousness is restored; and there may be signs of focal damage to the brain or of cranial-nerve palsies. With the passage of time, considerable improvement in symptoms is the rule, although enduring sequelae (*e.g.*, convulsive seizures, paralysis, and mental impairment) are not uncommon.

Treatment of craniocerebral injuries includes superficial cleaning and closing of wounds, removal of fragments of bone and dead tissue, and elevation of caved-in parts of the skull from underlying soft tissues. Nonsurgical treatment chiefly involves general care and control of increased intracranial pressure (*e.g.*, with such medications as corticosteroids).

Complications of head injuries. *Extradural hemorrhage.* Leaking of blood into the extradural space occurs in less than 1 percent of cases as a result of torn meningeal arteries (usually the middle meningeal). The mass of blood (hematoma) is often quite large, commonly extending over part of one hemisphere in the middle fossa (hollow) of the skull. The typical sequence is loss of consciousness at injury, a lucid interval of several hours, followed by relapse into coma and hemiplegia. In the ab-

sence of this sequence of events, angiography may be required to detect the hematoma. Death almost invariably ensues unless the clot is removed surgically.

Subdural hemorrhage. Bleeding into the subdural space comes from rupture of arachnoid vessels or of small vessels extending between the arachnoid and dural membranes (meninges). It occurs in about 2 percent of people who incur trauma to the head. The frequency of subdural hemorrhage is related to the severity of the injury, although it may occur in apparently mild injuries. Subdural hemorrhage is suspected when there are daily fluctuations in consciousness or when recovery fails to occur as expected. Echoencephalography, repeated electroencephalograms, and angiography can help in diagnosis; treatment consists of surgical removal of the clot.

Subarachnoid hemorrhage. Hemorrhage into the subarachnoid space is present in practically all cases of head injury of moderate or severe degree. Usually of no great consequence, the clot is generally absorbed spontaneously.

Intracerebral hemorrhage. Pinpoint hemorrhages, or small leaks of blood into the substance of the brain, are common in persons with severe injuries and do not require surgical therapy. The debris of large hemorrhages typically is removed surgically.

Infections. Infections of the bone, leptomeninges (purulent meninges), or intracerebral abscess are rare complications of head injury. Their diagnosis and treatment are the same as for other similar infections.

Cranial-nerve palsies. Damage to the cranial nerves occurs in about 5 percent of head injuries. The olfactory, facial, and acoustic nerves are most commonly affected (see above). Complete or partial recovery is the rule except for lesions of the olfactory nerve.

Focal cerebral lesions. Hemiplegia, aphasia, or other signs of localized (focal) damage to the brain may occur in sufferers of severe head injuries. The symptoms are usually related to contusion and laceration of the brain or to small hemorrhages into the substance of that organ.

Sequelae (residual symptoms) of head injuries. *Epilepsy.* Convulsive seizures occur in approximately 2 percent of those who suffer head injury. The incidence is much greater (more than 40 percent) when the brain itself is injured (*e.g.,* by penetrating objects such as shrapnel) or when the dura and arachnoid membranes are torn by compound fractures of the skull. Convulsions may occur immediately following injury but more commonly appear months or years afterward. They are similar to epilepsies arising from other causes except that focal seizures (*e.g.,* involuntary leg or arm twitching) are more common than are generalized convulsions. The treatment is the same as that for seizures from other causes and may include the use of sedatives and other drugs.

Major psychiatric disturbances. Transient psychotic episodes and some impairment of mental faculties are not uncommon after severe trauma to the brain, but long-continued psychotic episodes are rare. They account for less than 1 percent of admissions to psychiatric hospitals.

Posttraumatic syndromes. Headaches, dizziness, irritability, restlessness, inability to concentrate, increased sweating, depression, and other personality changes occur in more than one-third of victims of head injury. These symptoms do not seem to be related to the severity of injury to the brain, being attributed to minor derangement of cerebral function or to psychological reactions. Typically appearing shortly after the patient is allowed out of bed, the symptoms tend to decrease in severity and to disappear after three to six months.

Personality changes

Reassurance of the sufferer that the symptoms are not due to serious injury to his brain and that they will ultimately disappear is most beneficial. He is encouraged to participate in social and recreational activities and to return to appropriate work as soon as his symptoms permit. Mild sedatives and analgesics (*e.g.,* aspirin) are often of value in posttraumatic syndromes.

EPILEPSY (RECURRENT CONVULSIVE SEIZURES)

Epilepsy is a paroxysmal disorder of the nervous system recognized by recurring attacks of loss or impairment of consciousness with or without convulsive movements.

Classification. One way to classify sufferers of convulsive seizures is to consider those with demonstrable neural damage (symptomatic epilepsy) as essentially different from the rest (idiopathic epilepsy). This controversial view presumes no physical lesion as a basis for idiopathic epilepsy, almost as if the inherited tendency found for the disorder had no physiological repercussions in the nervous system.

More objective systems classify epilepsies according to the manifestations observed during seizures. This method serves to indicate the location of abnormal tissues (the neural focus of symptoms) and to show how their influence spreads in the nervous system. Distinctions of this sort are indeed of value in selecting treatment methods for each individual case.

Causes of convulsions. Convulsions may result from acute lack of blood sugar or oxygen for any reason, from disturbances of calcium and phosphorus metabolism and other metabolic disturbances, and from drugs and poisons. The common occurrence of seizures in patients with brain disease indicates that damage to cerebral tissue is the most common cause of what is called epilepsy.

The extensive list of diseases frequently accompanied by seizures includes developmental and congenital defects, birth injuries, infectious diseases of childhood, meningitis, encephalitis, cerebral trauma, tumours, abscesses, and degenerative diseases of the nervous system. Attacks may not appear for several or many years after initial damage to the brain.

Occurrence. Approximately 0.5 percent of people are subject to recurrent convulsive attacks. Males tend to be affected more frequently, and so-called idiopathic attacks commonly first appear before age two or at about the time of puberty. These peaks are related to congenital defects, birth injury, and diseases of childhood. In cases with onset in adult life, head trauma, brain tumours, and degenerative diseases contribute to the symptoms.

Symptoms. The characteristics of convulsive seizures are related to the site of the abnormal focus in the brain and to the rapidity of its neural discharge. They may be classified as grand mal (generalized) seizures, focal seizures (*e.g.,* Jacksonian epilepsy), generalized convulsions growing out of focal onset, petit mal (minor) attacks, psychomotor (partial) attacks usually related to damage to the temporal lobe of the brain, and epileptic equivalent (*e.g.,* outbursts of rage without convulsion). Grand mal or grand mal seizures with focal onset are the most common types of attack. Petit mal seizures are almost exclusively found in children; alternations in types of seizures displayed by the same person are common.

Characteristics of convulsive seizures

Diagnosis. Sufferers of convulsive seizures are routinely examined by roentgenograms of the skull, electroencephalography, brain scans, and studies of the cerebrospinal fluid. Pneumoencephalography or angiograms ordinarily are avoided unless brain tumour appears to be the cause of the convulsions. The electroencephalogram is especially valuable in classifying the type of seizure through its brain-wave patterns, in localizing focal abnormalities, and in following the course of the disorder (*e.g.,* in evaluating the effects of treatment).

Treatment. Epileptic seizures are treated by removing known precipitating factors, by methods of physical and mental hygiene, and by using anticonvulsant drugs. Removable causative factors (*e.g.,* operable brain lesions) are found in only a small percentage of cases, and treatment rests heavily on medication. Considerable care is taken in each individual case to find the best drug or combination of drugs and the proper dosage of each and to enlist the sufferer's cooperation in taking them and in avoiding such precipitating factors as alcohol and emotional stress. In grand mal, focal, or psychomotor seizures, such anticonvulsants as sodium diphenylhydantoin and sedatives (*e.g.,* phenobarbital), alone or in combination, are usually most effective; other specific drugs (*e.g.,* ethosuximide) are preferred in petit mal epilepsy.

Anticonvulsants and sedatives

TOXINS

Many poisons (toxins) affect the nervous system, including those produced by infectious micro-organisms and

viruses, compounds of many metals (*e.g.*, arsenic, lead, manganese, mercury, and thallium) and other substances used in industry, and many drugs (see also POISON AND POISONING; POISONOUS ANIMALS AND PLANTS; DRUG AND DRUG ACTION).

Bacterial toxins. *Diphtheria.* The bacterium that causes diphtheria produces a toxin that spreads from the site of infection in the throat or skin and affects the nerves that serve the muscles of the palate, larynx, and eyeball, also producing generalized polyneuritis. Widespread immunization and early use of effective treatment when infection occurs have made the neural complications of diphtheria rare in many countries.

Tetanus (lockjaw). The toxin produced by lockjaw bacteria leads to localized and generalized spasms of the muscles and convulsive seizures. The symptoms commonly appear about five to ten days after a wound has been contaminated by the organisms that cause tetanus. Widespread immunization of children, military personnel, and farmers and prophylactic (preventive) treatment (with tetanus toxoid or antitetanic serum) of any wounded person have reduced the number of cases. Narcotic addicts are particularly likely to develop lockjaw when they use contaminated hypodermic needles or other objects that break the skin.

Treatment of full-blown cases of tetanus is difficult and not reliably successful. Intravenous injection of antitetanus serum, administration of tranquillizers, general supportive measures, regulation of salt and water intake, and the use of such drugs as penicillin have lessened the mortality rate of the disease.

Botulism. The botulism bacterium makes a toxin that causes weakness of the muscles throughout the body. The toxin is readily destroyed by heat, and poisoning usually results from eating poorly cooked or canned foods (*e.g.*, string beans) contaminated with the bacteria. Symptoms appear within 12 to 24 hours after eating the tainted food and may be preceded by nausea, vomiting, or diarrhea. Difficulty in fixing (converging) the eyes on objects is usually the first symptom, followed by drooping of the eyelids, dilation of the pupils, paralysis of the eyeball muscles, difficulty in swallowing, and weakness of the muscles of the trunk and extremities. The mental faculties are preserved, but convulsions and coma may develop toward the end in fatal cases. Complete recovery is the rule, however, when small amounts of the toxin are eaten. In severe botulism, death typically occurs in four to eight days as a result of respiratory and circulatory failure. Botulism can be prevented by taking proper precautions in the preparation of canned foods and by discarding (without sampling) foods with a rancid odour in which gas has formed. Treatment consists of pumping out the stomach and giving a specific antitoxin.

Metallic poisoning. When metallic poisons are taken, neurological symptoms may or may not be preceded by digestive upsets.

Arsenic. Chronic ingestion of small doses of arsenic in any form causes a generalized polyneuritis that is usually not fatal; ingestion of large amounts or the injection of arsenical drugs may result in fatal hemorrhagic encephalitis. The specific treatment of arsenic poisoning involves the injection of a substance called British anti-lewisite (BAL).

Lead. Lead poisoning usually results from industrial exposure to the metal or by eating lead compounds (as in some house paints). The neurological complications of lead poisoning include polyneuritis with weakness of the forearm muscles (wrist drop) and what is called lead encephalopathy. The latter most commonly occurs among children who eat lead paint from walls or furniture and is characterized by lethargy, convulsions, coma, and increased intracranial pressure; lead encephalopathy has a mortality rate of about 20 percent. Mental retardation or recurrent convulsive seizures may be present in children who recover. Treatment includes measures to control seizures, efforts to reduce intracranial pressure, and the use of drugs (*e.g.*, edathamil disodium calcium).

Lead encephalopathy

Manganese. Chronic poisoning with manganese occurs most often in miners. The basal ganglia are mainly affected, producing a syndrome similar to that of Parkinson's disease. Indeed, as in Parkinsonism, the symptoms respond to levodopa (see above).

Thallium. Accidental ingestion of any form of this metal in rat poisons or by its absorption through the skin in some treatments for ringworm or from certain hair removers (depilatories) can produce thallium poisoning. Optic neuritis and generalized polyneuritis are the most common neurologic symptoms. Death usually occurs when large amounts of the metal enter the body, although the chances of survival are good in chronic poisoning. The polyneuritis usually clears up entirely, the most common of the sequelae being some reduction in visual acuity.

NEUROMUSCULAR DISORDERS

Weakness and atrophy of muscles frequently reflect diseases associated with lesions in the spinal cord, peripheral nerves, or the muscles themselves.

Amyotrophic lateral sclerosis. Of unknown origin, amyotrophic lateral sclerosis is characterized by muscular atrophy and fibrillation (trembling) associated with degeneration of the motor neurons in the spinal cord and medulla oblongata. Relatively common (five persons per 100,000), the disease affects men about twice as frequently as women; onset of symptoms usually is between the ages of 40 and 60.

A number of symptom combinations may develop. If the motor (ventral horn) cells in the spinal cord are most severely affected, the symptoms are those of progressive spinal muscular palsy; when the medulla oblongata bears the brunt, the syndrome of progressive bulbar palsy develops; if the upper motor neurons in the brain cortex are affected early, the syndrome may be described as primary lateral sclerosis.

In most cases there is weakness and atrophy of muscles, with hyperactive tendon reflexes and muscular irritability. In later stages (with severe atrophy of the muscles), normal reflexes are lost, the only evidence of corticospinal disease being extensor plantar (Babinski) response. Even the Babinski reflex fails to appear if the muscles of the toes are too atrophied to respond.

The course of amyotrophic lateral sclerosis is chronically progressive, showing paralysis of the muscles of speech and swallowing in its late stages. Death is usually from paralysis of the muscles of breathing or associated pneumonia. The average duration of the disease is about three years; when the site of initial lesion is the spinal cord, survival of ten or more years is observed in occasional cases. The course is rapid when the nuclei in the medulla oblongata are involved at the onset.

Progressive muscular dystrophy. Of hereditary origin, this disease is characterized by diffuse weakness and wasting of the muscles during early childhood. It has a frequency of occurrence of four cases per 100,000 people. Structural changes include degeneration of the muscles and replacement of muscle substance by connective tissue and fat. The child becomes clumsy in walking and has difficulty in climbing up and down stairs and in lifting heavy objects. Enlargement of weak muscles (pseudohypertrophy), when present, is most common in the calf muscles. Tendon reflexes usually disappear, while sensory and mental development remain normal.

There is a decrease in the excretion of creatinine (a substance deriving from muscle tissue) proportional to the loss of muscular mass and a drop in the ability of the body to store a related substance called creatine (*e.g.*, from meats).

As a rule, there is a gradual progression of the weakness, especially in the Duchenne form of muscular dystrophy, in which the child is usually incapacitated within five to 15 years. Death stems from associated infections or from involvement of the respiratory or heart muscles. In the other forms of muscular dystrophy, progress is slower, and survival into adult life (even to middle age) is not uncommon. No treatment is known to affect the progress of the disease.

Duchenne form of dystrophy

Myotonia congenita. "Myotonia" describes muscles characterized by prolonged contraction and slowness to

relax, particularly noticeable during efforts to make quick or forceful movements; *e.g.*, asking the person to forcefully grasp the examiner's fingers. Myotonia can also be observed by noting the mound of muscles at the base of the thumb when that place is tapped with a rubber hammer. Congenital myotonia is usually localized in a few muscles of the body, but many may be involved in some cases. A possible clue to the physiological basis of myotonia is that it tends to diminish in response to the administration of quinine and to increase when a drug called neostigmine is taken.

Myotonic muscular dystrophy. This hereditary disorder is characterized by myotonia and atrophy of the muscles (especially those of the face and neck) and by cataracts, early baldness, and atrophy of the gonads.

The disease is less common (by about 25 percent) than is progressive muscular dystrophy and tends to appear more often in males. Transmission of the hereditary tendency to show the full-blown disease is rare, because early gonadal atrophy prevents many sufferers from having children. Symptoms in the parents of most sufferers are usually confined to early development of cataracts (opaque lenses in the eyes). Weak facial muscles produce a characteristic hatchet face; the voice may be nasal; and there may be some slight difficulty in swallowing. The myotonia usually makes it difficult for the afflicted person to relax the hand muscles when an object is grasped. When the trunk muscles are involved, the sufferer may fall when making sudden movements and have trouble getting up. Mental retardation is a frequent concomitant, a number of victims being found in institutions for defectives.

Sufferers of myotonic muscular dystrophy rarely become badly incapacitated. The myotonia tends to lessen with the years, possibly as the victim learns to compensate by being careful to move more slowly and less emphatically. There is no known treatment for the muscular weakness or atrophy, although the myotonia responds to such drugs as quinine, diphenyhydantoin (also used in epilepsy), and procaine amide.

Myasthenia gravis. Characterized by weakness and abnormal fatigability of the striated muscles, myasthenia gravis is probably the result of a defect in the utilization of a chemical (acetylcholine) at the junctions between muscles and nerves. The thymus gland in the chest is enlarged or is the site of a tumour in half of the cases, and some abnormality (lymphoid hyperplasia) of the thymus is almost always present.

Occurring in about two persons of every 100,000, the onset of symptoms of myasthenia gravis may be at any age but is most common in the second or fourth decades. When symptoms begin before age 40, females are much more frequently affected, after which there is a slight preponderance of male sufferers.

Symptoms of myasthenia gravis

There is a characteristically progressive weakening of the muscles with each succeeding contraction, with strength returning after rest. Usually all of the muscles are affected equally: weakness may be confined to the trunk and extremities and rarely only to the eye muscles. The course may be punctuated by temporary improvement and worsening; more commonly death results in a few years unless treatment includes compounds such as neostigmine that enhance the effects of acetylcholine. Beneficial effects have followed the use of corticosteroids and removal of the thymus (particularly in young females).

Familial periodic paralysis. A rare and poorly understood disease, this disorder is characterized by recurrent attacks of weakness or paralysis of the voluntary muscles and loss of deep reflexes; affected muscles fail to respond even to direct electrical stimulation. Men are affected more than twice as often as women. In the hyperkalemic form, attacks are related to a fall in the potassium level of the blood. The first attack usually occurs about the time of puberty but may appear years before or after.

Attacks commonly begin after a period of rest, varying from a slight weakness to complete paralysis, usually of the muscles of the trunk and extremities, that lasts from a few minutes to a full day. The interval between attacks may be as long as a year, or several brief episodes may occur daily, potassium level returning to normal when symptoms stop. The short duration of many attacks makes it difficult to evaluate any therapy, but potassium salts are commonly given. Prophylactic treatment includes restriction of carbohydrate intake and the administration of potassium chloride by mouth.

Another (hyperkalemic) form of the disease has an early age of onset, tends to show milder and shorter attacks, and is associated with increased blood-potassium levels. Calcium gluconate and insulin are used in the treatment of an attack, while daily doses of dextroamphetamine are recommended as prophylaxis.

Myotonia (particularly of the upper eyelids) in association with familial periodic paralysis has been called paramyotonia congenita. In such cases the myotonia and the paralysis can be precipitated by immersing an affected limb in cold water or by the application of cold compresses to the eye.

This disorder and those already mentioned throughout this article represent only the better known fraction of nervous system diseases. Others are considered in such articles as HEADACHE; PAIN, THEORIES OF; and PSYCHIATRIC TREATMENT, CONCEPTS OF.

BIBLIOGRAPHY. More detailed discussions of diseases of the nervous system may be found in H.H. MERRITT, *Textbook of Neurology*, 4th ed. (1967); WALTER RUSSELL BRAIN, *Diseases of the Nervous System*, 7th ed. (1969); and W.R. GOWERS, *A Manual of Diseases of the Nervous System*, 2nd ed. rev., 2 vol. (1970, reprint of the 1893 edition).

(H.H.M.)

Nestorius

Nestorius was an early bishop of Constantinople whose views about the nature and Person of Christ led to a fierce controversy, resulting in the calling of the third ecumenical council to Ephesus in AD 431. The ensuing debate produced greater doctrinal clarification, on the one hand, and the creation of separate Nestorian churches, which still exist, on the other.

Birth and early career

Nestorius was born of Persian parents in the last quarter of the 4th century AD at Germanicia in Syria Euphratensis. Trained at Antioch, where he was ordained presbyter, he may have been a pupil of Theodore, bishop of Mopsuestia. He entered the nearby monastery of St. Euprepius and gained a wide reputation as a preacher. It is recorded that he had an excellent voice, was fluent of speech, and possessed fine eyes and red hair.

In 428, when the emperor Theodosius II invited him to become bishop of Constantinople, the situation in the capital was disturbed. An interregnum had followed the death of the previous bishop, Sissinius, and an unresolved rivalry between two candidates, Philip of Side, a church historian, and Proclus, who was himself to become bishop of Constantinople in 434. There was a large number of heretical groups whose existence affronted the orthodox. Preaching at his consecration on April 10, Nestorius declared: "Give me, O emperor, the earth purged from heretics, and I will repay you with heaven." On May 30 Theodosius issued an edict against heretics, and Nestorius was zealous in their suppression.

Nestorius had brought with him as his domestic chaplain a man named Anastasius, and on November 22 this man preached a sermon in which he objected to the title Theotokos (Mother of God) as applied to Mary. Many were scandalized, for the term had long been in use. Nestorius supported Anastasius and, on Christmas Day, began a series of addresses arguing that Mary was not Theotokos. He believed, incorrectly, that the title represented a recrudescence of Arianism—a belief, first propounded early in the 4th century by Arius, that Jesus was a creature and only semidivine.

To many people it seemed that Nestorius himself was denying the divinity of Christ and regarding him as a mere man who had been adopted by God as his son. He was charged with reviving the heresy of Paul of Samosata, bishop of Antioch *c.* 260–270. It was at this juncture that Cyril, bishop of Alexandria 412–444, chose to intervene and to set in motion the series of events that finally resulted in Nestorius' downfall.

Copies of Nestorius' sermons were circulating abroad and had given rise in Egypt to some dissension in the monasteries. Cyril, therefore, issued a circular letter, shortly after Easter 429, in which he attacked Nestorius' views. Cyril's motive was twofold. He sincerely believed that Nestorius was undermining the purity of the faith; Nestorius' refusal to style Mary as Mother of God seemed to Cyril to deny the unity of the God-man. Further, as bishop of Alexandria, Cyril was eager to belittle the rival see of Constantinople, as his immediate predecessor, Theophilus, had done in the case of John Chrysostom. Theological and political considerations were thus combined.

Toward the end of the summer, Cyril wrote to Nestorius, calling on him to admit his errors. Upon receiving only a curt acknowledgment and learning that Nestorius was supporting certain of his own clergy whom he had disciplined and who had fled to the capital, he sought the assistance of the imperial family, having previously approached Celestine, bishop of Rome (422–432). Celestine himself was not ignorant of the situation nor well disposed toward Nestorius, who had also received some of the followers of Pelagius—the denier of the doctrine of original sin—previously condemned in Rome. Celestine sent copies of Nestorius' writings to John Cassian, a leading theologian and one of the prime organizers of Western monasticism, asking him to submit a report. He then held a council in August 430 that condemned Nestorius.

Condemnation of Nestorius

In November Cyril summoned his bishops to Alexandria and wrote to Nestorius requiring him to anathematize 12 propositions on pain of excommunication. When his envoys reached Constantinople, they found that they had been forestalled. The Emperor had summoned a council to meet at Ephesus by Pentecost 431. It was Nestorius' hope that the result would be the condemnation of Cyril on the charges made against him by the Egyptian refugees.

Although the council was to begin on June 7, there was delay because of the late arrival of certain delegates—in particular, John, bishop of Antioch (428–441), who supported Nestorius. Cyril, having the ear of the local bishop, Memnon, decided to press on, as he feared John's influence, and so the first session opened on June 22. Nestorius refused to attend and was condemned in his absence. Four days later John reached Ephesus and proceeded to hold a countersession that itself deposed both Cyril and Memnon. On July 10 the three Roman legates representing Celestine arrived; Cyril again seized the initiative, and the condemnation of Nestorius was renewed. Both sides appealed to the Emperor. Intrigue followed upon intrigue, until Theodosius, influenced by his sister Pulcheria (whom Nestorius had offended), relegated him to his former monastery near Antioch, and the council was dissolved.

Nestorius remained in the monastery for four years (431–435) and spent his time writing about his case in a work entitled *Tragedy*, of which only fragments survive. He was not in a position to have much influence. Controversy continued, but gradually, in the interests of church unity, the "Orientals" (*i.e.*, the supporters of John of Antioch) and the Cyrillian party came to an understanding, the doctrinal issue being separated from the personal question of Nestorius' own fate. His continued presence near Antioch became an embarrassment, and John eventually recommended to the Emperor that he should be exiled. An imperial order was issued sending him to Petra in Arabia. Soon this was considered too lenient and he was transferred to the Great Oasis (now the Oasis of Khārijah) in the Libyan Desert (*c.* 436–437). Later still, marauders carried him off to Panopolis (Akhmīm), whence he was removed to Elephantine opposite to Syene (Aswān) and then back to Panopolis.

During his exile, Nestorius was not completely isolated. He received visitors and correspondence. He followed the course of the Eutychian controversy, a reaction to his own teaching, which involved the assertion that the humanity of Christ was swallowed up by his divinity and so ceased to exist. He wrote to the citizens of Constantinople in support of their bishop, Flavian (446–449), and of Leo, bishop of Rome (440–461), and saw in the treatment of Flavian by Cyril's successor, Dioscurus, a repetition of what he had suffered. He devoted time to a work that was to be a defense of his teaching and a history of his life. He entitled it the *Book of Heraclides of Damascus*, using a pseudonym to have it published. The sole treatise from his pen to have survived, it was discovered in 1895 in a Syriac translation.

Nestorius survived the death of Theodosius on July 28, 450, and it seems likely that the emperor Marcian had decided to allow him to return to civilization, but the news arrived after his death. Whether his death occurred just before, or shortly after, the opening of the Council of Chalcedon in 451 is not known.

Doctrines and significance

Nestorianism, as it was understood at the time, so insisted upon the full humanity of Christ's human nature that it was believed to divide him into two Persons, one human and the other divine. Nestorius repudiated such a view, and Cyril's interpretation was based on a misunderstanding.

What he actually taught was a prosopic union. The Greek term *prosōpon* means the self-manifestation of an individual that can be extended by means of other things —*e.g.*, a painter includes his brush within his own *prosōpon*. So the Son of God used manhood for his self-manifestation, and manhood was, therefore, included in his *prosōpon*, so that he was a single object of presentation.

While Nestorianism was crushed within the Roman Empire, it survived outside its frontiers. The Persian Church adopted it, largely to obtain the protection of their rulers by assuring them that their religion was not that of their enemies the Romans. It has continued until the present day, although its adherents are few in number, with groups being found in Iraq, the U.S.S.R., India, Iran, Syria, and North and South America.

BIBLIOGRAPHY. B.J. KIDD, *A History of the Christian Church to A.D. 461*, 3 vol. (1922), draws together the many known details into a comprehensive narrative. F. LOOFS, *Nestoriana* (1905), contains the majority of the surviving fragments of Nestorius' works. G.R. DRIVER and L. HODGSON (eds.), *Nestorius, The Bazaar of Heracleides: Newly Translated from the Syriac* (1925); and F. NAU, *Nestorius: Le livre d'Héraclide de Damas* (1910), both give translations of Nestorius' major extant work and provide full introductions and notes. A.R. VINE, *An Approach to Christology* (1948), discusses Nestorius' Christology. J. QUASTEN, *Patrology*, vol. 3, pp. 514–519 (1963), provides a comprehensive bibliography.

(J.G.D.)

Netherlands, The

One of the smaller countries of Europe, the Kingdom of The Netherlands is a parliamentary democracy under a constitutional monarch. The Netherlands has often been known in English as Holland (literally Hollow Land). This name, which was originally given to the medieval core of what later became the state and which is still used for two of its 11 provinces (Noord-Holland and Zuid-Holland), accurately reflects the country's character. The irregular outline of The Netherlands, not unlike a piece of a jigsaw puzzle, encloses some 15,892 square miles (41,160 square kilometres) of extraordinarily flat land lying between the North Sea on the north and west, the Federal Republic of Germany on the east, and Belgium to the south. The kingdom incorporates the former colonies of Surinam and The Netherlands Antilles as well.

Some 3,000 square miles of this land have been wrested from the sea by a process of careful water management, drainage, and reclamation that dates back to the 12th century, although the really large schemes were initiated in the 20th century. Much of the Zuiderzee, the great arm of the sea that formerly tongued into the country from the north, has been turned into *polderland*, land surrounded by dikes and artificially drained. The Waddenzee, the northern seaward sector, is now separated from the larger southern portion, known as IJsselmeer (Lake IJssel), by a huge barrier dam almost 20 miles long. Plans for reclaiming the Waddenzee itself are maturing, as is an equally ambitious project to cordon off the sea

The struggle against the sea

from the islands and deltas of the southwestern province of Zeeland.

The Netherlands was the home of more than 13,000,000 people in the early 1970s. Despite a government-encouraged emigration after World War II, which sent some 500,000 persons out of the country (mostly to the English-speaking countries of North America and the traditionally white nations of the then British Empire and Commonwealth), the country is today one of the most densely populated nations in the world. Partly as a result of the earlier emigration, about one-tenth of the citizens are now over 65 years old, but this fact has not prevented the capital, Amsterdam, from becoming one of the liveliest centres of the international counter-culture of Western youth. This independence of outlook, if it can be thus interpreted, has very strong roots in The Netherlands. Citizens of the 16th and 17th centuries rose time and again against foreign domination to found and secure their own republic. Dutch merchant ships sailed the world and helped to lay the foundations of a great trading nation characterized by a vigorous spirit of enterprise. The Netherlands contains the estuaries of the rivers Rhine, Maas, and Scheldt, and through their seaports the Dutch control a large share of the world's waterborne traffic. All this is a far cry from the phlegmatic composure that other Europeans have sometimes ascribed to the Dutch, though something of the traditional image of the country may be preserved in the neat, brick-built little towns, the countless windmills, and the magnificent colours of the spring tulip fields, even if the famous peasant dress—wooden clogs, the full-bottomed trousers of the men, and the white caps and wide skirts of the women—is not so widespread today.

With Belgium and Luxembourg, The Netherlands is a member of the economic union known as Benelux, in which foods, goods, and people can freely circulate; this union, in fact, served as a model for the larger European Economic Community (EEC), of which the Benelux nations are members. Since the formation of Benelux, production in The Netherlands has risen in quality and technical refinement. The country is also a member of the North Atlantic Treaty Organization (NATO) and the Organization for Economic Co-operation and Development (OECD). Of the former Dutch colonies, the Dutch East Indies became independent as Indonesia in 1949, while Surinam and the Netherlands Antilles were united with the European Netherlands in 1953, under the Statute of the Kingdom of The Netherlands. (For related information, see AMSTERDAM; HAGUE, THE; LOW COUNTRIES, HISTORY OF; and ROTTERDAM.)

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Relief features.* The landscape of The Netherlands is almost entirely flat. It is characterized by dunes and dikes protecting the most densely populated part of the country, which would otherwise be inundated. Indeed, 38 percent of the country would be swallowed by the sea if there were no man-made dikes. Villages and estates are surrounded by dikes and a system of drainage canals, while about half of the land area consists of polders, which intensify the flat appearance of the land.

There are several hundred polders, some made possible by the operation of windmills, which, despite their limited power, carried out the drainage in earlier times. But the larger polder areas, such as those of IJsselmeer, have been created by steam, diesel, and electric pumps. Fen, river, and sea polders lie on peat land along the rivers, but some polders above sea level, created by silting, are drained of excess water by sluice gates at ebb tide. The western part of The Netherlands lies below sea level, sometimes as much as 22 feet (6.7 metres) below. On such land, building is only possible after driving piles down to the sand layer.

Limburg, in the border zone of the Ardennes, is the only area that is not part of the lowland. Indeed, in the entire country the only relief contrasts are provided by the Torenberg, in the Veluwe region (340 feet [104 metres] high); a 315-foot (96-metre) elevation around Nijmegen; and the highest point in The Netherlands, the Vaalserberg, rising 1,050 feet (321 metres) in the south of Limburg.

The Zuiderzee

The Zuiderzee was originally an estuary of the Rhine. It then became a shallow inland sea, biting deep into the land, and eventually was hollowed into an almost circular shape by the action of winds and tides. Only in 1932 did the barrier dam, running northeastward to connect the province of Noord-Holland with that of Friesland, finally tame and divide the sea into the outer Waddenzee and the inland IJsselmeer. In the latter, four giant polders were constructed around a freshwater basin fed by the IJssel River and linked with the sea by sluices. Ninety percent of the land thus reclaimed is arable, and it is parcelled out in large plots that are rented to farmers. After World War II, additional development made the polders available for industrial, residential, and recreational purposes as well as for agriculture.

Further south, the disastrous spring tide of February 1, 1953, which flooded 400,000 acres of land and killed 1,800 people, called into existence the Delta Plan, which aims to close the sea inlets of the southwestern province of Zeeland. These huge delta works will shorten the coastline by 450 miles, combat the salination of the soil, and allow the development of the islands of Zeeland. The Oosterschelde (Eastern Scheldt) is already dammed; large sluices dominate the Haringvliet (river), and the embankment dikes along the Westerschelde (Western Scheldt) have been strengthened. Though future spring tide catastrophes may be avoided, the enterprise will doom the oyster culture and the shrimp and mussel fishing.

Glacial, riverine, and marine deposits

Soil. After the Ice Age (about 10,000 years ago), a large area in the north of what is now The Netherlands was covered by moraine (the glacial accumulation of earth and stones). In the centre and south, the Rhine and Maas unloaded thick layers of sand and gravel from the European mountain chains. Behind the coastal dunes, sea clay was deposited in the sheltered lagoons, and peat soil often subsequently developed in these areas. If the peat soil was washed away, sea clay was deposited again, and the alluvial layer had to stabilize through drainage and settlement.

Climate. The climate is temperate, with gentle winters, cool summers, and rainfall in every season. Southerly and westerly winds predominate, and the sea moderates the climate through onshore winds and the effect of the Gulf Stream.

The position of The Netherlands—between the area of high-pressure air masses centred on the Azores and the low-pressure region centred on Iceland—makes the country an area of collision between warm and polar air masses in summer, thus creating unsettled weather. In springtime, on the other hand, northern winds often cause night frosts. Winds meet with little resistance over the flat country, though the hills in the south diminish by more than half the 13-mile-per-hour wind velocity that prevails along the coast. Frost averages 60 days per year. July temperatures average about 63° F (17° C), and those of January average 35° F (2° C). The rainfall averages 31 inches (790 millimetres), with only about 25 clear days per year. The average rainfall is highest in summer (August) and autumn and lowest in springtime. The country is known—not least through the magnificent landscapes of Dutch painters—for its heavy clouds, and on an average day three-fifths of the sky is clouded.

Vegetation and animal life. The fauna of The Netherlands has been greatly affected by the activities of man; thus, the badger, hawk, river lobster, and marsh turtle have disappeared, while various fish species and the boar have been introduced, and the red-grouse population has been nurtured.

The Netherlands possesses the characteristic flora of the Atlantic district of the Euro-Siberian vegetation province to which it belongs. The flora is dominated by plants of the dunes (salt plants and water plants) and by plants of the woods, of cultivated land, and of meadows. In addition, the cultivation of ornamental plants, flowers, and bulbs has become a major industry.

The growth of urban centres

Urban and rural settlement. Urbanization in The Netherlands has taken place mainly in the 20th century. In 1900, over half the population was still living in villages or towns of fewer than 10,000 inhabitants. In 1972 this figure was only a little less than 20 percent. The population of the cities and towns is growing much faster than the population of the rural areas, for migration to the towns more than exceeds the rural areas' higher birthrate. The 1960s nevertheless saw a decrease in the downtown population of the large cities. These inner-city regions are now becoming economic and cultural centres, their populations having spread outward in search of newer housing and greater living space in the suburbs. Urbanization is concentrated in areas of industrial growth, especially around the harbours of the Nieuwe Waterweg (New Waterway) centred on Rotterdam or the Noordzeekanaal (North Sea Canal) centred on Amsterdam.

The bulk of the Dutch population is concentrated in the urban circle (known as the Randstad) comprising Rotterdam, The Hague, Leiden, Haarlem, Amsterdam, Hilversum, and Utrecht; in Noord-Brabant (in the three main cities of Breda, Tilburg, and Eindhoven); and in the Twente District around the city of Enschede. It is established policy to keep traditional towns and cities separated by strips of agricultural or recreational land. Even in the Randstad, public authorities try to maintain a "green heart."

THE PEOPLE

Ethnic and religious makeup. In origin, the Dutch people are a mixture of Frisians, Saxons, and Franks.

MAP INDEX

Political subdivisions

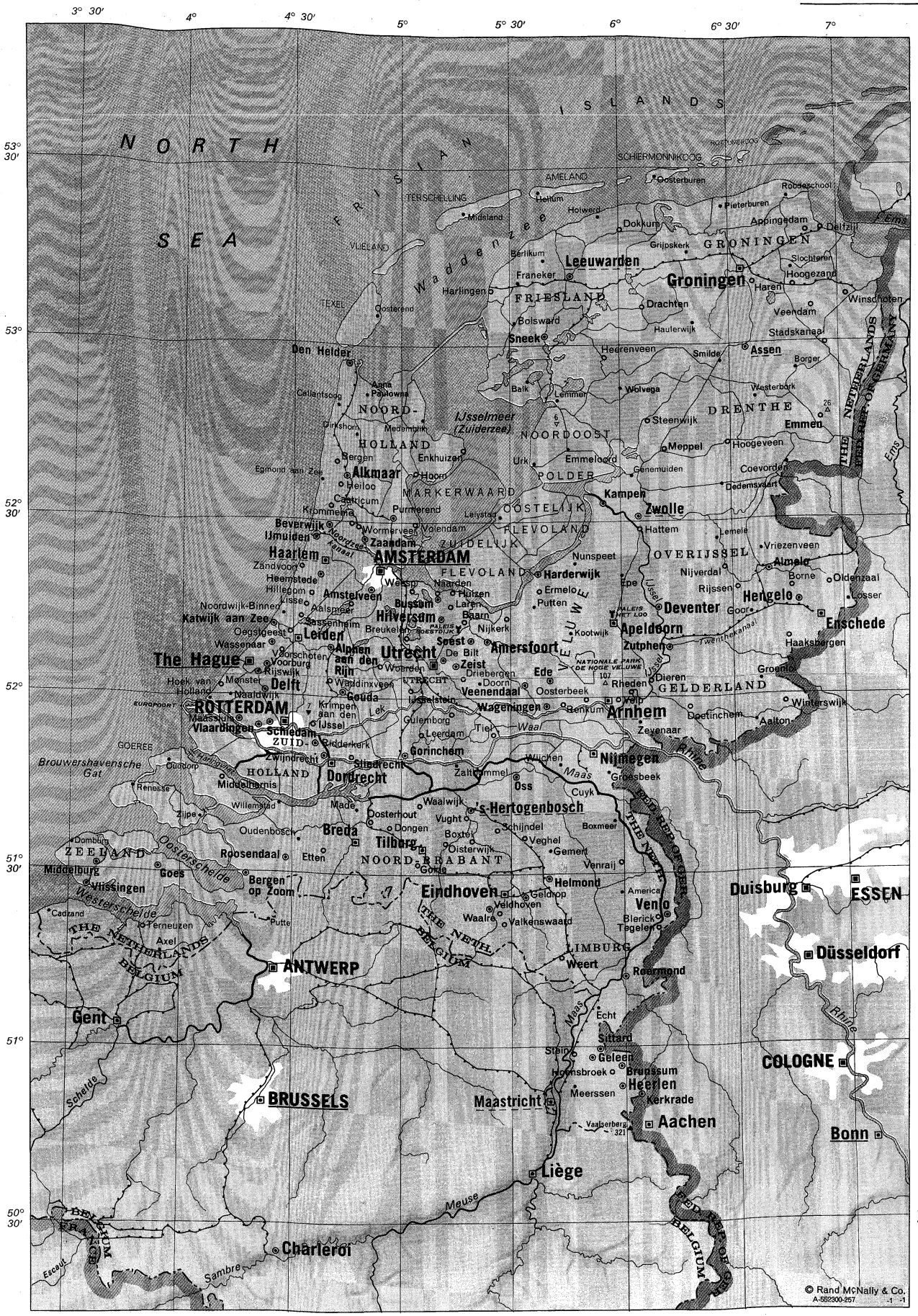
Drenthe 52·50n 6·35e
Friesland 53·10n 5·50e
Gelderland 52·00n 5·50e
Groningen 53·15n 6·45e
Limburg 51·15n 6·00e
Noord-Brabant 51·40n 5·00e
Noord-Holland 52·35n 4·55e
Overijssel 52·15n 6·30e
Utrecht 52·00n 5·15e
Zeeland 51·25n 3·45e
Zuid-Holland 52·00n 4·40e

Cities and towns

Aalsmeer 52·16n 4·45e
Aalten 51·56n 6·35e
Alkmaar 52·37n 4·44e
Almelo 52·21n 6·39e
Alphen aan den Rijn 52·07n 4·40e
America 51·26n 5·59e
Amersfoort 52·09n 5·24e
Amstelveen 52·18n 4·51e
Amsterdam 52·22n 4·54e
Anna Paulowna 52·52n 4·50e
Apeldoorn 52·13n 5·58e
Appingedam 53·18n 6·52e
Arnhem 51·59n 5·55e
Assen 52·59n 6·34e
Axel 51·16n 3·55e
Baarn 52·13n 5·16e
Balk 52·54n 5·34e
Bergen 52·40n 4·41e
Bergen op Zoom 51·30n 4·17e
Berlikum 53·15n 5·39e
Beverwijk 52·28n 4·40e
Blerick 51·23n 6·10e
Bolsward 53·03n 5·31e
Borger 52·55n 6·46e
Borne 52·18n 6·45e
Boxmeer 51·39n 5·57e
Boxtel 51·35n 5·20e
Breda 51·35n 4·46e
Breukelen 52·10n 5·00e
Brunssum 50·56n 5·59e
Bussum 52·16n 5·10e
Cadzand 51·22n 3·25e
Callantsoog 52·49n 4·41e
Castricum 52·33n 4·39e
Coevorden 52·40n 6·45e
Culemborg 51·56n 5·13e
Cuyk 51·44n 5·52e
De Bilt 52·06n 5·10e
Dedemsvaart 52·36n 6·28e
Delft 52·00n 4·21e
Delfzijl 53·19n 6·46e
Den Helder 52·54n 4·45e
Deventer 52·15n 6·10e
Dieren 52·03n 6·06e
Dirkshorn 52·45n 4·45e
Doetinchem 51·58n 6·17e
Dokkum 53·19n 6·00e
Domburg 51·34n 3·30e
Dongen 51·37n 4·56e
Doorn 52·03n 5·21e
Dordrecht 51·49n 4·40e
Drachten 53·06n 6·05e
Driebergen 52·03n 5·16e
Echt 51·06n 5·52e
Ede 52·03n 5·40e
Egmond aan Zee 52·36n 4·37e
Eindhoven 51·26n 5·28e
Emmeloord 52·43n 5·45e
Emmen 52·47n 7·00e
Enkhuizen 52·42n 5·17e
Enschede 52·12n 6·53e
Epe 52·21n 6·00e
Ermelo 52·17n 5·37e
Etten 51·34n 4·38e
Flushing, see Vlissingen
Franeker 53·11n 5·32e
Geldrop 51·25n 5·33e
Geleen 50·58n 5·52e
Gemert 51·34n 5·40e
Genemuiden 52·37n 6·07e
Goes 51·30n 3·54e
Goirle 51·32n 5·04e
Goor 52·14n 6·35e
Gorinchem 51·50n 5·00e
Gouda 52·01n 4·43e
Grijpskerk 53·15n 6·18e
Groenlo 52·03n 6·38e
Groesbeek 51·47n 5·55e
Groningen 53·13n 6·33e
Haaksbergen 52·09n 6·44e
Haarlem 52·23n 4·38e
Hague, The, see The Hague
Harderwijk 52·21n 5·36e
Haren 53·10n 6·35e
Harlingen 53·10n 5·24e
Hattem 52·28n 6·04e
Haulerwijk 53·04n 6·20e
Heemstede 52·21n 4·37e
Heerenveen 52·57n 5·55e
Heerlen 50·54n 5·59e
Heiloo 52·36n 4·43e
Hengelo 52·15n 6·45e
Hillegom 52·18n 4·35e
Hilversum 52·14n 5·10e
Hoek van Holland (Hook of Holland) 51·59n 4·07e
Hoensbroek 50·55n 5·55e
Hollum 53·26n 5·37e
Holwerd 53·22n 5·54e
Hoogeveen 52·43n 6·29e
Hoogezand 53·09n 6·47e
Hook of Holland, see Hoek van Holland
Hoorn 52·38n 5·04e
Huizen 52·17n 5·14e
IJmuiden 52·27n 4·36e
IJsselstein 52·02n 5·03e
Kampen 52·33n 5·54e
Katwijk aan Zee 52·13n 4·24e
Kerkrade 50·52n 6·04e
Kootwijk 52·12n 5·45e
Krimpen aan den IJssel 51·54n 4·35e
Krommenie 52·29n 4·45e
Laren 52·15n 5·14e
Leerdam 51·54n 5·05e
Leeuwarden 53·12n 5·46e
Leiden (Leyden) 52·09n 4·30e
Lelystad 52·31n 5·27e
Lemele 52·27n 6·25e
Lemmer 52·50n 5·42e
Leyden, see Leiden
Lisse 52·15n 4·33e
Losser 52·15n 7·00e
Maassluis 51·55n 4·15e
Maastricht 50·52n 5·43e
Made 51·41n 4·46e
Medemblik 52·46n 5·06e
Meerseen 50·53n 5·45e
Meppel 52·42n 6·11e
Middelburg 51·30n 3·37e
Middelharnis 51·45n 4·11e
Midsland 53·22n 5·16e
Monster 52·02n 4·10e
Naaldwijk 52·00n 4·12e
Naarden 52·17n 5·09e
Nijkerk 52·13n 5·30e
Nijmegen 51·50n 5·50e
Nijverdal 52·22n 6·27e
Noordwijk-Binnen 52·13n 4·27e
Nunspeet 52·23n 5·46e
Oegstgeest 52·10n 4·29e
Oisterwijk 51·35n 5·12e
Oldenzaal 52·19n 6·56e
Oosterbeek 52·00n 5·50e
Oosterburen 53·24n 6·10e
Oosterend 53·05n 4·52e
Oosterhout 51·38n 4·51e
Ouddorp 51·18n 3·56e
Oudenbosch 51·35n 4·31e
Oss 51·46n 5·31e
Pieterburen 53·24n 6·27e
Purmerend 52·31n 4·57e
Putte 51·22n 4·23e
Putten 52·15n 5·36e
Renesse 51·44n 3·46e
Renkum 51·58n 5·45e
Rheden 52·01n 6·02e
Ridderkerk 51·52n 4·36e
Rijssen 52·18n 6·30e
Rijswijk 52·04n 4·20e
Roermond 51·12n 6·00e
Roodeschool 53·25n 6·45e
Roosendaal 51·32n 4·28e
Rotterdam 51·55n 4·28e
Sassenheim 52·13n 4·31e
Schiedam 51·55n 4·24e
Schijndel 51·37n 5·25e
's-Gravenhage, see The Hague
's-Hertogenbosch 51·41n 5·19e
Sittard 51·00n 5·53e
Sliedrecht 51·49n 4·45e
Slochteren 53·12n 6·50e
Smilde 52·56n 5·27e
Sneek 53·02n 5·40e
Soest 52·09n 5·18e
Stadskanaal 53·00n 6·55e
Steenwijk 52·47n 6·08e
Stein 50·57n 5·46e
Tegelen 51·21n 6·09e
Terneuzen 51·20n 3·50e
The Hague ('s-Gravenhage) 52·06n 4·18e
Tiel 51·54n 5·25e
Tilburg 51·34n 5·05e
Urk 52·39n 5·36e
Utrecht 52·05n 5·08e
Valkenswaard 51·21n 5·28e
Veendam 53·06n 6·58e
Veenendaal 52·02n 5·34e
Veghel 51·37n 5·33e
Veldhoven 51·24n 5·24e
Velp 52·00n 5·59e
Venlo 51·24n 6·10e
Venraij 51·32n 5·59e
Vlaardingen 51·54n 4·21e
Vlissingen (Flushing) 51·26n 3·35e
Volendam 52·30n 5·04e
Voorburg 52·05n 4·23e
Voorschoten 52·07n 4·27e
Vriezenveen 52·25n 6·38e
Vught 51·40n 5·17e
Waalre 51·24n 5·26e
Waalwijk 51·42n 5·04e
Waddinxveen 52·03n 4·40e
Wageningen 51·58n 5·40e
Wassenaar 52·07n 4·24e
Weert 51·15n 5·43e
Weesp 52·17n 5·02e
Westerbork 52·51n 6·37e
Wijchen 51·48n 5·43e
Willemstad 51·42n 4·26e
Winschoten 53·08n 7·02e
Winterswijk 51·58n 6·44e
Woerden 52·05n 4·54e
Wolvega 52·53n 6·00e
Wormerveer 52·28n 4·46e
Zaandam 52·26n 4·49e
Zaltbommel 51·48n 5·15e
Zandvoort 52·22n 4·32e
Zeist 52·05n 5·15e
Zevenaar 51·56n 6·05e
Zijpe 51·38n 4·05e
Zutphen 52·08n 6·12e
Zwijndrecht 51·49n 4·39e
Zwolle 52·30n 6·05e

Physical features and points of interest

Ameland, *island* 53·25n 5·45e
Brouwershavensche Gat, *channel* 51·45n 3·50e
De Hoge Veluwe, Nationale Park 52·02n 5·55e
Frisian Islands (Friese Eilanden) 53·25n 5·45e
Goeree, *island* 51·50n 3·55e
Haringvliet, *channel* 51·46n 4·12e
Het Loo, Paleis, *palace* 52·14n 5·56e
IJssel, *river* 52·30n 6·00e
IJsselmeer (Zuiderzee), *sea* 52·45n 5·25e
Lek, *river* 52·00n 6·00e
Maas (Meuse), *river* 51·49n 5·01e
Markerwaard, *physical region* 52·33n 5·15e
Meuse, see Maas
Noordoost Polder, *physical region* 52·42n 5·45e
Europoort, *dock* 51·58n 4·00e
Noordzeekanaal (North Sea Canal) 52·27n 4·33e
North Sea 53·30n 4·00e
North Sea Canal see Noordzeekanaal
Oostelijk Flevoland, *physical region* 52·30n 5·45e
Ooster Schelde, *inlet* 51·30n 4·00e
Rhine, *river* 51·52n 6·02e
Rottumeroog, *island* 53·30n 6·30e
Schiermonnikoog, *island* 53·28n 6·15e
Soestdijk, Paleis, *palace* 52·12n 5·15e
Terschelling, *island* 53·24n 5·20e
Texel, *island* 53·05n 4·45e
Twenthekanal, *canal* 52·13n 6·50e
Vaalserberg, *mountain* 50·46n 6·01e
Veluwe, *physical region* 52·10n 5·55e
Vlieland, *island* 53·15n 5·00e
Waal, *river* 51·55n 4·30e
Waddenzee, *sound* 53·15n 5·15e
Wester Schelde, *estuary* 51·25n 3·45e
Zuidelijk Flevoland (Zuidelijke IJsselmeerpolders), *physical region* 52·22n 5·20e
Zuidelijke IJsselmeerpolders, see Zuidelijk Flevoland
Zuiderzee, see IJsselmeer


NORTH SEA
FRISIAN ISLANDS
Waddenzee
IJsselmeer (Zuiderzee)
NETHERLANDS
BELGIUM
FRANCE
GERMANY
Amsterdam
Rotterdam
The Hague
Utrecht
Eindhoven
Leeuwarden
Groningen
Arnhem
Nijmegen
Zwolle
Enschede
Apeldoorn
Deventer
Hengelo
Haarlem
Delft
Leiden
Dordrecht
Breda
Tilburg
's-Hertogenbosch
Amersfoort
Harderwijk
Alkmaar
Hilversum
Assen
Emmen
Meppel
Hoogeveen
Coevorden
Kampen
Vlissingen
Middelburg
Goes
Bergen op Zoom
Roosendaal
Helmond
Venlo
Roermond
Sittard
Maastricht
Heerlen
Aachen
Weert
Antwerp
Brussels
Ghent
Liège
Charleroi
Cologne
Bonn
Düsseldorf
Duisburg
Essen
Den Helder
Texel
Vlieland
Terschelling
Ameland
Schiermonnikoog
Oosterschelde
Westerschelde
Brouwershavensche Gat
Haringvliet
Volkerak
Maas
Rhine
Scheldt
Meuse
Sambre
Escaut
Ems
Dollard
Twentekanaal
IJssel
Waal
Lek
Gooi
Veluwe
Zuid-Holland
Noord-Holland
Noord-Brabant
Zeeland
Gelderland
Overijssel
Drenthe
Friesland
Groningen
Limburg
Flevoland
Oostelijk Flevoland
Zuidelijk Flevoland
Noordoost Polder
Markerwaard
Nationale Park De Hoge Veluwe
Paleis Soestdijk
Paleis het Loo
Europoort
Goeree
Schiermonnikoog
Rottumerplaat
Rottumeroog
Terschelling
Vlieland
Texel
Ameland
Oosterburen
Holwerd
Harlingen
Den Helder
Oosterend
Midsland
Hollum
Berlikum
Leeuwarden
Franeker
Bolsward
Sneek
Drachten
Dokkum
Grijpskerk
Appingedam
Delfzijl
Roodeschool
Pieterburen
Slochteren
Hoogezand
Haren
Winschoten
Veendam
Stadskanaal
Haulerwijk
Heerenveen
Smilde
Borger
Balk
Lemmer
Wolvega
Steenwijk
Westerbork
Emmeloord
Urk
Genemuiden
Dedemsvaart
Hattem
Lemele
Vriezenveen
Almelo
Borne
Oldenzaal
Nijverdal
Rijssen
Losser
Goor
Haaksbergen
Nunspeet
Ermelo
Putten
Elburg
Epe
Kootwijk
Baarn
Nijkerk
Soest
De Bilt
Zeist
Driebergen
Doorn
Veenendaal
Ede
Oosterbeek
Rheden
Dieren
Velp
Renkum
Wageningen
Groesbeek
Zevenaar
Doetinchem
Aalten
Winterswijk
Gronau
Zutphen
Laren
Huizen
Naarden
Bussum
Weesp
Amstelveen
Aalsmeer
Noordwijk-Binnen
Katwijk aan Zee
Oegstgeest
Wassenaar
Voorschoten
Voorburg
Rijswijk
Monster
Hoek van Holland
Naaldwijk
Wateringen
Pijnacker
Delfgauw
Schipluiden
Maasland
Vlaardingen
Schiedam
Krimpen aan den IJssel
Gouda
Waddinxveen
Alphen aan den Rijn
Woerden
Bodegraven
Reeuwijk
Gouderak
Ridderkerk
Zwijndrecht
Oudedorp
Middelharnis
Renesse
Willemstad
Made
Zuid-Beijerland
Oud-Beijerland
Strijen
Sliedrecht
Gorinchem
Leerdam
Culemborg
Tiel
Zaltbommel
Wijchen
Oss
Cuyk
Grave
Boxmeer
Veghel
Schijndel
Vught
Waalwijk
Oosterhout
Dongen
Oisterwijk
Boxtel
Gemert
Venraij
Helmond
Geldrop
Veldhoven
Waalre
Valkenswaard
Eersel
Oirschot
Best
Son
Nuenen
Gerwen
Nederwetten
Someren
Asten
Deurne
Helmond
Beek en Donk
Laarbeek
Budel
Weert
Nederweert
Roermond
Swalmen
Echt
Susteren
Sittard
Stein
Geleen
Hoensbroek
Brunssum
Heerlen
Kerkrade
Meerssen
Vaalserberg
321
Valkenburg
Simpelveld
Bocholtz
Gulpen
Wittem
Margraten
Eijsden
Vijlen
Epen
Cadzand
Terneuzen
Axel
Hulst
Sluis
Oostburg
Retranchement
Zuidzande
Aardenburg
Putte
Stekene
Sint-Niklaas
Lokeren
Dendermonde
Aalst
Ninove
Geraardsbergen
Oudenaarde
Zottegem
Deinze
Eeklo
Maldegem
Knokke-Heist
Blankenberge
Zeebrugge
Bruges
Ostend
Veurne
Dixmude
Roeselare
Tielt
Menen
Wervik
Comines
Armentières
Lille
Roubaix
Tourcoing
Valenciennes
Maubeuge
Mons
Binche
Soignies
Ath
Tournai
Leuze
Peruwelz
Quiévrain
Hensies
Dour
Frameries
Boussu
Quaregnon
Jemappes
Cuesmes
Hornu
Wasmes
Colfontaine
Pâturages
Quévy
Merbes-le-Château
Thuin
Lobbes
Erquelinnes
Beaumont
Chimay
Couvin
Philippeville
Florennes
Walcourt
Fosses-la-Ville
Sambreville
Auvelais
Jambes
Namur
Dinant
Rochefort
Marche-en-Famenne
Durbuy
Hotton
Barvaux
Comblain-la-Tour
Aywaille
Stavelot
Malmedy
Spa
Verviers
Eupen
Welkenraedt
Herve
Battice
Visé
Dalhem
Fléron
Chaudfontaine
Embourg
Beyne-Heusay
Soumagne
Fléron
Trooz
Pepinster
Theux
Jalhay
Sart
Waimes
Bütgenbach
Elsenborn
Büllingen
Amel
Saint-Vith
Malmedy
Stavelot
Trois-Ponts
Vielsalm
Houffalize
Bastogne
Neufchâteau
Libramont
Bertrix
Paliseul
Florenville
Virton
Arlon
Messancy
Aubange
Athus
Rodange
Differdange
Esch-sur-Alzette
Dudelange
Luxembourg
Mersch
Ettelbruck
Diekirch
Wiltz
Clervaux
Vianden
Echternach
Grevenmacher
Remich
Bitburg
Prüm
Daun
Gerolstein
Hillesheim
Mayen
Andernach
Neuwied
Koblenz
Bad Ems
Limburg
Wetzlar
Siegen
Olpe
Meschede
Arnsberg
Soest
Lippstadt
Paderborn
Detmold
Bielefeld
Gütersloh
Rheda-Wiedenbrück
Warendorf
Ahlen
Beckum
Hamm
Unna
Kamen
Lünen
Dortmund
Bochum
Herne
Gelsenkirchen
Recklinghausen
Castrop-Rauxel
Herten
Marl
Dorsten
Wesel
Xanten
Kleve
Goch
Emmerich
Rees
Bocholt
Borken
Ahaus
Gronau
Rheine
Lingen
Meppen
Papenburg
Leer
Emden
Norden
Aurich
Wittmund
Jever
Wilhelmshaven
Varel
Oldenburg
Delmenhorst
Bremen
Verden
Rotenburg
Soltau
Celle
Uelzen
Lüneburg
Winsen
Buxtehude
Stade
Cuxhaven
Otterndorf
Bremerhaven
Nordenham
Brake
Elsfleth
Westerstede
Leer
Papenburg
Meppen
Lingen
Rheine
Steinfurt
Ibbenbüren
Osnabrück
Melle
Bad Oeynhausen
Minden
Nienburg
Stolzenau
Hoya
Syke
Diepholz
Sulingen
Walsrode
Fallingbostel
Munster
Soltau
Buchholz
Harburg
Hamburg
Geesthacht
Ratzeburg
Mölln
Lauenburg
Schwarzenbek
Bergedorf
Reinbek
Ahrensburg
Norderstedt
Pinneberg
Elmshorn
Itzehoe
Wilster
Glückstadt
Brunsbüttel
Heide
Meldorf
Büsum
Husum
Niebüll
Tondern
Flensburg
Apenrade
Hadersleben
Sonderburg
Nordborg
Aabenraa
Rendsburg
Eckernförde
Kiel
Neumünster
Bad Segeberg
Lübeck
Travemünde
Timmendorfer Strand
Neustadt in Holstein
Oldenburg in Holstein
Eutin
Plön
Preetz
Bad Schwartau
Grömitz
Heiligenhafen
Fehmarn
Puttgarden
Burg
Marne
Meldorf
Tönning
St. Peter-Ording
Friedrichstadt
Schleswig
Flensburg
Kappeln
Schleimünde
Eckernförde
Kiel
Laboe
Plön
Eutin
Neustadt in Holstein
Timmendorfer Strand
Travemünde
Lübeck

The Netherlands, Area and Population

	area*		population†‡	
	sq mi	sq km	1960 census	1972 estimate
Provinces (*provincies*)				
Drenthe	1,023 1,037	2,649 2,685	312,000	380,000
Friesland	1,279 1,468	3,312 3,803	479,000	533,000
Gelderland	1,936 1,981	5,013 5,131	1,274,000	1,558,000
Groningen	883 939	2,287 2,432	475,000	527,000
Limburg	837 853	2,169 2,209	886,000	1,022,000
Noord-Brabant	1,898 1,971	4,915 5,105	1,496,000	1,850,000
Noord-Holland	1,025 1,124	2,656 2,912	2,057,000	2,274,000
Overijssel	1,470 1,516	3,806 3,927	804,000	946,000
Utrecht	513 539	1,328 1,396	681,000	827,000
Zeeland	684 1,060	1,772 2,746	283,000	316,000
Zuidelijke IJsselmeerpolders§‖	370¶ 417¶	959¶ 1,079¶	1,000	20,000
Zuid-Holland	1,101 1,287	2,851 3,333	2,707,000	3,013,000
Total The Netherlands	14,718♀♂ 15,892♀♂	38,119♀♂ 41,160♀♂	11,462,000♀□	13,270,000♀◇

*Where two figures are given, the first is the land area, the second the total area. †De jure. ‡According to country and province boundaries at the time of the census. §The Zuidelijke IJsselmeerpolders are not a province or first-order political subdivision. They are land that has been reclaimed from the IJsselmeer (Zuiderzee) and has not yet been incorporated into any of the provinces. ‖Since January 1, 1972, Dronten became a municipality and is situated in Zuidelijke IJsselmeerpolders. ¶Provisional. ♀Figures do not add to total given because of rounding. ♂Area includes 1,700 sq mi (4,402 sq km) of land not municipally classified. □Including nearly 6,000 persons not enumerated in any province but registered with the Central Register of Population. ◇Including nearly 3,000 persons not enumerated in any province but registered with the Central Register of Population.
Source: Official government figures.

Under the Romans, a German infiltration started that continued for centuries. The result was a predominantly Anglo-Saxon Frisia in the northeast, while the original Holland, further south, developed under Frankish influence.

Also of significance has been the attraction of countless refugees to the tolerant Dutch cities. Over the centuries, these have included southern Lowlanders, French Huguenots, Salzburgers, Swiss, and Portuguese Jews who left their respective homelands as a result of religious or political pressures. In addition, many Germans and non-Iberian Jews have also tended to immigrate to The Netherlands for more purely economic reasons. In the 20th century, Dutch-Indonesians and non-white peoples from Ambon, in the East Indies, and Surinam, on the northeast coast of South America, have added to the influx.

The heritage of Dutch Humanism

The Netherlands, a small state among big ones, has consistently preached freedom of thought. Dutch Humanism has always been a judicious mixture of piety and knowledge, epitomized in the development during the 14th and 15th centuries of a bourgeois civilization with a military and commercial spirit. These traditions, together with rigid Calvinism and a more tolerant Protestantism and Roman Catholicism, continue to characterize much of Dutch life. Even today, political, educational, and territorial divisions are along religious lines. Not even World War II and the unity forged under German occupation were able to break up the pattern, and only now are signs of change appearing.

Finally, even though in such a small country, individual regions have acquired their own distinctive personalities. The Zeelanders of the north are often thought of as possessing a certain asceticism in their life-style developed through centuries of struggle against the sea—in contrast to the more easygoing, almost Burgundian French, way of life in the south.

Demographic trends. In 1970 the population of The Netherlands passed the 13,000,000 mark, an increase of nearly 4,000,000 since World War II.

Live births, although increasing in absolute terms, have been declining relatively from 24 per 1,000 persons during the period 1945–54 to about 18 per 1,000 more recently. The mortality rate, moreover, down to 7.7 per 1,000 during the period 1955–64, had risen slightly by the early 1970s, when the life expectancy at birth was just under 71 years for males and over 76 years for females.

The average age of the Dutch is now almost 33, reflecting an increase of elderly people: those over 65 number about 10 percent of the population. The average family is shrinking in size, while the numbers both of households and of single persons multiply. Although the state of the national economy, which generally flourished in the 1950s and 1960s, has made an emigration policy unnecessary, a department of emigration is still maintained in the Ministry of Social Affairs and Public Health, and certain subsidies are still granted. Emigration to Canada, the United States, and Australia, which amounted to more than 30,000 persons per year in 1956 and 1957, has fallen to about 9,000 a year. South Africa still attracts between 1,700 and 1,800 persons yearly, and New Zealand attracts between 500 and 600.

As a result of the postwar government policy mentioned earlier, emigrants exceeded immigrants by about 10,000 each year from 1947 to 1954. During the years from 1955 to 1964, however, the economy and labour potential of western Europe attracted an increasing number of Spaniards, Yugoslavs, Turks, and Moroccans, so that the balance remained static. The latest figures indicate an immigrant surplus, a situation that reflects, in part, the growing number of immigrants from the former Caribbean colonies of Surinam and the Netherlands Antilles, who, as citizens of the kingdom, have free access to The Netherlands.

Concentration of population in the two western provinces

Nearly half the population lives in the two western provinces of Zuid-Holland and Noord-Holland and in the central province of Utrecht, in which the concentration is as high as 2,300 persons per square mile (895 per square kilometre). The four largest cities—Amsterdam, Rotterdam, The Hague ('s-Gravenhage), and Utrecht—are located in this area. In contrast, the three northern provinces—Groningen, Friesland, and Drenthe—are sparsely settled, with an average density of over 450 persons per square mile.

Figures on internal migrations suggest continuous emigration from Groningen and Friesland, in which industrialization has been relatively unsuccessful, accompanied by movement from Zuid-Holland and Noord-Holland toward Utrecht and the less densely populated provinces of Drenthe, Overijssel, Gelderland, and Noord-Brabant. The southernmost province, Limburg, suffers from a decaying coal-mining industry, and its inhabitants are increasingly forced to seek their living elsewhere. Although the current estimate of a total 17,100,000 inhabitants by the year 2000 is lower by 800,000 people than the 1967 estimate, The Netherlands will still have to absorb a growth of nearly one-third in a single generation. Attention will therefore have to be concentrated on regional planning and economic development, and both will demand a high rate of private and public investment. This burden, along with the support of those above 65, will fall on the working population. On the other hand, there will be relatively fewer schoolchildren. Some people have advocated an official birth-control policy, but traditional religious attitudes may prevent its fulfillment.

THE NATIONAL ECONOMY

Position in the world. Since World War II, The Netherlands has become a highly industrialized nation occupying a central position in the economic life of western Europe. Although agriculture now accounts for less than 7 percent of the national income, it remains a highly specialized contributor to Dutch exports. Because of the scarcity of mineral resources—with the important exception of natural gas—the country is dependent on large imports of basic materials. Given this situation, a continued trade deficit of something over 10 percent has to be offset by a favourable balance in transport and transit; in these areas the nation's fleet has become a major international carrier.

Dutch trading partners

Trade relations are mainly with Europe and North America, which supplied, in the early 1970s, over 80 percent of all imports, while receiving nearly 90 percent of all exports. The European Economic Community is, of course, the dominant factor in foreign trade, and, whereas in 1958 (just before the Common Market was set up) only 41.5 percent of Dutch exports went to West Germany, Belgium and Luxembourg, France, and Italy, in 1970 these nations accounted for over 60 percent.

Natural power sources. With increasing use of oil and natural gas, coal, which is cheaper when imported from abroad, is now little mined. Between 1958 and 1970, the number of coal miners fell from 48,000 to about 12,000—though rising productivity kept coal production about 40 percent of the 1958 amount. Coal mining is completely concentrated in southern Limburg.

The production of crude oil (a little over 1,600,000 tons a year) covers only a small part of Dutch requirements. The wells are located in southeastern Drenthe, near Schoonebeek, and in Zuid-Holland.

The discovery of natural gas during the 1950s has had a tremendous influence on the development of the Dutch economy. The gas field near Slochteren, in Groningen, holds proven reserves of more than 5,830,000,000,000,000 cubic feet (165,000,000,000,000 cubic metres); by 1971, annual production there and at some lesser fields had risen to some 88,000,000,000,000 cubic feet (2,000,000,000,000 cubic metres). About one-third of this production was exported to Germany, Belgium, and France, helping to improve the balance of payments in the area in which The Netherlands has usually had its largest deficit. An aluminum factory using natural gas as fuel was set up at Delfzijl, in Groningen. It became the core of an important industrialization project in a traditionally agricultural area.

The use of coal for producing coke is no longer an economical proposition in The Netherlands, and coke-oven gas has been superseded by natural gas as fuel for the chemical industry of south Limburg. Domestic natural-gas consumption for cooking and heating is also continuously rising.

Intense development of dairying and horticulture

Agriculture. Of the country's 9,000,000 acres (3,600,000 hectares), 5,300,000 are under cultivation: 3,300,000 acres as grassland, 1,700,000 acres for arable farming, and 300,000 acres for horticulture, especially tulips. Dutch cattle farming is highly developed: the milk yield per acre of grassland and the yield per cow are among the highest in the world. About 40 percent of the total milk production is exported after being processed into such dairy products as butter, cheese, and condensed milk. Large-scale crop cultivation is hardly possible in The Netherlands, and most fodder as well as cereals for human consumption are imported from such countries as Canada and the United States.

Horticulture carried on under glass is of special importance. The export of hothouse tomatoes, cucumbers, and lettuce increased sixfold from the 1950s to the late 1960s, and The Netherlands now contains almost half of the total European horticultural area under glass.

During the period from 1950 to 1966, the country's agricultural labour force decreased by half, and the number of farms dropped from 210,000 to about 145,000; in horticulture the number of holdings remained constant. During the same period, the total output of Dutch agriculture doubled—a clear indication of a rise in productivity. Fishing now accounts for only about 1 percent of the national income.

Industry. Dutch industrial development started relatively late, in 1870, and production rose even during the depression of the 1930s. Further development became a priority after World War II, when ascending population figures and growing farm-labour surpluses necessitated tens of thousands of additional jobs each year. A spectacular contribution to the solution of the employment problem was made by the chemical industry, whose annual output increased 10-fold between the late 1940s and the early 1970s—the greatest growth rate of any sector of Dutch industry.

Metal manufacturing, however, remains the most important industry. The biggest plant is the Royal Netherlands Blast Furnaces and Steel Works, at IJmuiden, with a capacity of 4,000,000 tons of crude steel per year. The electronics industry is also important: one large corporation alone accounts for one-tenth of all Dutch exports. The food, drink, and tobacco industries are a strong second, based as they are on the highly productive agriculture sector. Dutch industry has also been able to redevelop commercial aircraft since World War II, and automobile manufacturing has been expanded. In contrast, textile employment decreased over the 1960s, and several companies have been forced to close down entirely.

Banking

Finance. Commercial banking in The Netherlands is in the hands of a few large concerns. The Algemene Bank Nederland and the Amsterdam-Rotterdam Bank handle 60 percent of Dutch financial business between them, while the Coöperatieve Centrale Raiffeissen-Bank en Aangesloten Banken and the Coöperatieve Centrale Boerenleenbank specialize in agricultural credit. The more than 200 independent savings banks are centred on the state-owned Post Office Savings Bank. Since 1948 the central bank, Nederlandsche Bank, has also been state owned. Its function, defined by the 1948 Banking Act and the 1956 Credit Control Act, is to supervise the entire system. An independent agency, it has authority to strengthen traditional monetary controls.

The Dutch economy nevertheless remains basically in private hands. Postwar industrial development has been consciously stimulated by government policy, and state subsidies have been granted to attract industry toward the relatively underdeveloped northern provinces, as well as Limburg and certain other pockets of economic stagnation. Despite these subsidies, the western provinces remain the centre of new industry, especially in the services sector.

State ownership, employer organizations, and the role of labour. The state completely owns such corporations as Netherlands Railways and the Dutch State Mines and is a large-scale participant in others: the Royal Steel Works, the Royal Dutch Airlines (KLM), De Nationale Investeringsbank (Herstelbank), and the Gas Union, which distributes natural gas. The state thus has a direct, continuous influence on the management of some sectors of the economy, and especially on investment policy. Dutch employers are organized in three separate but closely cooperating organizations, one Catholic, one Protestant, and one nondenominational. Labour has a similar tripartite organization, comprising Catholic, Protestant, and Socialist groups, with some other, independent fringe organizations. Although the state has wide legal powers for intervening in wage disputes, government policy has not been able to put an effective check on inflationary rises in wages and prices. Consequent increases in the cost of living have been on the order of 7 percent annually from the late 1960s onward. A strong anti-inflationary policy was announced by the government formed after the elections of April 1971.

TRANSPORTATION

Transportation is of special importance in The Netherlands, since the country functions as a gateway for goods traffic between western Europe and the rest of the world. Goods flow through Dutch harbours, continuing their passage by river vessel, train, truck, pipeline, and—a recent development—by modern container units.

Shipping and air traffic. The maritime traffic of The Netherlands is about equal to the land traffic, and indeed the whole southern part of the North Sea may be likened to an immense traffic square, fed as it is by the Thames, Rhine, Maas, and Scheldt, with links into the hinterland of the Continent that make it the greatest commercial artery of the world after those of North America. Rotterdam has the country's best equipped modern harbour. Europoort, the region between Rotterdam and the sea, can easily be reached by the biggest oceangoing ships; it serves as an approach to Rotterdam harbour, itself handling more tonnage than any other harbour in the world. Rijnmond (the mouth of the Rhine) and IJmond (the

mouth of the IJ) are also equipped to receive big ships. In oil processing, too, Rotterdam is the world's biggest centre, having expanded from an annual capacity of 1,200,000 tons in 1930 to 137,200,000 in 1970. A trench 1,500 feet wide and 75 feet deep extends eight miles into the North Sea to facilitate the entry of supertankers with 200,000-ton capacities. The overall goods traffic in the sea harbours of Rotterdam and Amsterdam in fact increased sixfold over the 1951–70 period. The number of ships of the Dutch merchant fleet, which is among the 12 largest fleets in the world, decreased by 32 percent from 1960 to 1971, and the total tonnage decreased by 12 percent. The number of rivercrafts is probably unsurpassed by any other country.

Royal Dutch Airlines (KLM) initiated a scheduled service between Amsterdam and London in 1920 and now has a worldwide network. One subsidiary operates an airline in the West Indies, while another runs a modest domestic network between Amsterdam and the outlying provinces. Schiphol, Amsterdam's airport, and Zestienhoven, the airport at Rotterdam, both are of international importance. The Dutch government conducts a school for civil aviation and pilot training.

Inland transportation. As might be expected in so thickly settled and small a country, the road network is comprehensive. There are more than 57,000 miles (92,000 kilometres) of paved roads, half of which have separate lanes and two-level crossings. A rapid rate of expansion was planned for the 1970s and beyond, especially to handle the increasing number of private cars, already one to every six inhabitants.

Despite the competition offered by railroads, motor freight, containers, and pipelines, it appears that the important waterways will never be surpassed for traffic in freight. For this reason, much attention and financial support are devoted to the waterways in order to meet the growing international demand for efficient goods transport.

The continued importance of waterways

The Netherlands has about 3,500 miles of inland waterways linked with Belgian, French, and German systems. Besides such natural waterways as the rivers Maas, Rhine, Waal, and Lek, many artificial waterways—the Juliana Kanaal, the Amsterdam–Rijn Kanaal, and the Maas–Waal Kanaal—connect the major ports on the coast with the hinterland. The state-owned railroads are modest but well developed and in total length are about average for Europe. The combined effect of inland navigation and freight traffic on roads has prevented the expansion of goods traffic by rail. More than half of the rail network is electrified; elsewhere, steam has been replaced by diesel.

ADMINISTRATION AND SOCIAL CONDITIONS

The Netherlands and the former colonies of Surinam and the Netherlands Antilles form the Kingdom of The Netherlands, a constitutional monarchy—though Surinam and the Antilles are, in practice, autonomous and have their own governmental system. The monarchy is hereditary in both the male and female line, though the latter succeeds only in default of male heirs. By the early 1970s, the monarch had in fact been a queen since 1890.

Constitutional framework. The constitution, which dates from 1814, maintains that the head of state is inviolable, but it also embodies the concept of ministerial responsibility; it further provides that no government can remain in power against the will of parliament. The States-General (Staten-Generaal), as parliament is officially known, consists of two chambers: a directly elected Second Chamber of 150 members and a First Chamber, whose 75 members are elected by the members of the councils of the 11 provinces. Both chambers share legislative power with the government (officially known as the crown, defined as the head of state acting in conjunction with the ministers) and control government policy, but, whereas the First Chamber has only the power of veto in the legislative process, the directly elected Second Chamber, controlling the government, plays a dominant role.

Every four years, after elections for the Second Chamber have been held, the government resigns and a process of bargaining starts between various parliamentary leaders who hope to form a government that will control a majority in parliament. It usually takes a few months before a *formateur,* as the architect of such a coalition is known, can accept a royal invitation to form a government; the head of state then formally appoints the ministers. In the event of political crises resulting in the fall of the government before the end of a four-year period, the same process of bargaining takes place. The monarch, however, has the right to dissolve one or both chambers, at which time new elections are held.

Formation of the government

Local government. In local government, the most important institutions are the municipalities, of which there are 865. Their jurisdiction extends over the entire country (except for some areas of recently reclaimed land). Each municipality is run by a directly elected council of between seven and 45 members, depending on the size of the population. This council is presided over by a *burgemeester* (mayor), who is appointed by the government and serves as chairman of the executive, members of which are elected by and from the council. Where the councils' own ordinances are applicable, the municipalities are autonomous, and in many instances national legislation or provincial ordinances provide for the cooperation of municipal authorities.

The country is also divided into 11 provinces for purposes of local government—Noord-Holland, Zuid-Holland, Zeeland, Utrecht, Noord-Brabant, Limburg, Gelderland, Overijssel, Drenthe, Groningen, and Friesland. Their administrative system has the same structure as the municipal one: directly elected councils (*staten*) who elect the members of the executive, except for the chairman, who is appointed by the government. The main function of the provinces consists of controlling the municipalities within their borders, as well as controlling the district water-control boards (*waterschappen*), the public authorities that are responsible for water control within certain definite areas.

Political life. The Second Chamber, the provincial councils, and the municipal councils are elected according to a system of proportional representation. In general elections for the Second Chamber, it can take as little as 0.66 percent of the overall vote to get one of the 150 seats in the chamber. As a result, a large number of parties and political movements are represented in parliament. The three Christian centre parties (one Catholic and two Protestant) have slowly evolved a form of cooperation that included a common platform for the 1971 parliamentary elections. The Labour Party, the Democrats 1966 (D66—a party of constitutional reform and progressive, non-ideological politics), and the Radical Political Party (a progressive splinter, mainly from the Catholic People's Party) have also coalesced for electoral purposes, presenting a common program and a list of potential ministers as well. These two blocs have recently occupied, between them, more than two-thirds the total number of seats, thereby creating a certain measure of order in Dutch party politics.

Objections to the political system

Because of the complicated party system and the time-consuming and complicated process by which a government comes into being, constitutional and party reforms are advocated by many people. Dissatisfaction with "the system" expresses itself by widespread utterances of frustration about official politics, often embodied in extra-parliamentary movements.

Justice and the armed forces. In The Netherlands, the ordinary administration of justice is entrusted exclusively to appointed judges; trial by jury is unknown. There are 62 cantonal courts (*kantongerechten*), which exercise jurisdiction in a whole range of minor civil and criminal cases. More important cases are handled by one of the 19 district courts (*rechtbanken*), which also can hear appeals over cantonal court decisions. Appeals against decisions from the district courts are heard by one of five courts of appeal (*gerechtshoven*).

The Supreme Court (Hoge Raad) ensures a uniform application of the law, but it cannot suspend a law as being contrary to the constitution. In the legislative pro-

cess itself, the government and the parliament together pass judgment on the constitutionality of a bill under consideration. Laws that are contrary to an international agreement to which The Netherlands is a party can be abrogated by the courts.

The Dutch armed forces consist of an army, a navy, and an air force. All male citizens become liable to military service at the age of 18. Since 1967, the normal period of obligatory service has been fixed at 16 months, with an additional two months for officers and noncommissioned officers. The Netherlands is a member of NATO. The Dutch armed forces also are responsible for the defense of Surinam and the Netherlands Antilles.

The social sector. *Education.* The Dutch educational structure consists of a public and a private system. Private schools, as long as they conform to legally determined standards of such things as curriculum, examinations, and qualifications of personnel, are financially subsidized up to the highest level, according to the same rules in force for the public system. With a few exceptions, private schools are run by religious organizations, either Protestant or Catholic. Higher education is provided by three state universities: Leiden (established in 1575), Groningen (1614), and Utrecht (1636); other universities include that of the city of Amsterdam, which is the largest (1632), and two private universities, the Protestant Free University (Vrije Universiteit) in Amsterdam (1880), and the Catholic University of Nijmegen (1928).

In addition, higher technical education at the university level takes place in special institutions in Delft, Eindhoven, and Enschede. Private institutions for the teaching of economics at university level are established in Rotterdam (nondenominational) and Tilburg (Catholic). All these are financed by the state, and, although different rules have been considered, the right of entrance is determined by final examinations in secondary schools.

Housing. Since World War II, housing policy has been one of the most controversial topics of political and social life. Because of war damage and a stagnant building trade, a considerable housing shortage had developed by the early 1950s. The rate of population growth, moreover, had accelerated, and rising living standards meant that people expected more space per person and generally higher quality in housing facilities. Estimates of the existing housing shortage had to be revised upward continuously, but, with a building program of about 130,000 dwellings per year, an end to the shortage was in sight in the early 1970s.

The government's role in housing

Construction is heavily subsidized in The Netherlands. Since 1945, more than four out of every five new dwellings have been built with some form of public support. Subsidies are paid to municipalities and building associations, which are mainly concerned with the housing needs of lower income groups, or to private persons who conform to certain building standards. In urbanized areas, rent control is still maintained. Local authorities have power to set up a system of housing allocation, a power particularly exercised in metropolitan areas.

Social welfare. The Netherlands has an extensive system of social insurance. For instance, the state has direct supervision of old-age, widows', and orphans' pensions, to which all citizens are entitled. The enforcement of the Health Insurance Act, the Working Incapacity Insurance Act, and the Unemployed Benefit Act is entrusted to special associations set up by employers' and employees' organizations. The last two acts cover all employees, whereas participation under the Health Insurance Act is compulsory only for employees below a certain wage level, determined yearly according to the general level of wages and prices. Employees and self-employed persons also receive children's allowances, though rules and amounts differ for each group.

Since 1968, a special act has covered certain specialized or prolonged medical expenses for all citizens. About 70 percent of the population is covered by the Health Insurance Act, and most other persons are more or less covered by some form of private health insurance. A guaranteed minimum income is also maintained.

CULTURAL LIFE AND INSTITUTIONS

The cultural climate in The Netherlands is varied and lively, though the official life of the arts is mainly an upper class affair. There are nevertheless some signs of malaise, not only in concert life but also in the theatre and the fine arts. The interest of the public is declining, and the awarding of rather extensive subsidies by the administration has caused unrest among the public as well as among the artists themselves, many of whom are looking for new means of support. There is also a growth from a kind of artistic underground in which the young have begun to create new possibilities—with an international and anti-establishment flavour. These "subcultures" may not be recognized as part of the official arts, yet they are a significant development, often characterized by the fashions followed by the young.

The fine arts. There are 32 state-supported museums in the nation, the most famous of which are the Rijksmuseum at Amsterdam, the Mauritshuis at The Hague, and the Stedelijk Museum, also in Amsterdam; the first is noted for its magnificent collection of works by the great 17th-century Dutch master Rembrandt van Rijn, and the last is known for its collection of paintings by the late-19th-century painter Vincent van Gogh. Modern museums, with their progressive policies, attract large sections of the public. There are nearly 400 public collections of art and culture.

Connected with various cultural institutions are the Preservation Service for Monuments, which restores valuable houses and buildings and protects archaeological treasures; the State Office for the Documentation of the History of Art; and the Central Laboratory for Art and Science.

The history of painting could scarcely be written without including, besides those of Rembrandt and van Gogh, the names of the great painters of Holland, such as Jan van Eyck, the founder of the Flemish school in the 15th century; the allegorical Hieronymus Bosch in the 16th; portraitists such as Frans Hals and landscapists such as Jan Vermeer, Aelbert Cuyp, and Salomon van Ruysdael in the 17th; the 17th-century still-life artists such as Willem Heda and Willem Kalff; and the geometrically inclined Piet Mondrian in the 20th. Interest in painting in The Netherlands is still widespread, and such artists as Kees van Dongen, Willem de Kooning, and Karel Appel are internationally famous, as much in demand among collectors of contemporary works as classical masters are by the great museums of the world. Sculpture, on the other hand, has always been overshadowed by the more prominent pictorial developments. Thanks to a complicated system of public assistance and scholarships, young artists can indulge in their experiments with the certainty of a minimum income, while the government and municipalities regularly buy their works to adorn various administrative centres.

The problems of Dutch theatre and literature

Dutch theatre and literature have always been handicapped by the very tiny proportion of the human race that speaks Dutch: some 17,000,000 in Europe (among them the Flemings of Belgium); some 6,000,000 Afrikaans speakers in South Africa; and over 500,000 Dutch citizens abroad. This limits the possibility of Dutch writing making a worldwide impact; except for the translation of some Dutch novels and poems, Dutch literature, though very much alive, has received little of the attention accorded the universal visual vocabulary of Dutch painting.

Theatre companies are all private foundations, though the state compensates 40 percent, and the municipalities 60 percent, of any loss incurred by theatres. Players' salaries are legally determined, and performances of Dutch plays are encouraged, though the administration is not concerned with artistic qualities. Television has had some adverse effect on theatre life in The Netherlands.

Young Dutch film makers, less handicapped than their writing colleagues, have chalked up many successes abroad, usually with documentaries about Dutch painters. There is also a Netherlands Film Academy, which is building up a reputation far outdistancing the very limited production of feature films.

Musical life, on the other hand, is intense. Although private orchestras in The Netherlands are supported by the government, the latter makes no attempt to interfere in the operation of these organizations or the choice of repertory. The Amsterdam Concertgebouw orchestra is world-famous, and Dutch composers are encouraged. Other organizations, such as the three Dutch ballet companies, several choirs, and music schools, are also subsidized.

The press and broadcasting. The first European newspapers were printed in Amsterdam in 1620, and the Dutch press has a reputation for high-quality reporting; yet, despite ample advertising, it is plagued by financial depression. This situation stems from rising costs of newsprint and printing and the competition of radio and television in both news reporting and advertising. By 1971, half of the Dutch newspapers were operating at a deficit. Closedowns and mergers of old and respected newspapers (such as *Algemeen Handelsblad,* founded in 1828, and *Nieuwe Rotterdamse Courant,* founded in 1844) seem hardly able to check the crisis. During the previous 10 years, the number of newspaper publishers had sunk from 51 to 31. In the latter 1950s, there were 162 newspapers, with a daily circulation of 2,700,000; in 1971 there were fewer than 100 newspapers, though daily circulations had risen to about 4,000,000.

The Dutch broadcasting system is more complicated and may be considered unique, but above all it is democratic. Seven organizations, all under private initiative and with ideological aims, account for about 80 percent of radio broadcasting and 60 percent of television. These Catholic, Protestant, Christian, Socialist, and neutral stations all operate within the framework of the Netherlands Broadcasting Foundation (NOS), which is responsible for the actual transmission and for the joint programs filling the rest of the time. Church societies, political parties, and small representative groups are granted some reserved broadcasting time.

The government itself has no influence on the programs, and even permission to advertise is granted to a separate foundation, with a limited allocation of time. All broadcasting is financed by a licensing fee and by the yield from television and radio advertisements. These sums are paid into the Ministry of Cultural Affairs, Recreation and Social Welfare, which in turn pays the cost of the transmitters and allots the broadcasting systems a fixed amount per hour for radio and television programs.

THE OUTLOOK

The Netherlands is undergoing a period of intensive change. Although a growing population necessitates a policy of continuous economic growth, the combined demands on available space made by industry, traffic, housing, and recreation confront the authorities with complex decisions that can only be made within a framework of regional planning. At the same time, such decisions will have to take into account the country's evolution from a hierarchical democracy, built around religiously based reverence for authority, to a participatory democracy. The party system and the constitution also will undoubtedly have to be adapted to this change in social organization.

These problems can only be solved within a European context, and Dutch foreign policy, which has been based on the necessity for European integration since the end of World War II, will doubtless continue in that direction. Supranational policies are increasingly dictated in such fields as industrial development, the regulation of traffic, prevention of pollution, and protection of the natural environment. Actually, The Netherlands, densely populated and already highly developed economically, can hope to become something of a pioneer in solving the problems that confront industrialized Western civilization in the last quarter of the 20th century.

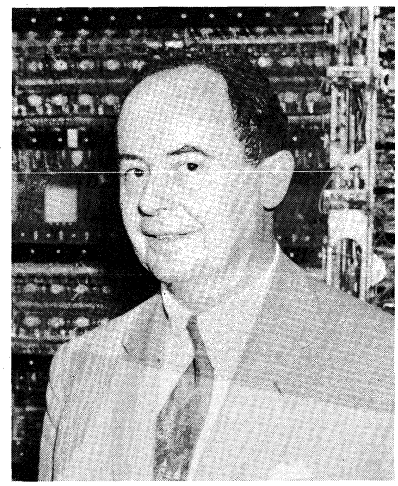
BIBLIOGRAPHY. The most extensive source of information on The Netherlands is *The Kingdom of The Netherlands* (1971), an official publication. The *Digests of the Kingdom of The Netherlands* are separate publications on the economy, the system of social security, and the constitutional organization in The Netherlands; the Government Information Service updates these publications annually. A complete survey of geographic conditions is published by the CARTOGRAPHIC INSTITUTE (BOOTSMA) under the title *Complete Geography of The Netherlands* (1971). Specialized works on drainage and water control are JOHAN VAN VEEN, *Dredge, Drain, Reclaim: The Art of a Nation,* 5th ed. (1962); see also two publications of the GOVERNMENT INFORMATION SERVICE: *The Zuyder Zee Works* (1967) and *The Delta Project* (1967). On demographic problems only works in Dutch are available; PIETER VAN DAALEN, *Dringen in Nederland* (1965); and J.P. VAN ROOIJEN, *Het Bevolkingsvraagstuk* (1961), more or less cover the field. HILBRAND BOSCHMA, *Fauna von Nederland* (1927); and H. HEUKELS, *Schoolflora voor Nederland* (1886, reprinted 1934), are good, if dated, works on Dutch flora. For a description of economic life the following works should be consulted: *The Textile Industry in Europe* (OECD, 1962), FOLKERT DE ROOS, *De Algemene Banken in Nederland* (1968); C.P.A. GEPPAART, *Onevenwichtigheden in de belastingwetgeving* (1968); *Politique des Transports aux Pays-Bas* (Éditions Techniques et Économiques, 1967). Dutch law and politics may be studied from the before-mentioned *Digests* and also from ERNST VAN RAALTE, *Het Nederlandse Parlement* (1958; Eng. trans., *The Parliament of the Kingdom of The Netherlands,* 1959). R.L. MORLAND, *Local Government in The Netherlands* (1958), can be a guide in that particular field. JOHN P. WINDMULLER, *Labor Relations in The Netherlands* (1969); and *Manpower and Social Policy in The Netherlands* (OECD, 1967), are the best introductions to Dutch labour politics; educational developments are well covered by *Educational Policy and Planning in The Netherlands* (OECD, 1967). The office for the documentation of art history publishes a *Bibliography of Dutch Art* (The Hague). *Broadcasting in The Netherlands* (NOS, 1971), explains the Dutch system of radio and television broadcasting.

(J.M.Lu./J.P.A.G.)

Neumann, John von

A mathematician, John von Neumann made important contributions to the sciences of quantum physics, logic, and meteorology that had practical applications in national defense and in the development of high-speed computing machines. His mathematical theory of games of strategy had a significant impact on economics.

Alan W. Richards



Von Neumann.

Von Neumann was born December 3, 1903, in Budapest, Hungary. Many legends about him, from childhood on, tell of his phenomenal speed in absorbing ideas and solving problems and of his equally phenomenal memory that retained everything he once thought through. He studied chemistry at the University of Berlin, and, at Technische Hochschule in Zürich, received the diploma in chemical engineering in 1926. The same year he received the Ph.D. in mathematics from the University of Budapest, with a dissertation about set theory. His axiomatization has left a permanent mark on the subject; and his definition of ordinal numbers, published when he was 20, has been universally adopted.

Von Neumann was *Privatdozent* ("lecturer") at Berlin

in 1926–29 and at the University of Hamburg in 1929–30. During this time he worked mainly on quantum physics and operator theory. Largely because of his work, quantum physics and operator theory can be viewed as two aspects of the same subject. He had a basic insight concerning vectors: the geometry of vectors in infinite-dimensional Euclidean space has the same mathematical characteristics as the structure of the states of a quantum-mechanical system.

In 1930 von Neumann was visiting lecturer at Princeton University; he was appointed professor in 1931. In 1932 he gave a precise formulation and proof of the "ergodic hypothesis" of statistical mathematics. His book on quantum mechanics, *Mathematische Grundlagen der Quantenmechanik*, published in 1932, is still a standard and one of the most inspiring treatments of the subject. In 1933 he became a professor at the newly founded Institute for Advanced Study, Princeton, keeping that position for the rest of his life. Meanwhile, he turned his attention to the challenge made in 1900 by a German mathematician, David Hilbert (*q.v.*), who proposed 23 basic theoretical problems for 20th-century mathematical research. Von Neumann solved a special case of Hilbert's fifth problem, the case of compact groups (see TOPOLOGICAL GROUPS AND DIFFERENTIAL TOPOLOGY).

Von Neumann algebras

In the second half of the 1930s the main part of von Neumann's publications, written partly in collaboration with F.J. Murray, was on "rings of operators" (now called Neumann algebras; see ANALYSIS, FUNCTIONAL). Of all his work, quite probably these concepts will be remembered the longest. Currently it is one of the most powerful tools in the study of quantum physics. An important outgrowth of rings of operators is "continuous geometry." Von Neumann saw that what really determines the character of the dimensional structure of a space is the group of rotations (see ALGEBRAIC STRUCTURES) that the structure allows. The groups of rotations associated with rings of operators make possible the description of space with continuously varying dimensions. In that context, for example, it makes sense to describe a space of dimension ¾. Hitherto, mathematicians had represented a given space only in terms of whole numbers. Abstracting from the "concrete" case of rings of operators, von Neumann formulated the axioms that make these continuous-dimensional spaces possible.

About 20 of von Neumann's 150 papers are in physics; the rest are divided more or less evenly between pure mathematics (mainly set theory, logic, topological group, measure theory, ergodic theory, operator theory, and continuous geometry) and applied mathematics (statistics; numerical analysis; shock waves; flow problems; hydrodynamics; aerodynamics; ballistics; problems of detonation; meteorology; and two nonclassical aspects of applied mathematics, games and computers). His publications show a discontinuous break from pure to applied around 1940.

During World War II, he was much in demand as a consultant to the armed forces and to the civilian agencies concerned with the problems of war. His two main contributions were his espousal of the implosion method for bringing nuclear fuel to explosion and his participation in the development of the hydrogen bomb. His political and administrative decisions were not usually on the liberal side. He continued to advocate atomic-bomb tests after the war.

The "minimax theorem"

The mathematical cornerstone of von Neumann's theory of games is the "minimax theorem," which he stated in 1928; its elaboration and applications are in the book he wrote jointly with Oskar Morgenstern in 1944, *Theory of Games and Economic Behavior*. The minimax theorem says that, for a large class of two-person games, there is no point in playing. Either player may consider, for each possible strategy of play, the maximum loss that he can expect to sustain with that strategy and then choose as his "optimal" strategy the one that minimizes the maximum loss. If a player follows this reasoning, then he can be statistically sure of not losing more than that value called the minimax value. Since (this is the assertion of the theorem) that minimax value is the negative of the one, similarly defined, that his opponent can guarantee for himself, the long-run outcome is completely determined by the rules.

In computer theory, von Neumann did much of the pioneering work in logical design, in the problem of obtaining reliable answers from a machine with unreliable components, the function of "memory," machine imitation of "randomness," and the problem of constructing automata that can reproduce their own kind. One of the most striking ideas, to the study of which he proposed to apply computer techniques, was to dye the polar ice caps so as to decrease the amount of energy they would reflect —the result could warm the Earth enough to make the climate of Iceland approximate that of Hawaii.

Among the many awards that von Neumann received are honorary doctorates from Princeton, Harvard, and Istanbul; a term as president of the American Mathematical Society; and memberships in several national scientific academies, including the United States National Academy of Sciences. President Eisenhower appointed him to the Atomic Energy Commission in 1955. He received the Enrico Fermi award in 1956, knowing that he was incurably ill. He died of cancer in Washington, D.C., on February 8, 1957.

The "axiomatic method" is sometimes mentioned as the secret of von Neumann's success. In his hands it was not pedantry but perception; he got to the root of the matter by concentrating on the basic properties (axioms) from which all else follows. His insights were illuminating and his statements precise.

BIBLIOGRAPHY. Von Neumann's *Collected Works* were published in six volumes between 1961 and 1963. A memorial issue of the *Bulletin of the American Mathematical Society* (vol. 64) was devoted to his life and works (May 1958). His last, incomplete, written work, *The Computer and the Brain* (published posthumously in 1958), has a touching preface by his widow.

Neuropteran

At one time the term neuropteran referred to many groups of insects that are now placed in other orders, so that, in a strict sense, the order Neuroptera now includes only the lacewings. Frequently, however, two other closely related insect groups are included also as neuropterans. These are the snakeflies (Raphidiodea), so-called for their body shape, and the dobsonflies and alderflies (Megaloptera). For completeness of discussion, therefore, all three groups, considered here as three separate orders, are included in this article.

General features

All three orders may have evolved from an early mecopteran (scorpionfly) ancestral stem, prior to the Trichoptera–Lepidoptera offshoot (see TRICHOPTERA). Carnivorous insects of varied structure and habit, both freshwater and terrestrial members of the three orders, are widely distributed with the exception of the snakeflies (which are confined to the Northern Hemisphere). Many members are important in the biological control of other insects and mite pests; others are "flies" of interest to anglers. Some of the most elegant and dainty insects are lacewings. There are more than 500 species of alderflies and dobsonflies, 80 species of snakeflies, and 4,000 species of lacewings.

The size of lacewings varies from 1.5 to over 35 mm in length and 2 to over 50 mm in anterior wing length. Lacewings are characterized by their many-veined pairs of wings, which give them a lacy appearance. Dobsonflies and alderflies, while similar in general appearance, have wings that appear heavier than those of lacewings; 15 to 30 mm in body length, their anterior wing length varies from 20 to more than 50 mm. Snakeflies are of medium size, 10 mm or more in body and anterior wing length, with wings similar to those of lacewings.

The Megaloptera and Raphidiodea differ from the Neuroptera in their prognathous adult heads (mouthparts anterior) and biting larval mandibles; the neuropterans have hypognathous adult heads (mouthparts ventral) and unique piercing–sucking larval jaws (composed of mandibles plus maxillae).

Natural history. *Life cycle of Neuroptera.* Neuropteran eggs may be laid loosely in light soil, cemented directly to a surface, or cemented on the end of a stalk. During the cementing process, the female places the tip of the abdomen on a surface and exudes a viscid fluid; she raises her abdomen slowly to draw the fluid into a slender filament. The fluid rapidly hardens, and the egg is attached by its posterior end to the top of the stalk.

Larvae hatch after 5–14 days unless the egg is in the overwintering stage. In some families a thickened portion of skin is used by the larva to break the egg; in others the egg splits. There are generally three larval stages in Neuroptera. The first lasts a few days, the second a few days or for the winter months, while the third varies, depending on the species, from weeks to months.

Neuropteran larvae are carnivorous and free-living with the exception of the aquatic family Sisyridae; these larvae are parasitic on freshwater sponges. Typically, a neuropteran larva sucks out the contents of its prey, leaving only a hollow skin. Although many lacewing larvae are nocturnal and need no camouflage, other species carry debris on bodies adapted for this purpose. In one family debris floats onto the hairs and is caught, while in another the larva takes debris in its jaws and places it on its back. In still another family larvae lie lightly covered in the soil; the lacewing larvae known as antlions dig conical pits in light, dry soil or sand. They vigorously throw out soil with their heads, then lie in the pit, body covered and jaws ready to grasp an ant or any other prey that may fall in. Should a captive attempt to escape by climbing a pit wall, the antlion will use its head to toss soil and cause the prey to fall back into the pit.

Larval feeding

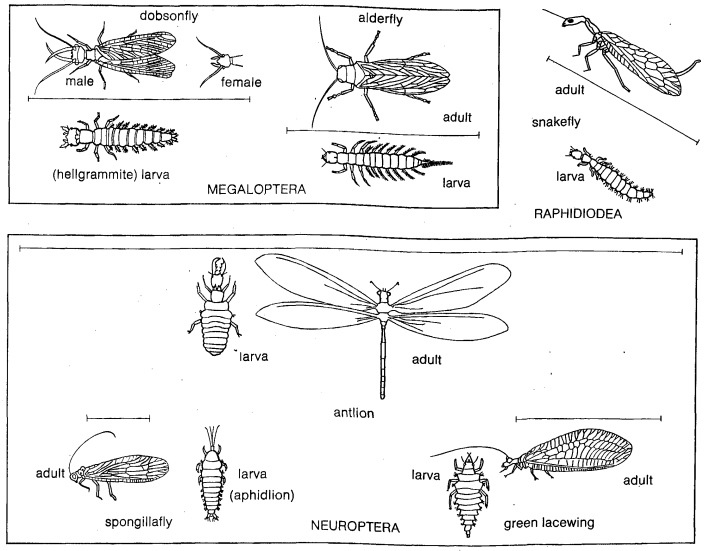
The neuropteran larva spins a double cocoon by exuding whitish or yellowish silk through its anus. First, a loosely woven cocoon is spun and fastened to a surface; then the larva spins a second tightly woven cocoon inside the first. This double construction is typical of neuropterans; the walls of the two cocoons may be closely attached or apart, depending on the species. The larva may spend the prepupal stage of several days or months within the cocoon before the pupal molt occurs. The limbs of the pupa are free. In a few species the pupa bursts from the cocoon, but most species use their functional mandibles to emerge. Adults appear either as the pupa leaves the cocoon or after it has reached a suitable position. Some species have two or more broods a year, although the life cycle does not exceed 12 months. Mating may occur anytime during the lifespan of the adult female; sometimes she holds fertilized eggs until weather conditions are suitable.

Life cycles of Megaloptera and Raphidiodea. Megalopteran females lay eggs above water level in masses of 3,000 or more. Larvae crawl into the water where they are very active. Commonly, alderfly larvae are associated with muddy bottoms of ponds and slow-moving streams, dobsonfly larvae with fast-flowing streams or rivers. The larvae are predaceous, usually nocturnal, and may leave the water to search for prey or to molt. Before pupation the larvae leave the water to form cells in damp, coarse soil beneath stones or debris. The pupa crawls from its earthen cell before the adult emerges. Raphidiodean females use slender ovipositors to lay eggs in the cracks of tree bark, and immature stages are found there.

Reproduction

Usually mating occurs at night. Sperm are passed either directly as spermatozoa or in a spermatophore that may project from the female after mating and be eaten by her, either wholly or in part. While megalopterans lay eggs in masses of several thousands, raphidiodeans lay them singly, and neuropterans lay 600 or 700 singly, separately in groups, or in batches.

Ecology. All of the larvae and adults are carnivorous and predaceous (except the parasitic Sisyridae), devouring enormous numbers of mites, insect eggs and larvae, ants, thysanopterans, psocopterans, and homopterans (especially aphids). For example, chrysopids consume large numbers of aphids and are a potent natural control mechanism. Larval feeding is continuous except during molts. Adults may also feed on dead insects, and females may feed on their own eggs and on spermatophores. The neuropteran family Sisyridae may have evolved from the closely related family Osmylidae whose larvae probe streamside mosses for prey.

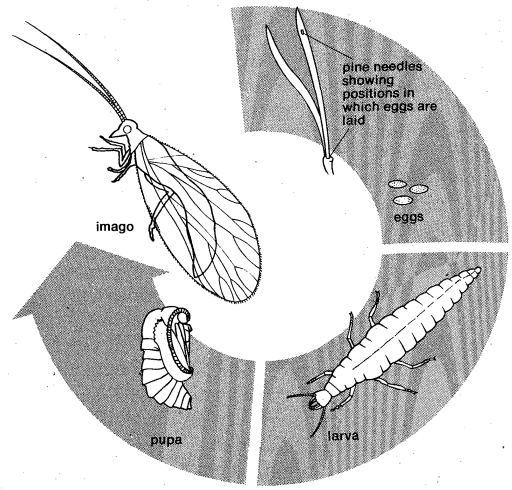


Diversity among neuropterans.
Line scales indicate approximate size of each insect.
From *Invertebrate Identification Manual* by Richard A. Pimentel, © 1967 by Litton Educational Publishing, Inc.; reprinted by permission of Van Nostrand Reinhold Company

Megalopteran larvae, a source of food for freshwater fish, are used as bait by anglers. Spiders capture both adults and terrestrial larvae in their webs. Neuropteran eggs and larvae are preyed upon by other neuropterans, other predaceous insects (*e.g.*, coccinellid, or ladybird beetle, larvae), ants, birds, and bats.

Form and function. *Adults.* Antennae are multisegmented in neuropterans and filiform in Megaloptera and Raphidiodea. The head is squarish or transverse in Megaloptera and Neuroptera, elongate in Raphidiodea; prognathous (mouthparts anterior) heads are typical of Megaloptera and Raphidiodea, hypognathous (mouthparts ventral) of Neuroptera. The four wings, which are similar in size, large or elongated, and translucent, contain many veins and crossveins and are sometimes coloured in brown or green shades. The ten-segmented abdomen lacks appendages (cerci); however, terminal claspers may occur. In both sexes, the last abdominal segment may be reduced, modified, or bear a pair of dorso-lateral groups of trichobothria (short, stiff sensory bristles). In females, abdominal terminalia are also variable; Raphidiodea have a long, thin ovipositor.

Immature stages. The larval head, prognathous in all three groups, may bear as many as seven simple eyes on each side or none. Megalopteran larvae have large man-



Life cycle of the brown lacewing *(Hemerobius stigma).*

Larval jaws

dibles, raphidiodean small, while larval jaws, the outstanding feature of neuropterans, may be short or long, straight or curved. In all neuropteran species each jaw is a sucking tube, mandible above and maxilla below, the two fitting together but separable. Since there is no true mouth, liquid food is sucked by a pharyngeal pump through the two jaws directly into the pharynx. The aquatic megalopteran larvae have lateral gill filaments and either a median caudal filament or anal prolegs; Raphidiodea lack abdominal processes. The terminal abdominal segments of neuropterans are modified into an anal proleg, with or without processes, which also functions as a spinneret (an organ for producing silk from internal glands); in several families the abdomen has special hairs or bristles (macrotrichia) for holding debris for camouflage.

Some pupae resemble adults, particularly the snakeflies. The hooks or ridges, aligned both forward and backward and evident on the dorsum of the abdomen of some pupae, may be used to maintain a position within the cocoon rather than to escape from it; mandibles and limbs are often used to aid in the escape.

Locomotion. Most larvae have walking legs except in those whose legs are modified for burrowing (Ithonidae and Myrmeleontoidea). The aquatic dobsonfly larvae have a pair of terminal abdominal prolegs that are important in moving in strong currents. The legs, abdominal filaments, and median tail filament of alderfly larvae bear long hairs, or setae, which help to propel them through their slow-moving water habitats. Snakefly larvae wriggle rapidly backwards to escape danger. Most neuropteran larvae have the tenth abdominal segment modified into a single anal foot for use in locomotion; a pair of hooked processes also assist in holding. Members of the three groups typically are weak fliers.

Evolution and Paleontology. The closely related dobsonflies, alderflies, and snakeflies may have arisen from a common stem that diverged from the ancestral mecopteran (scorpionflies) stem early in the Carboniferous Period (about 325,000,000 to 345,000,000 years ago); the Neuroptera may have diverged from the ancestral mecopteran stem later in the Carboniferous.

Fossil representatives of the modern families Sialidae and Corydalidae are represented in the Oligocene Baltic amber (about 26,000,000 to 38,000,000 years old). The present-day Raphidiidae are represented in the Oligocene Baltic amber. The present-day neuropteran family Psychopsidae was existent as early as the Triassic Period (about 190,000,000 to 225,000,000 years ago) in Australia. Other present-day neuropteran families, Coniopterygidae, Osmylidae, Sisyridae, Hemerobiidae, Chrysopidae, Myrmeleontidae, first occurred in the Oligocene Baltic amber, the Ascalaphidae in the Lower Miocene of France (about 36,000,000 years ago), and the Nemopteridae in the Miocene of Colorado.

Classification. *Distinguishing taxonomic features.* The adults of the orders Megaloptera, Raphidiodea, and Neuroptera have similar mouthparts, wings, and genitalia; they differ in shape of head and thorax. The presence of simple eyes (ocelli) and the morphology of antennae, head, wings (including venation), legs, and body, are criteria used to distinguish families.

Larval characters associated with the aquatic habit, such as abdominal filaments, separate Megaloptera and Sisyridae (Neuroptera) from terrestrial Neuroptera and arboreal Raphidiodea. The normal mandibles of Megaloptera and Raphidiodea separate these groups from Neuroptera, whose jaws are combined mandibles and maxillae. The form of the head, neck, mouthparts, legs, and body are used to separate neuropteran families.

Annotated classification. This is a combination of American and English classifications accepting three separate orders.

ORDER MEGALOPTERA

Adults medium to large; head prognathous; biting mouthparts; antennae threadlike (or filiform); 2 pairs of large, similar wings, held rooflike or nearly flat over abdomen in repose. Larvae elongated; biting mandibles; lateral abdominal gill filaments present; aquatic.

Family Sialidae (alderflies)

Wing expanse 20–40 mm; simple eyes (ocelli) absent; 4th tarsal joint of lower leg dilated and bilobed. Larvae with 7 pairs of lateral, segmented, abdominal filaments, median caudal filament; larvae without anal prolegs.

Family Corydalidae (dobsonflies)

Wing expanse 40–100 mm; 3 ocelli; 4th tarsal joint cylindrical; male mandibles sometimes enlarged. Larvae (called hellgrammites) with 8 pairs of unsegmented or imperfectly segmented, lateral, abdominal filaments; terminal filament absent; 2 anal prolegs.

ORDER RAPHIDIODEA (snakeflies)

Adults small to medium; head prognathous, elongated, tapering posteriorly; biting mouthparts; antennae filiform; prothorax elongated, cylindrical; 2 pairs of similar elongated wings, held rooflike over abdomen in repose; female with long, slender ovipositor. Larvae elongated; head and prothorax large, polished; biting mandibles; abdominal filaments and prolegs absent; arboreal.

Family Raphidiidae

Three ocelli; antennal joints smaller at bases.

Family Inocellidae

Ocelli absent; antennal joints cylindrical.

ORDER NEUROPTERA (lacewings)

Adults small to large; head hypognathous; biting mouthparts; antennae varied; 2 pairs of similar wings, held rooflike over abdomen in repose. Larvae elongated or broad; piercing–sucking jaws, elongated mandibles above, elongated maxillae below, form a hollow tube; abdominal filaments absent, except aquatic Sisyridae: paired abdominal prolegs absent; 10th abdominal segment modified as an anal proleg and/or spinneret.

Superfamily Ithonoidea

Adults large, mothlike; larvae with straight jaws, modified for subterranean predatory feeding.

Family Ithonidae (moth lacewings)

Adult wing expanse 30–79 mm; head small, closely set on prothorax; antennae long, filiform; costal areas of wings not broad. Larvae burrowing; legs adapted for digging, tibia fused to tarsus; mouthparts short, straight; maxillae enlarged.

Superfamily Hemeroboidea

Adults small to large; antennae filiform or moniliform (beadlike); posterior wings shorter than anterior; elongated larvae with jaws straight or curved.

Family Hemerobiidae (brown lacewings)

Adults small; antennae long, moniliform; ocelli absent; wings with few crossveins, hairy; female without projecting ovipositor. Larvae (called aphis wolves) elongated, with short incurved jaws; body smooth, fine hairs; free living on vegetation.

Family Sympherobiidae (brown lacewings)

Adults small, closely related to Hemerobiidae. Larvae spindle-shaped; head small; antennae short; jaws short, stout, slightly incurved.

Family Dilaridae (pleasing lacewings)

Small; antennae of male coarsely comb-shaped; head with 3 ocelli or ocellus-like tubercles; wings with numerous crossveins.

Family Psychopsidae (silky lacewings)

Adults large mothlike species; antennae short, wings broad. Larvae elongated, flat; head broad posteriorly, closely attached to prothorax; jaws incurved, large, sicklelike; often arboreal, under bark.

Family Osmylidae (osmylidflies)

Adults medium to large; head wider than long; antennae filiform, short; 3 ocelli; wings often with brown markings; claws with many teeth. Larvae elongated; long, slender, straight jaws, slightly upcurved; in margins of freshwater.

Family Polystoechotidae (large lacewings)

Adults medium to large; wing expanse 40–75 mm; antennae short. Larvae with short, sharp, incurved mandibles, maxillae stout, blunt; labial palpi, sensory appendages on labium (lower lip); leg 5-jointed; tarsal claws simple, slightly curved; knobbed structures (called empodia) between terminal elongated claws.

Family Sisyridae (spongillaflies)

Adults small, brownish, closely related to Osmylidae; ocelli absent. Larvae moderately elongated; hairy; antennae long, bristle-like; mandibles and maxillae elongated, with bristle-

like piercing structures; labial palpi absent; 1 tarsal claw per leg; aquatic; parasitic on freshwater sponges.

Family Chrysopidae (green lacewings, stinkflies)

Adults small to medium, yellowish, green, or grey, sometimes with markings; wing expanse 31–65 mm; head small, antennae long, filiform; iridescent golden eyes. Larvae (called aphid lions) moderately elongated, broad; long, sharp, incurved jaws; hairy; legs with long, trumpet-shaped empodium between claws; many with dorsal lumps on body, carry packets of debris dorsally; found on vegetation.

Family Apochrysidae (fragile lacewings)

Fragile, closely related to Chrysopidae; antennae filiform.

Family Berothidae (beaded lacewings)

Small, slender; antennae long, filiform; wings hairy, sometimes with seedlike scales; larvae thysanuriform, with short, straight jaws.

Family Trichomatidae

Small, slender, closely related to Berothidae; wings and body with long hairs.

Family Mantispidae (mantispids)

Adults small to large; wing expanse 10–55 mm; antennae short; prothorax greatly elongated; wings narrow; forelegs predatory. Larvae have 2 forms; 1st stage with squarish head; antennae with 3–5 joints; short, straight jaws; labial palpi with 3 or 7 joints; legs with 2 claws and empodium; active hunters; in presence of food reserves larvae become white, parasitic grubs with small heads and atrophied legs.

Superfamily Nemopteroidea

Adults highly specialized, related to Myrmeleontoidea; head prolonged in rostrum; posterior wings narrowed, elongated to twice anterior wing length. Larvae short, broad; with incurved jaws; varied in form, distinct from other neuropterans.

Family Nemopteridae (thread-winged or spoon-winged lacewings)

Adults delicate; head snoutlike; antennae short; posterior wings greatly elongated, ribbonlike or threadlike; often expanded distally to appear spoonlike. Larval antennae long, filiform; jaws incurved; mandibles with or without internal teeth; with or without an elongated neck formed by anterior portion of prothorax; if present, neck may be thin, cylindrical, 3 times length of head; remainder of thorax and abdomen broad, flattened; legs long, with 2 curved claws, empodium absent; free-living in dust, some cave dwelling.

Superfamily Myrmeleontoidea

Adults medium to large in size; antennae varied. Larvae short, broad; jaws incurved; mandibles with inner teeth.

Family Myiodactylidae

Adults resemble Osmylidae; antennae long, cylindrical.

Family Nymphidae (slender lacewings)

Adults slender; long, cylindrical antennae; wings long, slender. Larval neck slender; jaws incurved, mandibles with single internal tooth near midway.

Family Myrmeleontidae (antlions)

Adults with long, slender wings; bodies partly covered with fine hairs; dragonfly-like; antennae short, weakly clubbed or flattened distally. Larvae (called antlions or doodle bugs) short and broad; head small; neck bent to allow rapid upward and backward movement of head; mandibles strong, incurved; 3 internal teeth; posterior legs modified for backward burrowing, tibia fused to tarsus; abdomen hairy, plump, modified posteriorly for backward burrowing; lie in soil surface or at bottom of a self-made pit.

Family Stilbopterygidae (shiny wings)

Adults large, swift fliers; shining wings and bodies; dragonfly-like; antennae short, strongly clubbed; eyes large; larvae large, black, shiny; found in vegetable refuse.

Family Ascalaphidae (owlflies)

Adults large, dragonfly-like species, 40–50 mm in length; some rapid fliers; antennae long, slender, strongly clavate distally; eyes large, divided; head with long, fine hairs. Larval head squarish; mandibles elongated; 1 or several inner teeth; neck short, narrow; posterior legs with tibia fused to tarsus; body fringed.

Superfamily Coniopterygoidea

Tiny, somewhat separated from other Neuroptera; body, wings covered with white waxy powder; larvae straight-jawed; labrum large, covering the other mouthparts.

Family Coniopterygidae (dustywings)

Adult wing expanse 3–10 mm; antennae long, filiform wings similar or posterior wings reduced; wing venation reduced. Larvae, 5 simple eyes each side; 5-jointed antennae; jaws short or long, needlelike; legs with padlike empodium.

Critical appraisal. Standard English and American systems differ in their classification of Megaloptera and Raphidiodea as distinct orders.

The standard English classification is order Neuroptera, suborder Megaloptera, superfamily Sialoidea, family Corydalidae, family Sialidae, superfamily Raphidiodea, family Raphidiidae, and suborder Planipennia (true Neuroptera). In other systems, the suborder Megaloptera is termed suborder Sialodea, the combination Raphidioptera is used for Raphidiodea, and the name Neuroptera is used for a superorder containing the orders Megaloptera, Raphidioptera, Planipennia. The modern tendency is to separate the three main groups in the separate orders Megaloptera, Raphidiodea (or Raphidioptera), and Neuroptera (Planipennia), as in the classification above.

Although larvae of most families have been described, further breeding and detailed studies of life histories and morphology of immature and adult forms are required.

BIBLIOGRAPHY. C.T. BRUES, A.L. MELANDER, and F.M. CARPENTER, "Classification of Insects," *Bull. Mus. Comp. Zool. Harv. 108* (1954), keys to insect families; E.O. ESSIG, *College Entomology* (1942), a standard American text with the groups treated in separate orders; A.D. IMMS, *A General Textbook of Entomology*, rev. ed. by O.W. RICHARDS and R.G. DAVIES (1960), a standard English text; F.J. KILLINGTON, *Monograph of the British Neuroptera*, 2 vol. (1936–37), an excellent monograph; C.L. WITHYCOMBE, "Some Aspects of the Biology and Morphology of the Neuroptera, with Special Reference to the Immature Stages and their Possible Phylogenetic Significance," *Trans. Ent. Soc. Lond.*, pt. 3–4, pp. 303–411 (1924–25), a good source for information.

(K.A.J.W.)

Neutron

The neutron is one of two subatomic particles present in the nucleus of every atom (except of hydrogen), the other being the proton. These two particles, called nucleons, interact through the so-called nuclear force and bond in various proportions to form the different atomic species of the elements.

The mass of a neutron is slightly greater than the mass of the proton, and almost 2,000 times that of an electron; the proton has the basic positive electrical charge, whereas the neutron has no overall electrical charge. The neutron possesses an intrinsic angular momentum (a concept that relates the distribution of mass in a body and its angular velocity) and a magnetic moment; that is, it behaves like a minute magnet in ways that suggest that it is an entity of moving electrical charges. It appears, therefore, to have a rather complex structure. The free neutron, one not incorporated into a nucleus, is subject to radioactive decay of a type called beta decay. It can disintegrate, or decay, into a proton, an electron, and an antineutrino, a particle with no charge and no mass. Because of its beta decay, the neutron does not exist in nature in its free state but must be produced artificially. Because free neutrons are electrically neutral, they pass unhindered through the electrical fields within atoms and, therefore, constitute a penetrating form of radiation, interacting with matter almost solely through the relatively rare collisions with atomic nuclei.

Shortly after the discovery of the neutron in 1932, it was found that many elements may undergo a nuclear transformation, called fission, when bombarded by neutrons. In the same fission process a new "generation" of neutrons is produced. In 1942, it was demonstrated that neutrons were produced in the fission process in sufficient numbers to sustain a chain reaction in the world's first atomic pile (later called a nuclear reactor) at Chicago. This development led to the construction of the atomic bomb and to general realization of the tremendous capabilities of nuclear energy. After World War II, the neutron was shown to have great potentialities in nonmilitary fields. Subsequent developments led to the large-scale production of electrical power from nuclear energy and its use as the motive power for ships. The absorption of neutrons by nuclei exposed to the large neutron in-

tensities available in reactors has resulted in the production of large amounts of radioactive isotopes utilized in many fields, mainly as tracers, minute quantities of which can be followed, by the use of radiation detectors through various chemical, biological, and industrial processes. Radioisotopes are used in medicine to study normal tissue, to diagnose disease, and in some cases to treat various forms of cancer.

In addition, the neutron is an important tool in pure research. Since it is a basic constituent of all matter, knowledge of its properties and its structure is essential to an understanding of the structure of matter. Nuclear reactions induced by neutrons are valuable sources of information about the nucleus; an understanding of its interactions with matter reveals important data about the characteristics of the basic nuclear force.

HISTORY

In 1911, the British physicist Sir Ernest Rutherford demonstrated that most of the mass and all of the positive charge associated with an atom is concentrated in its centre. He proposed a basic model for the atom in which the positive charge resides in the nucleus (with a radius of 10^{-12}–10^{-13} centimetre), and the negative charge consisting of the electrons is distributed over a sphere with a radius comparable to the radius of the atom (about 10^{-8} centimetre). At that time, although the actual composition of the nucleus was unknown, it had been proposed that it might contain some electrons in addition to protons. As the properties of nuclear particles became better understood, it became clear, however, that free electrons could not exist within a space as small as that of a nucleus.

In experiments performed between 1918 and 1930, atoms of one element were transmuted into atoms of another by the use of high-speed alpha particles, the nuclei of helium atoms. In 1930, Walther Bothe and H. Becker, working in Germany, reported that a penetrating radiation resulted from the bombardment of beryllium with alpha particles. This radiation was interpreted as gamma radiation of very high energy—more penetrating than had been observed up to that time. Further studies of the properties of this radiation by Irene and Frédéric Joliot-Curie in France led to the discovery in January 1932 that its intensity was increased by passage through a screen containing a large hydrogen content. The Joliot-Curies first suggested that the protons in the screen were ejected in the process of gamma ray (γ-ray) scattering, but the kinetic energy associated with the gamma rays had to be unreasonably large in order to account for the observed effect.

Discovery

James Chadwick, working during this time at the Cavendish Laboratory in England, found that the radiations from the bombardment of beryllium by alpha particles could transfer kinetic energy to other light nuclei as well as they could to hydrogen. Late in February 1932, Chadwick published a paper in which he suggested that the protons (hydrogen nuclei) and other light nuclei were recoiling from elastic collisions with a particle of mass approximately equal to that of the proton. Charged particles react electrically with the atoms of matter into which they penetrate and are soon stopped; the postulated particle had to be electrically neutral in order to explain its high penetrating power. Chadwick called it the neutron, as suggested by Rutherford. Later in 1932 he showed that its mass is only slightly larger than that of the proton. It was then concluded that the atomic nucleus is composed only of protons and neutrons, both of which should be regarded as fundamental particles.

Within a few years after Chadwick's discovery, a large number of investigators throughout the world were studying the properties and interactions of the neutron. Its mass was more accurately determined. It was discovered that the (α, n) reaction, in which alpha particle bombardment results in the emission of a neutron, occurred for other light nuclei besides beryllium; and the inverse reaction (n, α) was also shown to exist. The latter indicated the neutron's ability to induce transmutations. In 1934, the Joliot-Curies produced artificially radioactive nuclei in an (α, n) reaction, and in the same year Enrico Fermi and his coworkers in Italy showed that radioactivity could be induced in a variety of elements through bombardment by neutrons. Further work by the Italian scientists indicated that neutron-induced radioactivity could be increased by reducing the velocities of the neutrons. In other words, it was found that slower neutrons were more likely to be captured by nuclei. This recognition was of great importance to the understanding of neutron reactions and led to the use of slow neutrons as the main tool in the investigation of nuclear structure until after World War II.

PROPERTIES OF NEUTRONS

Absence of charge. Perhaps the most important of the fundamental properties of the neutron is its effective lack of an electric charge. The proton, in comparison, has one unit of positive charge. As a result of this difference, neutrons and protons interact quite differently with matter. For example, a proton passing through matter interacts much more strongly because of the electrical forces between its charge and the negatively charged electrons in the atom and the positively charged nucleus and loses its kinetic energy rapidly. A neutron, on the other hand, loses kinetic energy only in its relatively rare collisions with an atomic nucleus. It therefore constitutes a highly penetrating type of radiation.

Mass. The most accurate determination of the mass of the neutron is obtained in nuclear reactions in which the masses and energies of all of the particles except the neutron mass are known. The first accurate determination was performed in Cambridge, England by the physicists Sir James Chadwick and Maurice Goldhaber in 1935, in a reaction in which a deuteron (an atomic nucleus of deuterium, a heavy isotope of hydrogen) was decomposed into a neutron and a proton by the absorption of gamma radiation (electromagnetic waves with very high frequencies). At the present time, the mass of the neutron is taken to be 1.67466×10^{-24}g, or 1.008665 atomic mass units. In comparison, the proton mass is 1.67236×10^{-24}g, or 1.007276 atomic mass units, and the mass of the hydrogen atom (one proton plus one electron) is 1.67327×10^{-24}g, or 1.007825 atomic mass units. (One atomic mass unit = 1 amu = $\frac{1}{12}$ of the mass of the C^{12} atom = 1.66028×10^{-24}g.)

Radioactivity. Because the mass of the neutron is larger than that of the proton, the free neutron can decay into a proton with the emission of an electron and an antineutrino.

Half-life

In this process, called beta decay of the neutron (from the symbol β for the type of electron emitted in radioactive decay), the maximum energy of the electron is 7,820,000 volts. Neutron beta decay was first observed in 1948. The most recent value for the half-life of the free neutron is 11.7 minutes (*i.e.*, one-half of any quantity of neutrons decays in 11.7 minutes). When a neutron is bound in a nucleus, however, electron emission is generally impossible, and the neutron is therefore stable.

Spin, magnetic moment, and statistics. The neutron has a spin; that is, it behaves somewhat as a top spinning on its axis, with the difference that only certain discrete values of the angular momentum associated with this spinning motion are possible. Not all possible values are allowed to the spin; only those that are multiples of a basic value, the quantum of action. As with the proton and electron, the value of the neutron's spin is ½ (in units of $h/2\pi$, in which h is Planck's constant, a physical constant that has the dimensions of energy times time. It must, furthermore, obey the Pauli exclusion principle, which says that two neutrons in the same atom cannot occupy the same quantum state. (The behaviour of a large number of neutrons is governed by the rules of one of the forms of quantum statistical mechanics, called the Fermi-Dirac statistics.)

The neutron acts like a small magnet, with north and south magnetic poles, when placed in a magnetic field; that is, it has an intrinsic magnetic dipole moment. The first accurate measurements of the neutron's magnetic moment were by Luis Walter Alvarez and Felix Bloch, United States physicists, in 1940. The presently accepted value is −1.91315 nuclear magnetons (one nuclear magneton = 0.505038×10^{-23} erg per gauss). The negative

sign means that the magnetic moment points in a direction opposite to that of the spin. The existence of a magnetic moment for the neutron is not consistent with the predictions of the theory developed by a British physicist, Paul Adrien Maurice Dirac, to explain the behaviour of electrons, and which should be applicable to all charged particles with spin ½. The neutron's magnetic moment is, therefore, considered to be anomalous, which implies that the neutron, being electrically neutral, must have a somewhat complex intrinsic structure of positive and negative charges.

Wave properties and classification. On the basis of the laws of quantum mechanics, which were developed to account for the fact that in some situations subatomic particles behave like waves, a wavelength can be associated with the neutron. In a general way, wave phenomena should be observed in the interaction between neutrons and matter when the wavelength is comparable to the size of the system with which it interacts. The wavelength is inversely proportional to the square root of its energy, and for computational purposes the value of the wavelength for a neutron with kinetic energy E can be written as

$$\lambda = \frac{2.86 \times 10^{-9}}{\sqrt{E}} \text{(centimetre)},$$

in which E is in units of electron volts. (The unit, electron volt [eV], which is equal to 1.6×10^{-12} erg, is an energy unit of convenient size in the world of elementary particles. Two commonly used multiples of this unit are 1 MeV $= 10^6$ eV $=$ 1,000,000 electron volts, and 1 keV $= 10^3$ eV $=$ 1,000 electron volts.) It is useful to classify neutrons according to their energy and wavelength, although any specific scheme is somewhat arbitrary. It is common to describe neutrons with energies in the range from 0 to 1,000 eV as slow, 1 to 500 keV as intermediate, 0.5 to 10 MeV as fast, and greater than 10 MeV as very fast. The slow neutron category can conveniently be subdivided into cold neutrons (energies below 0.002 eV), thermal neutrons (a most probable energy of 0.025 eV), and epithermal neutrons (energies greater than 0.5 eV). The wavelengths associated with these energies can be found by application of the above formula.

Structure. Valuable information about the properties of the basic nuclear force has been obtained from data on the scattering of neutrons by protons. According to current theory, many of these properties can be interpreted in terms of a picture in which both the neutron and proton are thought of as being surrounded by a "cloud" of charged or neutral pions, subatomic particles that react strongly with neutrons and protons, and are sometimes called pi mesons. When two nuclear particles (either two neutrons, a neutron and a proton, or two protons) are in close contact with each other, these pions are interchanged (transmitted) from one to the other and thus act as the "glue" that binds nuclear systems together. On the basis of such a theory, the neutron can exist part of the time as a proton and a negatively charged meson separated by about 10^{-13} centimetre. It therefore has an internal structure consisting of circulating electric charge that can give rise to its magnetic dipole moment. This picture implies furthermore that the neutron should have a small electrostatic interaction (electric attraction or repulsion) with electrons. The scattering of slow neutrons by atomic electrons has indicated the existence of various effects, the largest of these being due to the interaction between their magnetic moments. No effect specifically attributed to the internal structure of the neutron has, however, been observed in such experiments. Recent studies of the scattering of very energetic electrons from targets of hydrogen and deuterium, nevertheless, verify that the neutron has a complex internal structure that involves charged pions, although the details of the actual distribution of the charge are not at present well understood.

The neutron star. A major development in the field of astrophysics was the discovery (1967) of rapidly pulsating objects called pulsars that are sources of radio waves in space. Although the physical nature of these sources of radiation is not entirely certain, it has been thought that they might be so-called neutron stars. In the theory of stellar evolution, the neutron star is the name given to a star in one of its final stages of life. It is a small object (radius about six miles) and extremely dense. Although the exact constitution is unknown, it is thought to consist primarily of free neutrons held together, and thus stabilized against beta decay, by immense gravitational forces. If this conjecture concerning the pulsars is correct, then these stars constitute the first direct evidence of the existence within the universe of matter consisting almost completely of free neutrons.

NEUTRONS AS COMPONENTS OF NUCLEI

On the basis of present theory, the nucleus of any atom more massive than the lightest isotope of hydrogen is composed of one or more neutrons and protons. These particles, along with electrons, are the fundamental building blocks of the chemical elements. According to the neutron-proton model, the number of protons in the nuclear core is equal to the number of electrons in the atom and is therefore equal to the atomic number Z. The mass number A of the atom, defined as the closest integer to the actual atomic weight, is equal to the sum of the number of neutrons and protons in the nucleus since each of these particles contributes approximately one unit of atomic weight. The number of neutrons, N, in the nucleus is then equal to the difference between the mass number and the atomic number (*i.e.*, $N = A - Z$). Isotopes of an element all have the same number of protons and electrons, but differ from each other in mass number, and thus contain different numbers of neutrons. For example, the three stable isotopes of hydrogen each contain one proton; normal hydrogen has no neutrons and $A = 1$, deuterium has one neutron and $A = 2$, and tritium has two neutrons and $A = 3$.

The protons and neutrons in the atomic nucleus are bound together by the strong nuclear force. The strength of this binding, the so-called binding energy, can be deduced from the actual nuclear masses by use of the relationship describing the equivalence of mass and energy derived on the basis of the special theory of relativity. Thus, if ΔM is the decrease in mass that occurs when a number of neutrons, protons, and electrons combine to form an atom, the binding energy is

$$\Delta E = (\Delta M) \times c^2,$$

in which c is the velocity of light. The magnitudes of nuclear binding energies are very large indeed. For the case of an atom of helium, for example, the binding energy is 28 MeV, which means that it takes 28,000,000 electron volts energy to dissociate the helium nucleus into its four constituent particles. By comparison, it takes only a few electron volts to separate an electron from an atom. The tremendously higher energy involved in the case of nuclear transformations arises from the exceedingly strong forces that exist between the particles in the atomic nucleus.

Similarity to proton

In some circumstances it is useful to stress the similarities between the neutron and proton rather than their differences. In this respect, the neutron and proton can be regarded as two different manifestations of a single entity called the nucleon. The nucleon is therefore similar to an entire atomic nucleus, in that it can exist only in discrete quantum states; that is, it exists only in certain states characterized by definite values of charge and total internal energy. The neutron, with zero charge, is a state of the nucleon at a slightly higher energy (due to its slightly larger mass) than the proton, which is the nucleon state of +1 unit of charge.

NEUTRON PRODUCTION

For experimental purposes, free neutrons must be produced artificially. They can be released in nuclear reactions initiated by either charged particles or gamma radiation and in the nuclear fission process.

Energy. Neutron sources based on nuclear reactions provide monoenergetic neutrons (numbers of neutrons having the same energy) as well as neutrons having a broad continuous spectrum of energies. In any reaction, energy is either released or absorbed. Which occurs in a

particular case can be determined from a knowledge of the rest masses (a particle at rest has less mass than one in motion, due to relativity) of the atomic nuclei involved in the reaction along with the known energy-mass equivalence as expressed in the relationship, $E = mc^2$. If the energy associated with the sum of the nuclear masses of the projectile and the target is greater than the sum of the masses of the neutron and the final nucleus, then the reaction is exoergic and energy is released. If not, the reaction is endoergic, and energy must be absorbed in order for neutrons to be emitted. In endoergic processes the energy necessary for the reaction must be supplied by the kinetic energy of the relative motion of the target and projectile.

Yield. The yield of neutrons in a reaction depends on the nature and kinetic energy of the incoming nuclear particle. In a quantitative determination of the yield, it is necessary to know the reaction cross section. This quantity represents the probability that the nuclear particles will interact. In an elementary sense the cross section is the target area presented by a nucleus to an incoming particle, although this picture should not be taken too literally since the cross section depends not only on this area but also on the energy of the incident (incoming) particle and on the angle at which the final particles are emitted. The unit of cross section is called the barn (one barn $= 10^{-24}$ cm^2). The yield of a neutron-producing reaction is the number of neutrons produced for each particle incident on the nucleus.

Reactions initiated by charged particles. Most nuclear reactions that lead to the production of neutrons involve the bombardment of light atomic nuclei by incident protons, deuterons (heavy-hydrogen nuclei), or alpha particles. Other light projectiles such as ^{3}H (tritium) nuclei and ^{3}He (helium-3) particles are also used. These nuclear particles are usually accelerated to high kinetic energies in various kinds of nuclear accelerators, such as the Van de Graaff generator, the Cockroft-Walton generator, the cyclotron, and the proton synchrotron.

Symbols

A shorthand notation has developed for use in specifying nuclear reactions. The target nucleus and product nucleus are indicated by the chemical symbol of the element with a superscript giving the mass number A of the nucleus. (For example, ^{12}C represents the carbon isotope of mass 12.) The smaller incoming and outgoing particles are denoted n for neutron, p for proton, d for deuteron, α for alpha particles, and γ for gamma rays. Thus, the reaction ^{7}Li$(p, n)^7$Be means that one atom of lithium, mass 7, absorbs a proton, p, and emits a neutron, n, which leaves one atom of beryllium, mass 7.

Those reactions that yield the most copious supplies of monoenergetic neutrons are proton–neutron (p, n) and deuteron–neutron (d, n) reactions. The most common (p, n) sources are lithium and tritium, and these reactions are usually written: ^{7}Li$(p, n)^7$Be reaction (*i.e.*, $^7\text{Li} + p \rightarrow {}^7\text{Be} + n$) and ^{3}H$(p, n)^3$He reaction. Both processes are endoergic. High yields of monoenergetic neutrons with energies up to 4 MeV can be supplied from accelerated proton beams.

Most deuteron–neutron (d, n) reactions are exoergic; that is, energy is released in the interaction. These reactions can therefore supply neutrons of higher kinetic energy. The ^{2}H$(d, n)^3$He and ^{3}H$(d, n)^4$He reactions are frequently used for the production of monoenergetic neutrons at various energies between 2 and 30 MeV. Both of these reactions yield copious amounts of neutrons even at small deuteron kinetic energies.

Neutron yields in reactions induced by accelerated alpha particles are generally lower than in the reactions discussed above. Nevertheless, under certain conditions, the ^{9}Be$(\alpha, n)^{12}$C reaction, for example, is a useful source of monoenergetic neutrons. This reaction is, moreover, the principal process responsible for neutron production in radioactive materials used as neutron sources. In such sources the neutrons are produced in the bombardment of beryllium by the alpha particles emitted by naturally radioactive elements. In contrast to accelerator-produced neutrons, the neutrons emitted from radioactive sources are not monoenergetic.

Neutrons constitute an appreciable fraction of cosmic radiation as observed at sea level. Both slow neutrons and the very fast neutrons that arise from nuclear disintegrations in the atmosphere have been observed. From a practical point of view, such neutrons do not constitute a useful supply of neutrons at sea level because of their low intensities.

From the point of view of a radiation health hazard, it is fortunate that the intensities are low at sea level. These neutrons do, however, change some nitrogen atoms of the atmosphere into radioactive carbon atoms that have been found useful in radiocarbon dating (see DATING, RELATIVE AND ABSOLUTE).

Reactions initiated by gamma radiation. Neutrons can be produced in reactions initiated by gamma (γ) rays whose energy is greater than the energy with which the neutron is bound in the nucleus. This means that for most nuclei the γ-ray energy must be greater than about MeV. For beryllium and deuterium, however, the thresholds for neutron production by γ-ray absorption are 2.225 MeV and 1.66. MeV respectively. Some radioactive neutron sources utilize these low (γ, n) thresholds. Depending upon the particular radioactive material, approximately monoenergetic neutron sources with energies from 25 keV to about 1 MeV have been constructed.

A high-intensity (γ, n) neutron source can be obtained from the acceleration of large electron currents in linear accelerators, betatrons, and synchrotrons. When accelerated electrons strike targets made of heavy elements, they give rise to intense X-rays. The energy spectrum of this radiation is continuous up to a gamma-ray energy equal to the kinetic energy of the accelerated electrons. For electrons accelerated to energies in excess of the neutron-binding energies of nuclei, copious supplies of neutrons can be produced. The neutrons produced in such a source are not monoenergetic. The spectrum is relatively broad and resembles the energy spectrum that is obtained in the nuclear fission process.

Neutron energies in a reactor

Fission reactions. The nuclear reactor, or pile, is a very useful source of neutrons, especially of slow ones. The basic reaction involved is nuclear fission, during which neutrons are emitted. In the fission of uranium-235, ^{235}U, for example, an average of 2.5 neutrons are produced per fission. The spectrum of neutron energies is very broad—up to at least 8 MeV, with an average value of about 1.5 MeV. This is not, however, the energy distribution observed either inside or emerging from a nuclear reactor, because of the various slowing-down processes of neutrons in the reactor's moderating material (see below). The kinetic energies of the neutrons from a pile are classified as fast neutrons, epithermal and intermediate-energy neutrons, and thermal neutrons. From a practical viewpoint, the reactor is not a useful source of fast neutrons (although strong sources of fast neutrons can be obtained by bombarding ^{235}U by slow neutrons from the reactor), but both intermediate and thermal neutrons from a reactor are used extensively in nuclear studies.

NEUTRON CONTROL AND MANIPULATION

Moderation and diffusion. Beams of slow neutrons are usually obtained from a nuclear reactor, in which the fission reaction produces neutrons with kinetic energies up to 8 MeV. The process of reducing their energies into the slow range involves successive interactions with the many nuclei in the large mass of material (usually graphite or water) within the reactor. The career of a neutron in such a medium is usually divided into two parts. The first phase, called neutron moderation, is the process by which the neutrons are slowed down by elastic scattering collisions (see below) with the nuclei of the medium until they have attained thermal energies. A knowledge of the distribution of neutron energies and neutron positions within the moderating material during the slowing-down process is important to the design of nuclear reactors, as well as to the subsequent use of the neutrons. The second phase, referred to as thermal neutron diffusion, involves the subsequent collisions that the neutron encounters until it is finally captured by an atomic nucleus or it escapes from the medium. In this phase the loss of energy of the neu-

trons in collisions with nuclei is as probable as energy gains, and their behaviour is very similar to that of atoms in a gas. The individual neutron energies are distributed around an average value according to the so-called Maxwell distribution, a statistical distribution based on the principles of the kinetic molecular theory. Their behaviour is described by a diffusion equation, which is directly analogous to the basic differential equation of heat conduction.

Absorption. In the interaction between neutrons and bulk matter, the disappearance of the neutrons from the medium is governed by the probability of neutrons being absorbed by the nuclei of the medium. This probability is a mathematical one and is expressed as the cross section for the capture of neutrons; the larger the cross section, the greater the probability of absorption. For slow neutrons, the capture cross sections vary inversely as the square root of the kinetic energy, or velocity, of the neutron (so-called $1/v$ law, v being the neutron velocity), except at capture resonance (see below), where the probability for neutron absorption can become extremely large. Strong selective absorption by a nucleus over a narrow energy region, called resonance absorption, is superimposed over the $1/v$ absorption for slow neutrons.

Velocity selectors. The neutrons emerging from a mass of moderating material have a continuous distribution of kinetic energies in the thermal and intermediate energy regions. Some experiments require devices to separate neutrons of specific energies from the continuous spectrum emitted by the source. Such devices are called monochromators or velocity selectors. Mechanical velocity selectors, also called choppers, are utilized for neutrons in the thermal and somewhat higher energy range. A chopper is basically a high-speed rotating shutter designed to allow the passage of neutrons of only one specific velocity. Crystal velocity selectors select neutrons of a specific low energy by the process of reflecting neutrons from a single crystal of a metallic material that absorbs high-energy neutrons. A polycrystalline substance can also be used to filter out all short-wavelength neutrons to produce cold neutrons; that is, neutrons with kinetic energies less than 0.002 eV.

Choppers

Velocity selection can also be achieved by neutron time-of-flight techniques. Such methods have been utilized over a broad energy range that extends from thermal energies up to the fast neutron region. A short burst of neutrons is produced and their velocities determined by the measurement of the time taken to travel a known distance. Neutron sources can be pulsed charged-particle accelerators, which produce the neutron bursts in nuclear reactions, or reactors in which the emergent beam is broken into pulses by use of mechanical choppers. Flight times are measured by electronic techniques; by the use of multichannel methods, various neutron velocities can be selected simultaneously.

Neutron optics. Slow neutrons have wavelengths of about 10^{-8} centimetre, about the same size as the distances between atoms in crystals. They are, therefore, well suited for studies of the structure of crystals—that is, the pattern of the atoms in ordered matter. In the scattering of neutron waves by crystalline materials, the scattered wave interferes with the incident wave in much the same manner as in the case of X-ray scattering. The scattered intensities in different directions vary as a result of the particular relationship between the neutron wavelength and the location of the atoms in the crystal. The intensity patterns exhibited in this diffraction process can be analyzed to give details of the atomic structure of the material. The structure of organic crystals and of crystals with magnetic properties have been studied extensively by neutron-diffraction methods.

As with light waves, neutrons can also be reflected or refracted in passing across the boundary between two media. Slow neutrons, however, show these optical phenomena in the interaction with various metals—bismuth, lead, beryllium—rather than with optically transparent materials. Just as light exhibits the phenomenon of total reflection by a mirror, the total reflection of neutrons will also be observed if the neutron is incident on a neutron "mirror" (beryllium for example) at extremely small glancing angles. Such techniques allow the determination of the neutron index of refraction of a medium and have yielded valuable information on the interaction between neutrons and nuclei.

NUCLEAR REACTIONS PRODUCED BY NEUTRONS

When a neutron collides with an atomic nucleus, a variety of nuclear reactions can occur. The lack of charge enables even very low energy neutrons to initiate such reactions. In comparison, the positively-charged proton can interact with the nucleus only when its energy is high enough to overcome the repulsive force between its charge and the charge on the nucleus. The probability that a particular reaction will happen depends on the nature of the nucleus and the kinetic energy of the incident neutron. It is determined by the cross section for the reaction.

Energy dependence of cross sections. The manner in which a neutron interacts with a nucleus to initiate a reaction depends upon its kinetic energy. For slow neutrons the wavelengths are larger than nuclear sizes and therefore the probability for interaction is very large. When the neutron is captured it shares its energy with all the nucleons in the target and a compound nucleus is formed with a total energy, called energy of excitation, equal to the initial kinetic energy of the neutron plus the binding energy of the neutron in the compound nucleus. The compound nucleus, which can be thought of as a system at a high temperature, subsequently "cools down" by emitting a neutron, a γ ray, or a charged particle, and the final nucleus is left in one of its allowed quantum states. Which particle will be emitted depends on the properties of the compound and final nuclear states. For incident neutron energies, such that the total excitation energy corresponds to the actual energy states of the compound nucleus, a compatability called nuclear resonance occurs, and the probability for capture becomes very large. The cross section for absorption or for a particular mode of decay exhibits peaks in energy, which are fairly well defined. The occurrence of resonance behaviour increases for intermediate and higher energy neutrons. These neutrons have wavelengths comparable in size to nuclear dimensions and are therefore useful probes of the details of nuclear structure. In the fast and very fast energy region, neutron wavelengths become smaller than the average distance of separation of the nucleons in the nucleus, and there is a good chance that the neutron will pass through (or close to) the nucleus without being captured at all. The cross sections do not appear to depend sensitively on the internal structure of the nucleus. An almost perfect analogy is the scattering and absorption of light by a semitransparent medium. In this sense, the nucleus acts like a "cloudy crystal ball." At these energies, furthermore, the neutron has a good chance of interacting with only a single nucleon, which is then emitted in what is called a direct nuclear reaction. If it is absorbed to form a compound nucleus, the nuclear temperature is so high that individual capture resonances are not observed. Instead, the system behaves as if many quantum states are excited at the same energy, and the chance that many particles will be emitted in the decay becomes large. Some of the nuclear reactions induced by neutrons are discussed in the following sections.

Elastic scattering and total cross sections. The simplest process that occurs when a neutron strikes a nucleus is elastic scattering. It occurs for all nuclei at all neutron energies. In this process, the sum of the kinetic energies of the nucleus and the neutron after the collision is equal to the kinetic energy of the neutron before the collision. (It is assumed that the target nucleus is initially at rest and thus has no kinetic energy.) Experimentally, the nucleus in its lowest state of internal motion experiences a recoil, and the neutron with reduced energy is re-emitted in a different direction.

Elastic scattering is the dominant neutron reaction at intermediate and high energies for light- and medium-weight nuclei. At those energies at which the cloudy-crystal-ball model (see above) of the nucleus becomes important, the reaction is sensitive to the shape of the nucleus,

and the effective sizes of nuclei (*i.e.*, their radii) can be determined from measured scattering cross sections.

Total cross section

A given nucleus generally has a different cross section for each type of nuclear reaction that can take place, and the sum of the cross sections for all processes is called the total cross section. Thus, the total cross section is a measure of the number of neutrons that are removed from an incident beam in passage through matter. The difference between the total cross section and the scattering cross section is the absorption cross section.

Radiative capture. At low neutron energies, in the thermal and near-thermal region, the most probable reaction following neutron capture is the emission of gamma rays. At those energies for which resonances occur, the cross sections are exceedingly large. For indium, for example, the capture cross section at a resonance at 1.4 eV is more than 100 times that at nearby nonresonant energies. As the neutron energy is increased into the intermediate region, the neutron–gamma (n, γ) reaction becomes less probable because of increased competition with resonant elastic scattering, at least for light nuclei. For heavier elements, on the other hand, radiative capture is still the predominant process. Much of present knowledge of the energy-level structure of medium-weight and heavy nuclei has been obtained in studies of the (n, γ) reaction.

Emission of neutrons. Inelastic scattering results from a collision in which the incident neutron transfers some of its kinetic energy into internal energy of the target nucleus and a neutron is re-emitted at very much reduced energy. Such a process requires that the kinetic energy of the incident neutron be greater than the energy of the lowest excited state of the nucleus. Depending on the particular target nucleus, the neutron must have a minimum energy of between about 0.05 and 5 MeV before inelastic scattering can occur. After the inelastic process has taken place, the target nucleus can decay to a state of lower energy by the emission of γ rays.

In the very fast neutron energy region (kinetic energies of 10 MeV or higher), neutron-induced reactions frequently lead to the subsequent emission of two neutrons. In light nuclei, this (n, $2n$) reaction takes place almost as often as neutron scattering or charged-particle emission. For heavy nuclei, the emission of two neutrons is considerably more probable than is charged-particle emission.

Emission of charged particles. Reactions that are initiated by neutrons and result in the emission of one charged particle are the inverse of those in which neutrons are produced by charged-particle bombardment. The most frequently encountered neutron-induced reactions are those in which protons and alpha particles are emitted. The process of proton emission occurs only after the absorption of energy (that is, the process is endoergic, except for a few lightweight target nuclei), and the cross sections for such reactions are appreciable only for fast neutrons for which the available energy for the charged particle is large. In (n, α) reactions energy is often released and hence they can be initiated by slow neutrons.

Fission reactions. Nuclear fission is a mode of disintegration in which a heavy nucleus is split into two fairly massive intermediate-weight fragments with the release of a large amount of energy (about 200 MeV). One or more neutrons can also be released in the process. Some elements fission spontaneously or, as with many uranium and plutonium isotopes, fission can be induced by slow neutrons. For such heavy elements, fission occurs with a high probability after the capture of a neutron. For example, the fission cross section for the uranium-235 isotope, ^{235}U, is about 83 percent of the total slow neutron absorption cross section. For some heavy nuclei, however —notably ^{238}U and ^{232}Th—fission occurs only after the capture of neutrons with kinetic energies greater than about 1.5 MeV.

Other reactions. At ultrahigh neutron energies (between 50 and 500 MeV), multiple particle emission can take place. The reaction appears to proceed in two stages. In the first, the incident neutron interacts with individual nucleons in the nucleus, and these are emitted by a direct reaction. In the second step, the remaining nucleons in the system redistribute their energies to form a very highly excited compound nucleus, which decays by the emission of gamma rays and one or more nuclear particles.

DETECTION OF NEUTRONS

Methods of detection

In experiments that involve neutrons, one of the most important problems is their detection. Since the neutron is electrically neutral, it does not produce direct ionization. Neutron detection must therefore be based on secondary effects that result from their reactions with nuclei. For fast neutrons, the most important detection techniques make use of the ionization by light nuclei recoiling after elastically scattering neutrons. Other methods for detecting both fast and slow neutrons rely on neutron-induced charged-particle reactions in which the ionization produced by the charged particle is detected on the observation of fission fragments from neutron-induced fission or on a measurement of the radioactivity induced in a suitable nucleus by neutron capture or some other neutron reaction.

The design of apparatus for neutron detection involves a number of considerations. The technique to be used is selected on the basis of the application, the efficiency with which the secondary radiation is collected, the response to neutrons as a function of their energy, the sensitivity to background effects, and the response to neutrons sharply defined in energy.

Detection of fast neutrons; recoil detectors. In the process of elastic scattering, the positively charged target nucleus struck by a neutron with kinetic energy E will recoil with an energy E' as shown by the formula in which m and M are the masses of the neutron and recoiling nucleus, respectively, and θ is the angle between the directions of motion of the recoil nucleus and the incident neutron:

$$E' = \frac{4\,mM}{(m+M)^2} E\cos^2\theta.$$

For most efficient detection, a large recoil kinetic energy is desirable. The formula shows that the largest recoil energies occur when m and M are equal, so that scattering from protons (that is, from hydrogen nuclei) is preferred, since the mass of a proton is almost identical with that of a neutron. The main disadvantage to recoil proton detection is that the recoil particles will have all energies from 0 to E for incident monoenergetic neutrons with energy E.

Typical proton recoil detectors are ionization chambers and proportional counters filled with hydrogenous gases, scintillators made of hydrogenous compounds, photographic emulsions, and cloud and bubble chambers. A single gas recoil counter is useful for the simple detection of neutrons. To determine the neutron energy as well, combinations of two or more recoil counters have been used to measure both the energy of the recoiling proton and its direction of emission. The simultaneous detection of energy and angle of emission of the recoil proton is possible in photographic emulsions and in cloud and bubble chambers, since in these devices the path travelled by a charged particle appears as a visible track. In recent years, fast-neutron recoil counters based on the scintillation method of detection have been widely used. These devices are usually organic crystals or organic liquids and plastics. The light output from the scintillators, which is generated from the passage of a recoil proton, is amplified by a photomultiplier tube and its associated electronic circuitry. Recent advances in such techniques have made it possible to reduce the sensitivity of the device to background radiations, particularly γ rays. These counters are well adapted to the investigation of complex neutron spectra by methods that utilize velocity selection by time of flight.

Scintillation counters

Detection of fast neutrons. *By nuclear reactions.* In order for a neutron-induced reaction to be useful as a detector over a fairly wide energy range, the reaction cross section should be large and should vary smoothly with neutron energy. There should, furthermore, be few if any alternative reactions. Generally, therefore, the choice is limited to reactions involving light nuclei. Both the $^3He(n, p)^3H$ and $^6Li(n, {}^3H)^4He$ reactions satisfy these requirements in the fast-neutron region moderately well,

although there are some background effects. Both have been employed as fast-neutron detectors—^{3}He as a gas-filled proportional counter, and ^{6}Li in the form of a thin foil. The $^{10}B(n, \alpha)^{7}Li$ reaction, which is widely applied for neutron detection at low energies, has a prohibitively small sensitivity to fast neutrons. If, however, proportional counters containing boron bifluoride, BF_3, gas are immersed in a moderating medium such as paraffin, a detector that has a uniform response for neutrons from very low energies up to about 5–6 MeV can be constructed. This "long counter" is used widely in applications that involve the simple detection of neutrons.

The fission reaction utilizing uranium, ^{238}U, and thorium, ^{232}Th, as targets can serve as detectors for fast neutrons with energies greater than about 1.5 MeV. Since the energy release in the fission reaction is extremely large, the ionization pulses from the fragments are great enough to be distinguished from any background effects, and efficient neutron detection is possible.

By radioactivity. Fast neutrons can be detected by radioactivation techniques. Either the fast neutrons are slowed down to thermal energies and then captured in a material that becomes radioactive, or the fast neutrons themselves produce radioactivity in a nuclear reaction. Processes such as (n, p) and $(n, 2n)$ reactions, which are sensitive only to neutrons in the MeV range, frequently lead to radioactive final nuclei and can serve as such fast-neutron detectors. The radioactivity is detected after neutron irradiation by measuring the subsequent alpha, beta, and gamma radiation with suitable counters.

Detection of slow neutrons. *By nuclear reactions.* The most widely usable detector for slow neutrons is based on the reaction $^{10}B(n, \alpha)^{7}Li$ reaction. In this reaction, a large amount of energy is released. Hence the ionization caused by the alpha particle is easily observed, and background radiations can be eliminated. The cross section at thermal energies is very large and is a smooth function of energy up to about 1 keV.

The standard detector for slow neutrons has for many years been the BF_3 gas proportional counter or a similar counter with ^{10}B used as a wall coating of the counter tube. For time-of-flight experiments with slow neutrons, however, improved time resolution (because of faster time response of scintillators) can be obtained by the use of various boron scintillators, including scintillating liquids with boron dissolved in them. Other scintillating materials which contain ^{6}Li are also used for slow-neutron applications. The $^{6}Li(n, {}^{3}H)^{4}He$ reaction, which has been mentioned in connection with fast-neutron detection, has a large cross section for thermal neutrons as well. Although their efficiencies are lower than for boron counters, detectors with good time resolution have been constructed.

Some heavy nuclei have very large cross sections for fission by slow neutrons. Fission detectors consisting of thin foils of ^{233}U, ^{235}U, and ^{239}Pu are efficient detectors of neutrons in this energy range.

By radioactivity. The neutron radiative-capture process can be used as a detector of slow neutrons. In this application, either the γ rays following neutron capture are detected or (if the final nucleus is radioactive) the induced activity can be measured. To measure the radioactivity, the most important requirements are that the half-life be in a convenient range (a few minutes to several days) and that the capture cross section be large. Capture in manganese, indium, and iodine, for example, satisfy these requirements and are frequently used.

BIBLIOGRAPHY

Technical: General surveys of neutron physics include B.T. FELD, *The Neutron*, vol. 2 of *Experimental Nuclear Physics*, ed. by E.G. SEGRE (1953); and E. AMALDI, "The Production and Slowing Down of Neutrons," in *Encyclopedia of Physics*, vol. 38/2, ed. by S. FLUGGE (1959). Specific aspects of neutron physics are covered in *Fast Neutron Physics*, 2 vol., ed. by J.B. MARION and J.L. FOWLER (1960); and D.J. HUGHES, *Pile Neutron Research* (1953). I. KAPLAN, *Nuclear Physics*, 2nd ed. (1963), is a textbook at the college level.

Non-technical: D.J. HUGHES, *The Neutron Story* (1959); A. ROMER, *The Restless Atom: The Awakening of Nuclear Physics* (1960); B.L. COHEN, *The Heart of the Atom: The Structure of the Atomic Nucleus* (1967); C.J. CRAVEN, *Our Atomic World* (1963); and C.E. SWARTZ, *Microstructure of Matter* (1965), are particularly recommended.

(A.J.El.)

Nevada

Located in a mountainous region that includes vast semi-arid grasslands and sandy alkali deserts, Nevada is a part of the Old West that in 1864 became the 36th state of the United States. Its name was derived from the Spanish *nevada* ("snow clad"), a reference to the high mountain scenery of the Sierra Nevadas on its southwestern border with California. Other neighbouring states sharing many of the geographical features that make Nevada the nation's most arid state are Oregon and Idaho to the north, Utah to the east, and Arizona to the southeast. The 1970 population of about 489,000 living amid 110,540 square miles (286,299 square kilometres) made Nevada, after Alaska and Wyoming, the most sparsely populated state in the nation. During the 1960s, however, its percentage gain in population was the largest of the 50 states. Its capital, Carson City, remained the smallest state capital in the nation.

Location and general character

The Nevada of the 1970s appears far removed from the days when Virginia City was a fabled frontier town, thriving upon the rich silver mines of the Comstock Lode. Many frontier qualities persist, though subtly transformed into a sophisticated urban environment. The prospector digging against odds to find a bonanza has been replaced by the fortune seekers in the gambling casinos of Las Vegas and Reno, and the erstwhile "saloon diversions" have evolved into the lavish nightclub entertainment.

While the great majority of Nevadans live in the two main cities—more than one-half of them in the Las Vegas metropolitan area and one-fourth in that of Reno—the vast undeveloped lands of the state provide a largely unknown resource for the future. Combined with the major scientific activity related to the federal government's atomic-research facilities, the modern cities and desert reaches make Nevada a fascinating and unique phenomenon among American states. For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; NORTH AMERICAN DESERT; and SIERRA NEVADA RANGE.

THE HISTORY OF NEVADA

Anthropological evidence indicates that prehistoric Indian settlements existed in Nevada more than 20,000 years ago. Cave-dwelling tribesmen left picture writings on rocks in southern Nevada, and Basket Makers and Pueblo Indians also flourished there. Explorers of the early 1800s found Mohave, Paiute, Shoshoni, and Washoe Indians at various locations within Nevada.

Explorers and settlers. Missionaries and fur traders were in the vanguard of exploration. The missionary travels of Francisco Garcés from New Mexico to California in 1775–76 were imitated by other Spanish Franciscans. In 1825 Hudson's Bay Company trappers explored the northern and central region, and two years later Jedediah Smith led a party of Americans into the Las Vegas Valley and across the Great Basin. By 1830 the Old Spanish Trail brought traders from Santa Fe and Los Angeles, and in 1843 and 1845 John C. Frémont's military explorations publicized the Great Basin and the Sierra Nevada region. During the 1840s American pioneers followed the Humboldt Valley-Donner Pass route to the Pacific Coast, and the Forty-Niner Gold Rush greatly expanded migration through Nevada to California.

Nevada, which came under United States sovereignty through the Mexican Cession of 1848, was a part of California until it was incorporated into the newly organized Utah Territory in 1850. In 1849 a settlement was made at Genoa (then Mormon Station) in Carson Valley, but population remained meagre until the discovery of the Comstock Lode in 1859. From that time on Nevada ceased to be merely a highway for gold seekers on the way to California. Thereupon Virginia City became the most famous of all the Western mining camps, and

Discovery of mineral wealth

the rapid influx of prospectors and settlers resulted in the organization of Nevada Territory in 1861.

Simultaneously, the American Civil War gave strategic importance to the new territory. President Lincoln realized that Nevada's mineral wealth could help the Union, and he also needed a Northern state to support the proposed antislavery amendments to the Constitution. Although Nevada Territory had only about one-fifth of the 127,381 persons required for statehood, Nevadans were encouraged to seek admission to the Union. Congress accepted the proposed state constitution and voted for statehood in 1864.

Mining and cattle-ranching decades. Throughout its early decades Nevada's economy was dependent on mining and ranching. The rich Comstock mines reached a maximum annual output of $36,000,000 in silver in 1878. During the 1870s, however, the federal government limited the role of silver in the monetary system, causing a decline in silver prices, the closing of many mines, and the decay of thriving communities into ghost towns.

As mining declined, cattle ranching became a major industry. But beef prices were unpredictable, high railroad rates were burdensome, and severe winters often killed thousands of cattle. In the late 1880s many cattlemen were forced into bankruptcy. Depressed in mining and ranching, the state's population dropped from 62,000 in 1880 to 47,000 in 1890.

Prosperity returned only after the turn of the century, when rich silver ores were discovered near Tonopah, major copper deposits around Ely, and a major gold strike occurred at Goldfield. Thousands of miners answered the lure of those bonanzas, and railroads built extensive branch lines to bring in equipment and haul out the ore. Accessible railroads and reduced rates also encouraged cattlemen to renew large-scale production. Irrigation of fertile river valleys produced sizable hay crops. Thus assured of winter feed, ranchers further expanded herds in the upland regions. World War I demands for Nevada's beef and metals kept the boom going, but the failing markets of the 1920s brought the return of depression.

During its first three decades, Nevada was strongly oriented to Republican control. Reflecting the lax standards in national politics, the state was often manipulated by corrupt politicians. Mineowners and ranchers frequently subsidized government officials, and there were accusations that rich men had bought seats in the United States Senate. Monetary issues became paramount in the 1890s, and the Free Silver Party swept four consecutive state elections. By 1900, however, the traditional two-party system was again in control, and since then Nevada has voted consistently in line with national trends.

Creation of a modern economy. During the Great Depression of the 1930s, when agriculture and mining were in acute distress, Nevada began its transition to a modern economy. After legalization of gambling in 1931, as well as establishment of a six weeks' residency to qualify for divorce, the state developed resort and tourist-oriented businesses. Construction of Hoover Dam on the Colorado River substantially aided the economy of southern Nevada, and its cheap hydroelectric power opened the way for manufacturing. The importation of hydroelectric power from Bonneville Dam on the Columbia River and piped-in natural gas also has brought industrial development in the northwestern region.

Advent of the resort and scientific communities

In the 1950s the establishment of the Nevada Test Site by the federal Atomic Energy Commission expanded employment opportunities and stimulated the development of technical industries within the state. Overshadowing the new industrialization, and fundamentally responsible for the current prosperity, is the diversification and expansion of the tourist trade to include not only the gaming and entertainment facilities of the Reno and Las Vegas areas but also the scenic and recreational opportunities throughout the state.

THE LANDSCAPE

Surface features and waters. Most of Nevada lies within America's Basin and Range Province, or Great Basin, whose topography is characterized by rugged mountains, flat valleys with occasional buttes and mesas, and sandy desert regions. More than 30 north–south mountain ranges cross the state; the highest elevations are Boundary Peak, at 13,140 feet (4,005 metres) and Wheeler Peak, at 13,061 feet (3,981 metres). The southern area of the state is within the Mojave Desert, and the lowest elevation, 470 feet, is in the Colorado River Canyon.

The state's rivers depend for the most part on the melting of winter snows and spring rainfall, and almost all drain into lakes that have no outlets or into shallow sinks that in summer evaporate into alkaline mud flats. The Humboldt, the largest of Nevada's rivers, provides the only major east–west drainage system. The Truckee, Carson, and Walker rivers, which rise in the Sierra Nevadas, serve extensive irrigation and reclamation projects. The Muddy and Virgin rivers in southern Nevada are related to the Colorado River system.

Several lakes provide scenic and recreational attractions. Lake Tahoe on the California–Nevada border is particularly notable for its clarity, depth, and scenic beauty. Pyramid, Walker, and Winnemucca lakes are remnants of an ancient sea. In relation to its area, however, Nevada has comparatively little surface water. The increasing demands of urbanism, industry, and agriculture are exhausting both underground and surface resources, and scarcity of water is an increasing concern. The impounded waters of Lake Mead, extending for 117 miles behind Hoover Dam, provide significant relief for the southeastern area.

Scarcity of water

Climate. The high Sierras along the western boundary often cause clouds of Pacific origin to drop their moisture before reaching Nevada, thus producing a semi-arid climate. The driest regions are in the southeast and near Carson Sink, where annual rainfall seldom exceeds four inches (100 millimetres). The northeast averages eight–nine inches of precipitation, whereas that of the northwestern mountains often reaches 24 inches. Temperatures vary as widely. In the north July temperatures average 70° F (21° C) and in the south 86° F (30° C). In January the averages range from 24° F (−4° C) in the north to 40° F (4° C) in the south. The northern and eastern areas have long, cold winters and short but relatively hot summers, whereas in southern Nevada the summers are long and exceedingly hot, the winters brief and mild. Regional differences are pointed up by variations in the growing season: Las Vegas has 239 days; Elko, only 103; and Reno, 155.

Vegetation and animal life. Despite aridity and rugged terrain, Nevada shows considerable variety in vegetation. In the lower desert areas, mesquite, creosote, greasewood, yucca, and more than 30 varieties of cacti abound, while sagebrush and Joshua trees flourish at the higher elevations. Throughout the state, particularly after the spring rains, more than 2,000 varieties of wild flowers have been identified. Mountain forests contain pine, fir, and spruce as well as juniper and mountain mahogany. The piñon pine is characteristic in the high mountain regions, and the rare bristlecone pine—one of the oldest living species of trees—is native to the Toyaibe Range.

The animal population includes those best adapted to temperature extremes and lack of moisture. Among the larger animals are bighorn sheep, several varieties of deer, and the pronghorn antelope. Rabbit and other rodents are abundant; the desert harbours such reptiles as geckos, horned toads, tortoises, and sidewinder rattlesnakes; and predators such as coyote and bobcat are common. The permanent bird population is limited, but there are seasonal visitations by a great variety of migratory birds. Game birds include sage grouse, pheasants, and quails; and Nevada's rivers and lakes contain large quantities of bass, trout, crappy, and catfish.

THE PEOPLE OF NEVADA

Ethnic and religious groups. The great majority of Nevadans are of European ancestry, about 95 percent of whom were born in the United States. Some 6,000 persons trace their ancestry to Basques recruited as sheep-

herders from their ancient Pyrenean homeland, and about 3,000 Mexican-Americans live in the southeast. Descendants of the Paiute, Shoshone, and Washoe Indians live on several reservations. Blacks, mostly in the Las Vegas and Reno areas, make up less than six percent of the state's populace.

The two predominant religious groups are the Mormons, who migrated westward from neighbouring Utah, and Roman Catholics, each with about one-fifth of the population. Other Protestant denominations and a Jewish minority account for the remainder of the religious adherents.

Contemporary demography. During both the 1950s and 1960s, Nevada's population grew by more than 70 percent, though it remains only 47th in size among the 50 states. In spite of a birthrate slightly above, and a death rate slightly below, the national averages, this growth was largely the result of a steady influx of people from other states. The impact of this immigration has been felt most strongly in Las Vegas and surrounding Clark County, which more than doubled in population in each decade after 1940, and in Reno and surrounding Washoe County. In 1970 about 80 percent of the Nevadans were urban, 73 percent of them engaged in the booming and diversifying economies of those two metropolitan areas.

Immigration and urbanization

THE STATE'S ECONOMY

Although the traditional bases of Nevada's economic life, mining and agriculture, remain important, they were far overshadowed in the 1970s by manufacturing, governmental activity, and especially tourist-related services. The state's labour force in 1970 enrolled nearly one-half of the population, indicating an extraordinarily low percentage of dependent persons, and the per capita income was the fourth highest in the nation.

Mining, agriculture, and forestry. One of the richest mineral regions of the nation extends eastward from California across Nevada and into Arizona. Even with 75 percent of its lands unexplored for its resources, in 1970 Nevada ranked high in mineral production, which provided one-third of the total value of goods produced in Nevada. Copper is the most commercially valuable of the metals, and the state's annual output ranks fifth in the nation. The Ruth Mine in White Pine County is the largest open-pit copper mine in the world. Manganese ore is mined in the southeast, and tungsten, antimony, and iron deposits in central and northwestern Nevada. The Cordero Mine in Humboldt County is the second largest source of mercury in the United States, and total mercury production ranks second only to that of California. Nevada's silver output ranks seventh in the nation, and commercial production of gold occurs only at the Carlin Mine in Eureka County, second in yield nationally. Petroleum was discovered in Nye County in 1954.

Continuity of the traditional economy

Nevada's agriculture is largely dependent upon irrigation. Even in the river valleys, farmers and ranchers pump additional underground water for their crops and livestock. About 600,000 acres are classified as croplands, compared with more than 7,000,000 acres utilized for fenced pastures, meadowlands, and open-range grazing. In recent decades farms and ranches have increased in acreage while declining in number.

Farms are devoted mainly to forage crops or truck farming, though southeastern farms also produce cotton. Livestock ranching, however, is the main source of agricultural income. The large cattle and sheep ranches are chiefly in Elko, Humboldt, and Lander counties, where many ranchers graze their herds for part of the year on leased public lands. Most of the beeves are shipped to California or the Middle West for fattening prior to marketing. Dairy and poultry farms have become important in western and southeastern Nevada, and horse ranches are also being developed in the same regions.

Approximately 12,000,000 acres, or 17 percent of Nevada's total acreage, is devoted to forests and woodlands. More than 5,000,000 acres have been designated as national forests, and private holdings support only a small-scale lumber industry. Aside from lumber production, the forests are of importance for the conservation of water and wildlife and in providing recreational opportunities.

Manufacturing and electrical resources. Throughout the 1960s manufacturing expanded and diversified, and industrial employment rose by nearly 60 percent. Most of the larger enterprises were located in Clark or Washoe counties. In order of value the leading product groups are: primary metals; cement, stone, glass, and ceramic clay; foodstuffs, including meat and dairy products; fabricated metals and machinery; printing and publications; and lumber and wood products. The primary industry is copper smelting, and the largest industrial complex is located in Henderson, where major factories process titanium ore and produce industrial chemicals.

Nevada's electric power capacity for industrial and domestic use doubled between 1955 and 1966. Approximately one-half of the electricity is generated by hydroelectric plants. Although Hoover and Davis dams are major power sources, federal allocations do not adequately serve the energy needs of the state. Steam-generating plants near Las Vegas provide supplementation for southern Nevada, while northwestern Nevada imports hydroelectric power from California.

Federal government. About 87 percent of Nevada's land is owned by the federal government. Following establishment of the Nevada Test Site by the Atomic Energy Commission in the 1950s, a complex of research and development enterprises, mainly in the aerospace, civil-defense research, biomedical environmental protection, and electronics fields, grew up in the Las Vegas area. By the 1970s they had come to rival the similar industries in California and in the Boston and Washington, D.C., areas. The test site itself is a major centre for underground nuclear detonation and for 15 years, until 1972, for nuclear rocket development. Thousands of military personnel are stationed also at Nellis Air Force Base and Fallon Station (Naval Auxiliary Air Station).

The tourist-related industries. In 1970 tourism and its related activities brought Nevada 20,000,000 visitors, contributed 30 percent more income than mining, agriculture, and manufacturing combined, and employed about one-third of the work force. Although millions of persons visited Lake Mead and other recreational and scenic areas, the focus of the tourist industry was on several attractions that are unique to Nevada among the American states.

Legalized gambling and related lures

The 24-hour-a-day gaming casinos bordering the "Strip" in Las Vegas are the most publicized aspect of the legalized-gambling industry. Most such activities are conducted within the law and have contributed to the state's treasury, but there also have been intermittent investigations into the involvement of underworld crime syndicates in Nevada gambling. Important adjuncts to the casinos are the luxury hotels and nightclubs that have made Las Vegas—and, to a lesser extent, Reno and Lake Tahoe—a major centre of live entertainment in the nation.

In addition, Nevada law long has permitted divorces after only brief residence and marriages without waiting periods. Nevada's divorce law attracts thousands of persons annually from across the country, while the ease of marriage is taken advantage of mainly by persons from neighbouring states. The existence of legal prostitution is a unique circumstance in the sparsely populated rural counties of central Nevada.

Labour and unions. The labour organizations of the 1870s and 1880s were related to the mining and railroad industries, but a series of unsuccessful strikes and the radical reputation of the Industrial Workers of the World retarded the movement in the 1900s. Recovery came very slowly, and not until 1939 did more than 30 percent of the Nevada work force hold union membership. Nevertheless, substantial state legislation supported the working population. In 1913 workmen's compensation insurance was initiated, state employment offices were established in 1923, and unemployment compensation was begun in 1937.

In recent decades union organization has had a substantial resurgence. Restaurant workers and the entertainers

have built large, influential unions paralleling those of the teamsters, operating engineers, plumbers, carpenters, and sheet-metal workers. By 1970 labour-union membership had exceeded 50,000, and most Nevada urban workers—excluding the 17,000-plus employees of the gaming industry—were covered by union contracts.

Transportation. Its vast size makes Nevada heavily dependent upon air transportation. The state is served by seven national airlines and numerous supplemental carriers. There are 81 airports and airfields, and both Las Vegas and Reno have been designated as international ports of entry.

Three major railroads cross the state, while short lines serve as feeders where truck competition has not caused discontinuance. Nevada's public roads include primary and secondary highways as well as municipal and rural roads. Two of the nine federal highways are part of the interstate system. Several national bus lines and trucking lines serve the state.

The three major transportation and trade centres are Reno (the principal distributive centre for northwestern Nevada and northeastern California), Elko and Ely (in northeastern Nevada), and Las Vegas (the commercial centre for southern Nevada and nearby areas of Utah and Arizona). Warehousing and trucking industries flourish because of Nevada's strategic geographical location and the "free port" tax exemption for goods continuing in transit.

ADMINISTRATION AND SOCIAL CONDITIONS

State and local government. Nevada is governed under its original constitution, adopted in 1864 but since amended in many respects. The chief officials, including the governor, lieutenant governor, attorney general, secretary of state, controller, and treasurer, are elected to four-year terms. In addition to the usual departments and agencies supervising areas of public concern, the state Equal Rights Commission oversees areas of discrimination of various kinds, while the Gaming Control Board oversees operations of the gambling industry.

Executive, legislative, and judicial branches of government

Nevada's bicameral legislature comprises a Senate of 20 members elected for four-year terms and an Assembly of 40 members elected for two-year terms. It convenes in January of odd-numbered years.

The highest judicial body is the Supreme Court, composed of a chief justice and four associate justices. There are also 12 district courts, subdivided into departments on a population basis. Cities and townships have courts staffed by municipal judges and justices of the peace. All judicial offices are subject to nonpartisan elections.

Local government comprises 17 counties, 16 cities and towns, and 57 townships. Since Nevada traditionally has been rurally oriented, the county remains the primary unit of local administration. Each county has a public administrator, board of commissioners, district attorney, sheriff, and other officials. Cities and towns are incorporated under charters granted by the legislature, most of them with a mayor–council form of government.

Finance. Nevada's fiscal policies have been markedly conservative. The constitution rigidly limits both taxation and indebtedness. The bonding capacity cannot exceed one percent of the total assessed valuation of real property in the state, and the maximum tax rate on real estate is $5 on each $100 of valuation. Even more unusual is the absence of state taxation upon inheritances and all types of income.

Special taxation, however, provides about 70 percent of governmental income, with the balance coming from federal grants and subventions. A general sales tax is the chief source of state revenue, followed by various levies upon gambling and privilege excises. Principal expenditures are for education, health and welfare, and highways.

Political life. Nevada's politics has been middle-of-the-road. Through 1970 the Republicans and Democrats each have carried the state 13 times in presidential elections, and the state's electoral vote has been cast for the victor in every presidential campaign since 1912. State elections have shown a similar pattern. Of the 20 governors elected since statehood, nine were Democrats, three were Free Silverites, and nine were Republicans. The legislature usually is divided, with Democrats frequently in control of the Assembly, and Republicans customarily having a majority in the Senate. In the 1960s legislative overrepresentation of the rural counties offset the numerical advantage of the Democrats, but the redistricting that became effective in 1972 opened the way for a major shift in political control.

Education. The public school system is controlled by an elected eight-member Board of Education, which delegates administrative responsibilities to an appointive superintendent of public instruction. Local school districts, coextensive with the counties, receive supplementary funding from the state. School attendance is compulsory for those between the ages of seven and 17; in 1970–71 more than one-fourth of the state's population was enrolled in public schools.

Institutions of higher education

The University of Nevada originally was established at Elko in 1874 under the provisions of the Morrill Land-Grant College Act; 12 years later it was moved to Reno. In 1951 an extension branch was established in Las Vegas, which since has become the autonomous University of Nevada, Las Vegas. Elko Community College is within the university system, and the 1971 legislature established similar two-year colleges for the western and southern regions. To supplement campus instruction, the Desert Research Institute and the Agricultural Experiment Stations provide statewide research services.

Welfare, health, and housing. Nevada's social welfare programs and its custodial institutions are administered by the Department of Health, Welfare, and Rehabilitation. Old-age and welfare allotments are given to those qualified on the basis of need, and public assistance is available for the blind and other handicapped individuals. Support payments are provided for dependent children, and a state orphanage is located near Carson City. The state mental hospital is in Sparks, and in 1970 a mental-health facility was established in Las Vegas. Penal and rehabilitation institutions include the state prison in Carson City, a girls' training centre at Caliente, and an industrial school for boys at Elko.

Health care, housing, and public safety are responsibilities of local government or private enterprise. In 1971 Nevada had 20 general hospitals, which provided slightly more than 2,000 beds for patients, and there were 109 physicians and 47 dentists per 1,000 of population. An open-housing law was passed by the 1971 legislature, and limited beginnings were made in low-cost public housing.

CULTURAL LIFE AND INSTITUTIONS

Arts and entertainments. Nevadans traditionally have mingled rural conservatism and the individualism of the Old West. During its first century the population was relatively small and dispersed, and cultural values remained those of an agrarian society. With the sophisticated resort industries and the influx of residents, however, Las Vegas and Reno developed marked metropolitan characteristics. Not only has the economy diversified but cultural activities have burgeoned. Recognizing this trend, the 1967 legislature established the Nevada State Council on the Arts to coordinate and stimulate cultural activities.

Both major cities have well-established programs in the performing arts. The universities sponsor lectures, concerts, and theatrical productions, while the tourist industry regularly features the most famous entertainers from America's show business. Both Reno and Las Vegas support symphony orchestras and have commercial and public art galleries. Varied handicrafts are promoted by the Artists and Craftsmen's Guild of Nevada and by the Handweavers and Spinners Guild. Traditional Indian arts and crafts have been revived on reservations and in urban colonies.

Preservation of the past. Nevada's frontier heritage is commemorated by annual pageants and festivals, which provide natives with both sentimental nostalgia and additional commercial opportunities. Perhaps the best known is Helldorado Week in Las Vegas each May, for which

townspeople wear Western garb and mobilize a series of rodeos and parades. A unique ethnic event is the Basque Festival in Elko, and the Reno Rodeo is an outstanding Fourth of July celebration. The entire state observes its anniversary, Admission Day, on October 31, highlighted by a picturesque parade and costume ball in Carson City.

Old West festivals

Nevada has five museums and 43 libraries. The Nevada State Museum, located in the old federal mint building in Carson City, emphasizes the mining industry and mineral collections. Anthropological artifacts are featured at the Lost City Museum in Overton, at the University Museum in Las Vegas, and at the Southern Nevada Museum in Henderson. The Mackay School of Mines Museum, on the Reno campus of the state university, is oriented to metallurgical, mineralogical, and geological specimens. The Nevada Historical Society, also in Reno, has pioneer mementos, the most complete holding of Nevada newspapers, and a sizable historical reference library. The library of the University of Nevada, Reno, has the largest collection of books in the state, while the Nevada State Library at Carson City is notable for excellence in legal fields.

Scenic and historic sites. State and federal governments maintain six historic monuments, three major parks, and several developed recreational areas. The Valley of Fire State Park, near Overton, is widely known for its brilliantly coloured rock formations, Indian pictographs, and petrified forest. Mormon Station Historic State Monument, in Genoa, includes the site of the first permanent settlement in Nevada, and Cathedral Gorge, near Caliente, displays red and gold rocks that resemble church spires. The famed Death Valley area lies on the border between Nevada and California, and Lehman Caves National Monument, near Baker, has underground pools and unique limestone formations.

Communications media. Of the 31 newspapers published in Nevada, 23 appear as weekly journals. The largest dailies are the *Las Vegas Review Journal,* the *Las Vegas Sun,* Reno's *Nevada State Journal,* and the *Reno Evening Gazette.* Nine television channels and 32 commercial radio stations blanket the state.

Prospects. Nevada's future appears to be related closely to an ever accelerated growth and diversification. A continuing influx of population and industrial expansion will have a special impact on the southern region. Consequently, Las Vegas and Clark County will increasingly dominate the politics and economy of the state. Problems that already loom conspicuously on the horizon include: conservation and augmentation of water resources, control of environmental pollution, elimination of disadvantages for ethnic minorities, a broader base for state and local taxation, and governmental consolidation for the metropolitan areas. It is likely that, during the 1970s, Nevada will move farther from its rural past and continue to develop in the patterns of urban America.

BIBLIOGRAPHY. The FEDERAL WRITERS' PROJECT, *Nevada: A Guide to the Silver State* (1940), provides a thorough but dated overview of the state. For a comprehensive bibliography, see RUSSELL R. ELLIOTT and H.J. POULTON, *Writings on Nevada* (1963). The best historical survey is JAMES W. HULSE, *The Nevada Adventure,* rev. ed. (1969); but still useful are JAMES G. SCRUGHAM (ed.), *Nevada: A Narrative of the Conquest of a Frontier Land,* 3 vol. (1935); and RICHARD G. LILLARD, *Desert Challenge* (1942). For special topics, see ELEANORE BUSHNELL, *The Nevada Constitution,* rev. ed. (1968); JOHN A. BRENNAN, *Silver and the First New Deal* (1969); RUSSELL R. ELLIOTT, *Nevada's Twentieth-Century Mining Boom* (1966); MARY ELLEN GLASS, *Silver and Politics in Nevada* (1970); JOHN KOONTZ, *Political History of Nevada,* 5th ed. (1965); ELIOT LORD, *Comstock Mining and Miners* (1883); STANLEY W. PAHER, *Nevada Ghost Towns and Mining Camps* (1970); and WILLIAM WRIGHT, *The Big Bonanza* (1947).

(R.J.Z.)

New Brunswick

New Brunswick is a Canadian province located on the eastern seaboard of the North American continent. One of the four original provinces making up the national Confederation in 1867, it ranked eighth of the ten Canadian provinces in population a century later. Together with Nova Scotia and Prince Edward Island, it forms the regional grouping known as the Atlantic (Maritime) Provinces. Though it has contributed heavily to the national life of Canada, its importance has receded as the country has expanded toward the west and north. Economically, it has failed to keep pace with the more central provinces of Ontario and Quebec. On the other hand, its beautiful natural environment—an attraction for tourists, hunters, and fishermen—has remained relatively unspoiled. It has an oblong shape, about 200 miles from north to south and 150 miles from east to west. Its area is 28,354 square miles (73,437 square kilometres), and it extends from 45° to 48° north latitude and from 63° to 69° west longitude. To its west, across the United States border, lies the state of Maine; to the north it is bounded by Quebec, on the east by the Gulf of St. Lawrence and Northumberland Strait, and on the south by the Bay of Fundy. Separating the latter features, the narrow neck of the Isthmus of Chignecto joins the province to Nova Scotia. (For related information see CANADA; CANADA, HISTORY OF.)

HISTORY

The province was first known to Europeans as a portion of a seaward region called Acadia, inhabited by Micmac and Malecite Indians of the Algonkian-speaking group. French traders and settlers made but slight impact on this mainland part of Acadia until, in the mid-18th century, British expansion on the neighbouring Nova Scotian peninsula compelled the French authorities to make a precautionary military move. A French fort at Beauséjour, on the 17-mile-wide Chignecto Isthmus, became the point of the principal confrontation and fell to British assault in 1755. The expulsion of the French-speaking Acadian people followed immediately, and for several years the British hunted down parties of refugees in the inner recesses of New Brunswick's forests and rivers.

By a British proclamation of 1763, the area later to be known as New Brunswick was incorporated into Nova Scotia. Development was slow. A sprinkling of New England settlers continued to move to the St. John River, where the first English settlement had taken place at Maugerville in 1762. A few hundred Acadians returned, establishing themselves on the northern and eastern coastlines. The American Revolution had important consequences for the area when some 14,000 Loyalists, mostly from the middle states of the new republic, settled on the banks of the St. John and St. Croix rivers. This decisive event was followed by the separation of the area from Nova Scotia and the establishment of the province of New Brunswick in 1784.

Early settlement

Early Loyalist enthusiasm was soon diminished. In the last years of the 18th century, the war of the French Revolution prevented the British government from giving New Brunswick the kind of encouragement its people expected, and not until about 1808 did the terms of trade alter in favour of the province and the population start to grow. This was primarily a result of the decision of the British government to grant tariff preferences to the timber resources of the North American colonies, a situation brought about by Napoleon's control of Baltic products. The Loyalist expectation of a society of landed gentlemen, living from the fruits of agriculture, gave way to the realization that New Brunswick possessed little arable land and the subsequent development of an economy based on the forests.

Material progress continued rapidly until the mid-19th century, reaching a climax during the so-called age of wood, wind, and water, when the province gained a reputation for the quality of its ships and sailors. Responsible home government was granted by Britain in 1848. The Reciprocity Treaty with the United States (1854) and opportunities offered by the U.S. Civil War stimulated trade and shipping. New Brunswick entered into confederation with Canada in 1867 only with reluctance, largely because of its general satisfaction with its status as an almost completely self-governing colony within the British Empire.

Union with Canada

New Brunswick has had an unspectacular history in its existence as a province of Canada. Forest products have continued to be the basis of its economy, and secondary industry has failed to develop to significant proportions. In comparison with national growth, wealth and population have progressed but slowly. Until the restrictions by American immigration authorities in the 1920s, large numbers of young people left for New England, where economic prospects appeared brighter. Since the 1940s there has been a comparable exodus to central Canada. By the 1970s, average income in the province was about two-thirds that of Canada as a whole.

THE LANDSCAPE

Forests cover some 87 percent of the province, a proportion that has been increasing since 1950. Pine, once profuse, has steadily diminished, and the principal growths today are maple, birch, spruce, fir, and tamarack (American larch). Much reforestation has taken place. The climate, with temperatures ranging from −30° F (−34° C) in winter to 90° F (32° C) in summer, and with 40 to 45 inches (1,000 to 1,125 millimetres) of rainfall and up to 100 inches of snow annually, nourishes rapid forest growth. New Brunswick is also a land of lakes, rivers, and small streams. The St. John and St. Croix rivers drain the western half of the province, the east being served by the Restigouche, Miramichi, and Petitcodiac. Ranges of hills, reaching a maximum altitude in Mt. Carleton, 2,690 feet (820 metres), cross the province from west to east in both the north and south.

Most of the soil of New Brunswick is acid, stony, and poorly drained, although there are occasional pockets of good cattle-raising country, and the upper St. John Valley is noted for the production of potatoes. Villages are market centres or sites for sawmills. The population has become increasingly urbanized, the proportion exceeding 50 percent by the 1970s. Of the rural dwellers, less than 10 percent are engaged in farming. The largest city, Saint John, with a metropolitan area population exceeding 130,000, is a manufacturing centre, national port, and terminus of Canada's two great railways. Moncton, with a population of nearly 50,000 by the early 1970s, is a manufacturing and communications centre. Fredericton (25,000), the provincial capital since 1785, is the commercial centre of central New Brunswick.

Patterns of settlement

The dense forests of the province shelter an increasing moose population. During the autumnal open season, thousands of deer are pursued by brightly-coated hunters, many of them from the sprawling urban areas elsewhere on the continental seaboard. Rivers and streams abound in fish, but by the 1970s the famous Atlantic salmon had greatly diminished in number, the victim of pollution from industrial plants and from the aerial spray directed against the spruce budworm that infested the forests.

THE PEOPLE

New Brunswick is notable for its large French-speaking minority, principally descendants of those returning Acadians of the 18th century who settled on the northern and eastern shores of the province. In the northwest these have been substantially reinforced by French Canadians from Quebec. Against strong historical odds, the Acadians have not only preserved their identity in an English-speaking environment but have, over the years since confederation, increased from 16 to nearly 40 percent of the population. Northern and eastern counties are predominantly French-speaking. Settlement has followed the coastline and rivers, the northern interior being almost barren of population.

Origins

The first English-speaking groups moved in during the 18th century, from the New England region farther south. Between 1810 and 1850 the population grew rapidly, from 30,000 to 193,800. This was a result of the introduction of English, Scots, and Irish groups, who came with the intention of farming but were for the most part absorbed into forest industries. Numerically most important and most identifiable, especially in Saint John and on the Miramichi, were the Irish. Following confederation small colonies of Scots (and one of Danes) were established on the upper St. John. Since 1945 the population has become slightly more cosmopolitan, the most notable addition being a contingent of Dutch. Indians, scattered through the province on reservations, numbered less than 4,000 by 1970. A black minority, from colonial times, amounted to 5,500, some of them descendants of slaves of the Loyalists.

The admittedly modest natural resources of the province and the heavy emigration attendant upon the greater opportunities available elsewhere have been responsible for a slowing of natural population growth, a factor given fresh emphasis in recent years. At the 1971 census the total population was 635,000, an increase of about 18,000 over the census figures of 1966. This was in spite of a comparatively high birth rate of 19.9 per 1,000 and a death rate of 7.9. As a result of its high birth rate, and unlike the other Atlantic Provinces, which had a high proportion of elderly people, New Brunswick was close to the national average.

New Brunswick, by the early 1970s, had nothing to approach the national minority problems afflicting the bilingual communities of Quebec and the racial strife of the metropolitan areas of the nearby United States. Historically, relations between the two main national groups within the province have been amicable, though in recent years, under the stimulus of the Assumption Society and other cultural organizations, Acadians have become more conscious of an underprivileged status in the operation of the economy, in government, and in education. Inequalities have, to a great extent, been removed by a government that, for ten years after 1960, had an Acadian premier and a majority of Acadian ministers. The University of Moncton (1963; originally founded as St. Joseph's University in 1864), has become an important Acadian cultural centre, and French has been made an official second language of the province, with minimal protest from the English-speaking majority. The economic and cultural plight of the Indian and black minorities began to generate some attention during this period.

Minority consciousness

Religiously, the province is roughly 51 percent Roman Catholic. The largest Protestant denominations are the Baptist (18 percent), which gained many adherents in the early 19th century, the Anglican, and the United Church of Canada (each about 14 percent).

THE ECONOMY

Two-thirds of provincial production is based on forest industries and the manufacture of pulp and paper, a proportion tending to grow in recent years. Seven large paper mills and a multitude of sawn-lumber establishments account for the principal exports. Many farmers contribute to the industry by maintaining woodlots, pulpwood frequently being an important cash crop. On the Bay of Fundy and along the northeastern coast, fishing is a major industry, with lobsters being the most valuable product. By 1970, total landings exceeded $16,000,000 annually. The most spectacular development of the 1960s and early 1970s has been the advance of mining for base metals in the northeast, value being almost $100,000,000. Production of electric power from hydro- and steam-generating sources, greatly encouraged by government in hope of attracting secondary industry, amounted by 1970 to nearly 4,000,000 kilowatt-hours. Owing to the seasonal factors on which most industries are based, the rate of unemployment, at times reaching 10.8 percent, is considerably above the national average.

The long-recognized fact that New Brunswick is an underprivileged region has impelled governmental activity to raise average income closer to the national standard. The traditional method has been to alleviate isolation of the province from large markets by favourable adjustment in freight rates. Results, however, have been insignificant, and more recently the federal government has embarked on heavy spending programs, a notable example of which is the June 1970 agreement of Canada's Department of Regional Development to spend $62,000,000 over a two-year period on what is called infrastructure—improvement of roads and construction of new facilities for communication—all in the hope of

Development measures

sparking industrial development. Saint John and Moncton have been designated growth areas to receive special attention. This planning operation has been supplemented by work of The New Brunswick Development Corporation, a provincial agency whose principal ambition is to build a multi-industry complex at Saint John, based on the possibility of a greatly increasing share of North America's seagoing traffic. The port's deepwater facilities offer hope that new-type containerized cargo vessels will enable it to expand its trade, perhaps at the expense of New York and Montreal.

TRANSPORTATION

The two great national railway systems, Canadian National and Canadian Pacific, have slowly yielded their primacy to transportation by road and air. The port of Saint John, however, still depends on rail traffic for its main sustenance, principally by means of the Canadian Pacific line running through Maine to Montreal. Passenger service has steadily diminished as a result of the development of airports at Saint John, Moncton, and Fredericton, supplemented by half a dozen smaller centres in the north. An autobus service accounts for most internal passenger traffic. As a major port, Saint John accounts for a high proportion of Canadian external trade, especially in the winter season when Montreal is closed off by ice in the St. Lawrence. Small freighters transport produce of the northern lumber and mining industries to Europe and elsewhere.

From the 1950s on, the road system has been modernized. In recent decades emphasis has been given to improvement of secondary roads and to improved snow clearance in the winter season.

ADMINISTRATION AND SOCIAL CONDITIONS

Government follows the constitutional forms of a Canadian province within the framework of the British North America Act. The one-chamber legislature at Fredericton, consisting of 58 members and representing 15 counties and six cities, has control over education and other internal services but shares jurisdiction with Ottawa over several matters. The executive consists of the premier and 13 cabinet members. Since 1945 the area of provincial jurisdiction has tended to enlarge at the expense of the federal. Under the constitution elections are to be held every five years but in practice take place every four years. Major political parties are the Liberal and Conservative, third parties never having succeeded in making an important impact. Liberals held power from 1960 to 1970, a Conservative government taking office in November of the latter year.

Control over many social services—family allowances, old-age pensions, unemployment insurance—is in whole or part shared with the federal government. Approximately 50 percent of provincial revenues are derived from Ottawa by subsidy, under the British North America Act and subsequent federal-provincial agreement. For the development of its program of social welfare, the provincial government is heavily dependent on Ottawa for financial assistance and has been compelled to raise rates of taxation to levels higher than those in most Canadian provinces.

By its Equal Opportunity Programme (1967), the government sharply curtailed powers of municipal administration, taking exclusive control over assessment of properties and collection of taxes, the object being to equalize living standards throughout the province. Six cities, 21 towns, and over 90 incorporated villages are dependent for revenues on provincial grants. School boards, administration of justice at the local level, and welfare assistance are all under the control of provincial officers to a much greater extent than formerly applied. For the 1970s, centralization has been the formula in vogue.

Great disparities in standards of living between wealthier counties of the St. John Valley and those of the predominantly Acadian northern and eastern shores have been a target for governmental action. In spite of heavy expenditure, plans to move poverty-stricken residents of Gloucester County, an area of chronic unemployment, to more favoured locations have produced negligible results. Yet the opening of new trade and technical schools in all sections gives hope for the future. Marginal farms, once prosperous, have become deserted, and more and more people are becoming dependent on secondary industry and its ancillaries; yet their material condition has not greatly improved. A general complaint has been that thousands of young people, educated by the government at heavy cost, have left the province. To meet inflationary conditions, the Minimum Wage Act of 1965 has steadily been modified in an upward direction, and the cost of living is almost that of the national average. A socialized medical system came into operation on January 1, 1971. Housing construction is dependent on interest rates set by federal authorities at Ottawa. Perhaps because of their close contact with a beautiful natural environment, a general awareness of depressed living conditions has not, as yet, become coupled with bitter and widespread social discontent.

Depressed areas

The Canadian Armed Forces maintain Camp Gagetown, a training base and staging area for army personnel, in the lower St. John valley. Containing, as it does, some 5,000–10,000 troops and their dependents, it is an important ingredient in the provincial economy. At Chatham there is a base for fighter craft of the air force.

CULTURAL LIFE AND INSTITUTIONS

For a small province, New Brunswick enjoys a remarkable and varied cultural life. At Fredericton, home of the first native Canadian school of poetry, that of William Bliss Carman (1861–1929), a celebrant of nature, and Sir Charles Roberts (1860–1943), a regionalist influenced by Ralph Waldo Emerson, the tradition is strongly sustained. The Lord Beaverbrook Art Gallery at Fredericton—a legacy of the newspaper magnate, who spent his formative years in the province—contains many old masters, housing treasures of English and Canadian schools of painting, including the famous Krieghoff Collection. Many artists of national reputation, attracted by the magnificent setting, reside in the province. Theatre New Brunswick, endowed by the Beaverbrook Foundation and directed from The Playhouse at Fredericton, carries on year-round activity. There are many regional festivals of music, the most distinguished being those of the Acadians, who have produced choral groups that have been highly successful in international competition. Sculpture in wood and work in the ceramic arts are vigorously pursued. The province maintains an agency to foster cultural activity.

Education

Four universities—those of New Brunswick at Fredericton and Saint John (established 1785 and the oldest in Canada); Moncton, which is bilingual; Mount Allison, at Sackville; and St. Thomas, Fredericton—all support creative arts. New Brunswick and Moncton are under public control. Mount Allison is a private corporation, while St. Thomas, working in association with New Brunswick, is Roman Catholic. All are financially supported by the provincial government. The New Brunswick Museum, Saint John, oldest in Canada, contains many treasures reminiscent of pioneer days. The Provincial Archives are located at Fredericton. The historic past is commemorated by the construction of King's Landing, 20 miles north of Fredericton, a replica of a pioneer village; while at Ft. Beauséjour a massive work of archaeological investigation and a national historic park remind visitors of the 18th century conflict between French and British.

By the early 1970s there were five daily English-language newspapers and one French-language newspaper. Their services are supplemented by eight weeklies. The Canadian Broadcasting Corporation has television stations at Saint John and Moncton, the latter providing service in French. An independent station operates at Moncton. Ample radio broadcasting is provided from a dozen stations.

PROBLEMS AND PROSPECTS

Since 1867, material progress in New Brunswick has been disappointing, but largely in relation to development elsewhere in North America. Confederation failed to provide

large markets for provincial produce in the interior of the continent. The larger and more centrally located secondary industry of Ontario and Quebec gave cruel competition to provincial energies, sometimes bringing them close to economic extinction. The decline of wooden shipbuilding brought an end to the major role New Brunswick had played in overseas markets, and the rise of the pulp and paper industry in the 1920s failed to compensate for lost advantages. Since 1937 the federal government has accepted the principle that wealthier provinces should assist the less fortunate ones. Consequent federal action up to the decade of the 1970s has enabled the province to keep from losing further ground, although hopes for a substantial reversal of fortunes spring eternal in the local consciousness. In the early 1970s, these were based on brighter prospects for the future of those Atlantic ports enjoying deepwater facilities. The recent increase of income from the tourist trade also brought the belief that this, too, might become a principal provincial asset.

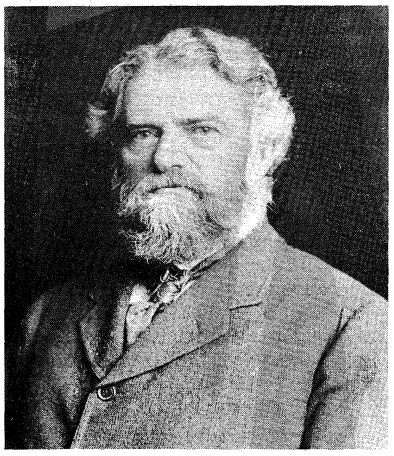
BIBLIOGRAPHY. For the discovery of the site of what later became New Brunswick, see HENRY HARRISSE, *Notes pour servir à l'histoire . . . de la Nouvelle-France et des pays adjacents, 1545–1700* (1872); and MORRIS BISHOP, *Champlain: The Life of Fortitude* (1948). For its early history as a colony of France and then of Great Britain, see BONA ARSENAULT, *Histoire et généalogie des Acadiens*, 2 vol. (1965; rev. Eng. trans. of the historical part, *History of the Acadians*, 1966); and G.G. CAMPBELL, *The History of Nova Scotia* (1948). For confederation, see P.B. WAITE, *The Life and Times of Confederation, 1864–67* (1962). For a comprehensive history, see JAMES HANNAY, *The History of New Brunswick*, 2 vol. (1909); and W.S. MACNUTT, *New Brunswick, a History: 1784–1867* (1963). For the economy, see *The New Brunswick Economy* (1955), which was presented to the Royal Commission on Canada's Economic Prospects. See also *The New Brunswick Almanac* (annual).

(W.S.MacNu.)

Newcomb, Simon

Simon Newcomb was perhaps the greatest American astronomer of the 19th century, although in depth and originality of mathematical invention he was excelled by George W. Hill, his associate for 15 years. Newcomb was born in Wallace, Nova Scotia, on March 12, 1835. His father, an itinerant country schoolteacher, taught Simon to count at the age of four, and before he was five he was spending several hours a day making calculations in addition and multiplication; before he was seven he had finished the arithmetic book, including the extraction of cube roots.

By courtesy of the University of Chicago; photograph, Yerkes Observatory, Williams Bay, Wis.



Newcomb.

Newcomb had little or no formal education. At the age of 16 he was apprenticed to a quack herb doctor in Salisbury, New Brunswick. After two or three years he ran away to join his widowed father who had settled in the United States, in Maryland. In the libraries at Washington, D.C., Simon found the first full opportunity to indulge his intellectual curiosity. After avidly exploring many technical fields he concluded that his principal talent lay in mathematics. He was especially attracted to the *American Ephemeris and Nautical Almanac*, an annual handbook for astronomers, containing predicted positions of the principal celestial objects and other astronomical phenomena. Of this work he says, "Its preparation seemed to me to embody the highest intellectual power to which man had ever attained." He thereupon applied for employment in the American Nautical Almanac Office, then at Cambridge, Massachusetts, and became a computer there in 1857. He also enrolled in the Lawrence Scientific School of Harvard University, receiving a degree in 1858. In 1861 he applied for and received a commission in the corps of professors of mathematics in the United States Navy and was assigned to the United States Naval Observatory at Washington, where he worked for more than ten years determining the positions of celestial objects with the meridian instruments and for two years with a then new 26-inch refractor telescope.

Major work: solar-system motions

In 1877 Newcomb was put in charge of the American Nautical Almanac Office, then in Washington, where almost at once he commenced the great work that he had had in his mind for some years, and that was to occupy the greater part of the rest of his life: the calculation of the motions of the bodies in the solar system. Reaching the compulsory retirement age for captains in 1897, he later received the then unusual distinction of retirement with the rank of rear admiral. He died in Washington on July 11, 1909, and was buried in Arlington National Cemetery.

In 1884 Newcomb had obtained the additional appointment of professor of mathematics and astronomy, which he held until 1893, at Johns Hopkins University, Baltimore, Maryland, continuing, however, to live in Washington. For many years he was editor of the *American Journal of Mathematics*. He was one of the founders of the American Astronomical Society and was its first president (1899–1905). Newcomb received honorary degrees from ten European and seven American universities and was a member of 45 foreign societies. He was awarded the gold medal of the Royal Astronomical Society (1874), the Huygens Gold Medal of the Holland Society of Science (1878), the Copley Medal of the Royal Society (1890), the Bruce Gold Medal of the Astronomical Society of the Pacific (1897), the Schubert Prize of the Imperial Academy of Sciences, St. Petersburg (1897), and the Sylvester Prize of the Johns Hopkins University (1901). He was elected a member of the National Academy of Sciences in 1869, serving as home secretary, 1881–83; vice president, 1883–89; and foreign secretary, 1903 until his death.

Newcomb's most important work appeared in the *Astronomical Papers Prepared for the Use of the American Ephemeris and Nautical Almanac*, a series of memoirs that he founded in 1879 with the object of giving "a systematic determination of the constants of astronomy from the best existing data, a reinvestigation of the theories of the celestial motions, and the preparation of tables, formulae, and precepts for the construction of ephemerides, and for other applications of the same results." Of 36 articles filling approximately 4,500 quarto pages in the first nine volumes, he was sole or principal author of 25. Among them were his tables of the Sun, Mercury, Venus, Mars, Uranus, and Neptune, and Hill's tables of Jupiter and Saturn, which were in use throughout most of the world for calculating daily positions of the objects from 1901 to 1959, and even afterward for the Sun, Mercury, Venus, and Mars. This series of *Papers* is remarkable for its sustained high quality. Hardly anything in them has proved to be incorrect, and at mid-20th century they were still worthy of the attention of any student of celestial motions.

Development of system of astronomical constants

Possibly Newcomb's most far-reaching contribution was his inauguration, jointly with A.M.W. Downing, then superintendent of the British Nautical Almanac Office, of a worldwide unified system of astronomical constants, which was later to lead to the outstandingly successful scheme of international collaboration among the princi-

pal almanac makers of the world that survived two World Wars with increasing vigour. Newcomb and Downing were impressed by the "confusion which pervaded the whole system of exact astronomy, arising from the diversity of the fundamental data made use of by the astronomers of foreign countries and various institutions in their work" (Newcomb, *Reminiscences of an Astronomer*, 1903). A conference of the directors of the national ephemerides of the United States, Great Britain, France, and Germany, was held in Paris in May 1896. It resolved that beginning with 1901 a certain set of constants, substantially Newcomb's, should be used by all the ephemerides. The decision even included some work of Newcomb's that was not to be finished for several years. Although Newcomb was attacked by some astronomers at home for decisions that were in their opinion premature, time has proved the wisdom of his course. A similar conference, held at Paris in 1950, decided unanimously that the system of constants adopted in 1896 was still preferable to any other for practical use.

BIBLIOGRAPHY. NEWCOMB wrote a number of books, including *Popular Astronomy* (1878), *The Stars* (1901), *Astronomy for Everybody* (1902), and *Compendium of Spherical Astronomy* (1906), some of which have been translated into as many as seven foreign languages. He also wrote on finance and economics and published some fiction. An autobiography, *Reminiscences of an Astronomer*, was published in 1903. A bibliography of his life and works (541 titles) is given by R.C. ARCHIBALD, *Mem. Natn. Acad. Sci.*, 17:19–69 (1924).

(G.M.Ce.)

Newfoundland

The most easterly portion of the North American continent, Newfoundland is a Canadian province thrusting into the bleak North Atlantic Ocean far from the nation's industrial heartland. Its position gives it strategic importance in transportation and communications between North America and Europe. The Strait of Belle Isle separates the island of Newfoundland to the south from mainland Labrador, which comprises nearly three-fourths of the province's total area. The wildly jagged and seasmashed coasts that, with all indentations, stretch for more than 10,000 miles (16,000 kilometres) form an apt symbol of the staunch independence of its people, who refused to become Canadians until 1949. Above the continental shelf fringing the coasts, but especially about the Grand Bank to the south of the island, are some of the world's finest and most extensive fishing grounds. Their existence has been the single most crucial factor in shaping the history and the character of the province from its founding to the present. (For information on related topics, see CANADA; CANADA, HISTORY OF; and NORTH AMERICA.)

HISTORY

The official discoverer of Newfoundland was the Genoese-Venetian navigator Giovanni Caboto (*c.* 1450-*c.* 1498/1499), better known to history as John Cabot, who reached the island in 1497 while sailing under the English flag. The Grand Bank doubtless had been known to fishermen earlier, but Cabot's enthusiastic reports opened the way for international rivalries over the region. Early in the 16th century, English, French, Basque, and Portuguese fishermen were contesting for catches; by 1600 England and France were the chief rivals for Newfoundland.

Early restrictions on settlement. Several attempts made at colonization during the 17th century met the inveterate hostility of the English fishermen, and, after 1634, of the English crown. In 1699 Parliament prohibited settlement of the island except as necessary to maintain the cod fishery. This remained the basic governing directive for Newfoundland for more than 150 years, during which time the island became virtually a private fiefdom ruled by commercial fishing interests.

The rationale for this policy was twofold. First, under the prevailing concepts of international trade, known as the mercantile theory, the fishing of the Grand Bank remained a mother-country industry rather than that of a colony, with definite economic advantages accruing to England. Second, the region was looked upon as a "nursery for seamen," a training ground for the crews that were to maintain England's maritime superiority into the distant future. This latter idea also led the French in 1662 to found Placentia on the southeast coast as the seat of a royal governor and a base of operations. From Placentia the French destroyed every English settlement on the island in actions which took place in 1696, 1705, and 1708. Treaties in 1713 and 1783, the latter marking the armistice ending the American Revolution, recognized British sovereignty but granted French fishermen the right to land and dry their catches along portions of the northern and western coasts. This "French shore problem" remained a thorny one for the islanders until 1904.

French raids

During the 17th and 18th centuries the island's population increased despite repressive legislation, the denial of property rights and social services, and a particularly vicious judicial system managed by the captains of fishing vessels. Britain found it necessary to appoint a naval governor in 1729 and to establish a proper judicial system in 1792. Both governor and court remained only for the summer fishing season, however, and not until 1824, with the appointment of a resident governor and council, was a settled colony finally acknowledged.

In 1832 local agitation led to the addition of a popularly elected assembly. In 1855 the island was granted full responsible government after a campaign that reignited a bitter underlying sectarianism. As the hold of the transient fishermen over the island's life weakened toward the end of the 18th century, many of the prosperous fishermen settled in as merchants, and government officeholders increased in number. Large groups of labourers were imported for the cod fishery, 11,000 coming in the peak year of 1814–15 alone. These immigrants, mainly Irish Catholics, continued to stand for the rights of fishermen and servants against the new and largely Protestant middle class, creating a sectarian spirit that persisted in Newfoundland's political life well into the 20th century.

Sectarian rivalries

Toward confederation. Between 1864 and 1869, Newfoundland was involved in discussions leading to Canadian confederation. The idea of actually joining the confederation, however, was rejected at the polls both then and after negotiations were reopened in 1895. New radical social policies were heralded with the formation of the People's Party by Edward Morris in 1907 and the Fishermen's Protective Union by William Coaker in 1908. The political rivalry of Morris and Coaker was submerged in the common efforts of World War I.

For nearly 150 years, rights to the coasts of Labrador, an ill-defined geographical entity, were hotly debated. By 1825, British tribunals had given Newfoundland jurisdiction over the Atlantic face, Canada over the coasts on the Gulf of St. Lawrence. But the inland boundaries and administration of Labrador were to remain unsettled until 1927, when Newfoundland was awarded the huge inland region in a definition that said that the "coasts" reached to the headwaters of rivers emptying eastward into the Atlantic.

The depression that afflicted the Western world after 1929 completely overwhelmed Newfoundland's already struggling economy, which had been overtaxed by the war and by the cost of building a transisland railway. In 1933 a royal commission recommended the suspension of the existing government and the inauguration of government by a commission appointed by and responsible to the government of the U.K. In 1934, Newfoundland came under the rule of such a government. The commission was finally able to re-establish the colony's finances on a firmer foundation.

Following World War II a national convention was elected to recommend possible forms of future government to be voted upon by the people. The convention failed, however, to recommend anything beyond a continuation of the existing government or a return to the pre-1934 constitution. The British government, nevertheless, added the confederation question to the ballot. The first of two referenda held in 1948 was inconclusive; the second gave a clear majority to confederation. On March 31, 1949, Newfoundland entered the Canadian Union.

Confederation contested

THE PHYSICAL ENVIRONMENT

Landscapes. Newfoundland and Labrador must be treated as two separate physiographic regions. The island is roughly triangular in shape with two large peninsulas hooking into the sea from the northwest and southeast corners. It is part of the Appalachian geological province of North America, in which the landforms run from southwest to northeast. The land in the west of the island rises abruptly from the sea or from a narrow coastal plain to a maximum height of 2,651 feet (808 metres) to form the edge of a plateau that slopes away to the northeast to elevations that do not generally rise above a few hundred feet. The plateau region is gently undulating and dotted with thousands of lakes and ponds; the coastal terrain is hilly and rugged. The submersion of coastal areas has created numerous bays and offshore islands.

Labrador lies north and east of the province of Quebec on the Labrador-Ungava Peninsula. Geologically, it is part of the edge of the Canadian Shield, the rugged mass comprising some of the world's oldest rock. In the Torngat Mountains in the extreme north, the land rises abruptly from the sea to heights of over 5,000 feet. Deeply indented and dissected by many rivers, first glaciated then flooded, the coastline presents a rugged, forbidding front to the Atlantic. Inland, the plateau is undulating and lake spotted.

Climate. In general, Newfoundland has cold but not severe winters and warm to cool summers. The mean July temperature of the coast ranges from 40° F (4.4° C) in northern Labrador to 59° F (15° C) on the island's south coast. In the southern interior the July mean is just above 60° F (16° C). January means in the southern portion of the island are above 20° F (−6.7° C), on the coast of Labrador approximately 10° F (−12.2° C) and 0° F (−17.8° C) in the south and north, respectively, and in interior western Labrador about −15° F (−26° C). Extreme low temperatures in southern Newfoundland rarely reach 0° F, but in western Labrador readings below −40° F (−40° C) are not uncommon. The annual precipitation varies from 55 inches (1,397 millimetres) in the southern parts of the island to about 17 inches (432 millimetres) at Cape Chidley. In the northern regions, as much as half of the annual precipitation occurs as snow, and snowfalls of more than 100 inches a year are not uncommon. In the south, the snowfall usually accounts for only 20 percent of the total precipitation.

Midlatitude storms, moving across Canada and up the Atlantic seaboard, have a pronounced effect on Newfoundland's climate. Warm air drawn in on the southern side of the disturbances contributes to the high precipitation in the southern part of the island. Northeast and east winds prevail in advance of each storm, blow across the cold Labrador Current, keep coastal summers cool, and retard the coming of spring. Northwest winds, following in the wake of each storm, carry in cold Arctic air that forces temperatures even lower than the latitude and marine location would suggest. To the south of the island the cool air of the Labrador Current mixes with the warmer air over the Gulf Stream to produce frequent dense fogs, which are most common over the Grand Bank and along the south and southeastern portion of the coast.

Forest cover

Plant and animal life. With the exception of the tundra region of northern Labrador and the barren reaches of higher altitudes and of some coastal regions, much of the province is forested. The principal species are conifers, of which balsam fir and black spruce are the most abundant. In most parts of the province, but particularly on the island, conifers are mixed with such leaf-bearing species as the paper and yellow birches and a wide variety of hardwood shrubs. The best stands of forest occur in areas of deep and well-drained soils, while in less-favoured areas much of the forest growth is stunted. In others, repeated fires and subsequent erosion have created barrens that produce a great variety of small woody plants, many of them producing edible fruit. Where glacial action has created poor drainage conditions, there are extensive boglands of organic material that vary in thickness from a few inches to 30 feet.

The moose is the most plentiful of the large wild animals found in the province. Other species include the caribou, the black bear, and the polar bear, and several species of seals. Small fur-bearing animals are typical of the northern coniferous forest and the tundra of Labrador.

Large seabird colonies inhabit the islands and headlands along the coast. Migratory ducks and shorebirds and wading birds in great numbers frequent the coast in their varying seasons. Inland, the principal game species are ptarmigan and snipe. Several species of ducks and geese maintain breeding populations. Birds of prey such as the peregrine falcon, the osprey, and the bald eagle find a respite from the environmental deterioration that has afflicted them in other habitats.

THE REGIONS OF NEWFOUNDLAND

The Avalon Peninsula. Newfoundlanders see their province as comprising nine regions, distinct in their natural settings and evidences of human presence. The most densely populated part of the island, the Avalon Peninsula on the southeast coast, contains the capital city, St. John's. Most of the coastal terrain is barren, the result of human deforestation. Apart from St. John's and a few small towns such as Harbour Grace, Carbonear, and Placentia, most of the peninsula's villages typically are located on a safe harbour and have evolved without any reference to town planning. The houses are generally of two-story frame construction, brightly painted, and surrounded by small gardens and meadows. They are located in sheltered positions often adjacent to such waterfront premises as stores, stages, wharves, and the fish-drying platforms called flakes. The helter-skelter arrangement of houses and other domestic buildings and of gardens, meadows, pastures, and woodlands, though sometimes shocking to persons accustomed to the precise streets, squares, and blocks of orthodox North American towns, represents, to others, a delightfully picturesque and romantic nonconformity. Only after about 1950 was the idea of the "planned community" introduced into town design throughout the island.

Patterns of settlement

Bonavista and Notre Dame bays. Northward from the Avalon, in the region of Bonavista and Notre Dame bays on the north coast, the landscape is somewhat softened. The exposed headlands and offshore islands are still barren, but in the deeper bays and estuaries the size and variety of the forest cover attest to deeper soils. The rocky shoreline is interrupted by stretches of sandy beach, and the fishing village occasionally is replaced by a farming community. Many densely forested islands shield the mainland from the worst effects of the northeast winds and create hundreds of miles of virtually landlocked waters. Many old communities such as Fogo and Bonavista cling to the more exposed locations, where the exigencies of the fishery demanded that they be built. These are being outstripped in growth, however, by the newer towns located in more favourable settings and engaged in the more recent industrial developments in paper manufacture, sawmilling, transportation, and distribution.

The Great Northern Peninsula. The region of the Great Northern Peninsula stretches northward toward the Labrador coast. On the west, magnificent sandy beaches and a narrow coastal plain dissected by numerous lakes and rivers are set against the scenic Long Range Mountains, which rise abruptly to heights of over 2,600 feet. On the eastern slope of the Long Range are good stands of commercial forest. Settlements in this region are more widely spaced along the coast than in other regions, with St. Anthony the largest town.

The west coast. The west coast region stretches southward from Bonne Bay, at the base of Great Northern Peninsula, to Cape Ray on the Cabot Strait. The northern section is well forested, while from St. George's Bay to the Codroy Valley the terrain is the best suited to agriculture in the province. The principal population centres are Corner Brook, a city based on a large pulp and paper complex, and Stephenville, originally developed in conjunction with a U.S. air base that is being converted to a centre for light industry.

The south coast. The most spectacularly rugged section of the island's coastline is probably the south coast, stretching from Cape Ray eastward to Fortune Bay. Here, the land generally rises abruptly from the sea, and the rivers are short and turbulent, but the head of Bay d'Espoir and the rivers and valleys that enter it are well forested. Many small settlements that once dotted the coast have been abandoned, and the population is now centred in fewer but larger communities such as Channel-Port-aux-Basques at the western extreme of the region, which is a major transshipment centre for freight and is the port of entry for rail and road traffic from mainland Canada.

The Burin Peninsula. Almost all settlement is coastal and directly related to the fishery in the Burin Peninsula region. The larger towns are Grand Bank, Fortune, and Marystown, all centres for the fresh-fish-processing industry, and St. Lawrence, which produces Canada's total supply of the mineral fluorspar, an important agent in metallurgy.

The island interior. The interior region is largely uninhabited. The pulp and paper industry at Grand Falls, however, has fostered the development of such towns as Windsor and Bishop's Falls, while mining and international air traffic have led to the growth of the towns of Buchans and Gander respectively.

Coastal and western Labrador. Labrador is composed of two very different regions. The coast south of the Hamilton Inlet, like the island of Newfoundland, is settled with scattered fishing villages. North of the inlet the five permanent settlements have predominantly native Indian and Eskimo populations. The only large towns are Goose Bay and Happy Valley, which developed around a military airfield. In western Labrador the twin mining towns of Labrador City and Wabush are modern, well planned, and prosperous, and Churchill Falls is the site of one of the world's greatest hydroelectrical developments. Here, as everywhere in the province, fishing provides a major recreation for all classes and ages.

THE PEOPLE AND THEIR WAY OF LIFE

Dominance of British peoples

Population composition. Newfoundland's people represent an extremely homogeneous group. Of the 522,000 at the 1971 census, about 95 percent are of British origin, the English, Irish, and Scottish and Welsh making up percentages of 75, 17, and 3, respectively. Nearly 4 percent are of French and Channel Island extraction. The native Indian and Eskimo peoples together comprise only 0.3 percent of the total. The homogeneity is accentuated by the fact that the vast majority of those persons of English ancestry trace their origins to the four western counties of Devon, Dorset, Somerset, and Cornwall. Historically, such cleavage as has existed has been along denominational rather than linguistic or ethnic lines. The three principal denominations are Roman Catholic, Church of England, and United Church of Canada, representing about 36, 29, and 21 percent of the total population, respectively. The birthrate has been falling steadily since 1950, reaching 24.0 per 1,000 of population in 1970, while the deathrate for the same year was 5.9 per 1,000. The net result is at present a natural increase of about 2 percent. Well over half of the population is below 25 years of age.

Since the beginning of the 20th century, emigration to the U.S. and, more recently, to other Canadian provinces has removed some of the pressure of the population. The exodus has probably, on occasion, reached 15,000 in a single year.

There were about 1,300 coastal villages, of which fewer than 100 had more than 500 inhabitants. The 20th-century trend toward urbanization has been accentuated by a vigorous government policy of resettlement designed to concentrate the people in fewer but larger and, in theory, more viable communities. Though hundreds of small villages remain, about 54 percent of the population is classified as urban, with nearly half of that number living in the metropolitan areas of St. John's and Corner Brook.

The provincial economy. *Mineral and forest products.* Newfoundland has been fortunate in being able to meet changing demands for products through exploitation of its rich mineral, forest, and marine resources. Historically, the fisheries brought Newfoundland into being, and until the beginning of the 20th century they were virtually the only industry. During the late 19th century mineral deposits were found both on the island and in Labrador, and the demands for paper gave great impetus to forestry. Mining has assumed the major role, its exports totalling nearly four times the value of pulp and paper products, more than five times that of fish products. This striking preponderance of mineral production is a direct consequence of the exploitation of western Labrador's vast iron reserves.

Power resources. The enormous hydroelectric-power resources of both the island and Labrador have been of the utmost importance in Newfoundland's industrial development. The Churchill Falls development in western Labrador, the first phase of which was completed in 1972, produces upwards of 7,000,000 horsepower of electrical energy. This great potential, as well as Newfoundland's position midway on the main shipping routes between North America and Europe, is crucial in planning the province's economic future. Attempts are being made to develop industries that require inexpensive electrical power and ease of ocean shipping and whose products can be marketed both in Europe and in North America. A phosphorus plant constructed at Long Harbour in the late 1960s represented a major step in this direction.

Energy potential

Modern fishery. Technical innovation has changed the nature of the fisheries. For centuries, the industry revolved about salted, dried codfish, caught in inshore waters and prepared for market primarily by many small family-run facilities. The majority of today's catch, comprising a wide variety of species, is taken by large trawlers in offshore waters and fresh frozen in large plants located strategically along the coast.

These developments reflect an increasingly complex and more broadly based economy. Secondary manufacture is of growing importance, and there is a rapid associated growth of service related industry. The relative values of these components of the economy are increasing rapidly, as is the absolute value of total production.

Transportation. Until the closing years of the 19th century, the sea was about the only highway facilitating transport and communication around the coast. A narrow-gauge railway was built later from St. John's to Channel–Port aux Basques along a route that touched the innermost parts of the major bays. Subsequently, branch lines were developed to serve regional centres of population and to connect with coastal steamships that carried passengers, freight, and mail to the hundreds of villages not served by the railway. Today a road system reaches nearly all communities, and a first-class paved highway crosses the island, more or less following the rail line. Some of the secondary roads are paved, but most have gravelled surfaces.

Connections with the mainland

The island is linked to mainland Canada by rail, car, and passenger ferries operating between Channel–Port aux Basques and Argentia on the south coast and North Sydney, Nova Scotia. National, international, and regional air service is provided by various lines.

Scheduled flights to tiny settlements throughout Labrador are of special importance during the frozen winters. In addition, a railway runs south from the western towns of Labrador City and Wabush to Sept-Îles, Quebec, on the Gulf of St. Lawrence. Further small road networks exist around principal population centres, and a short stretch of road runs alongside the Strait of Belle Isle.

Administration and social conditions. In most ways, administrative organization and social conditions in Newfoundland are typical of those in other Canadian provinces, though there was generally a slower development of institutions because of the unique history of the province.

Governmental organization. Under the terms of union with Canada, Newfoundland is represented in the Canadian Parliament by six senators and seven members of the House of Commons. The provincial government, completely autonomous in those areas reserved to its jurisdic-

tion by the British North America Act of 1867, consists of a lieutenant governor representing the crown and a unicameral legislature of 42 members elected by universal suffrage. The government is organized on the British parliamentary model. The executive council, or cabinet, is chosen from among the elected members of the party representing the majority in the House of Assembly. The Liberal Party has won every election since confederation, with the official opposition, the Progressive Conservative Party, unable to win more than 6 of the 42 electoral districts.

Local government was slow to evolve in Newfoundland. Not until 1888 was the first municipal government elected in St. John's, and the next town council was not created until 1938. By the 1970s, most of the settlements were incorporated as towns or as components of rural district organizations. Elected councils are empowered to levy taxes, and, under the general aegis of the provincial Department of Municipal Affairs and Housing, are responsible for the management of local concerns.

Justice. The Supreme Court of Newfoundland consists of a chief justice and other judges, sitting in St. John's but travelling annually on circuit to other parts of the province. Inferior district courts are established in federal electoral districts, and magistrates' courts serve many local communities.

Education. As in much of Canada, Newfoundland's educational system is a mixture of state and church activity. Historically, educational services were provided by the churches, which were then permitted to retain administrative controls even when the government assumed the major financial burden. The rights thus acquired were, at the time of confederation, enshrined in the provincial constitution, with the result that the provincial Department of Education was divided into five virtually autonomous denominational divisions. Throughout the province, more than 300 school boards duplicated schools and services, creating an increasing public outcry and, finally, a determination by the principal denominations themselves to rationalize their operations. In 1969 the Department of Education was reorganized along functional lines, school districts were consolidated with a drastic reduction in the number of school boards, and the principal Protestant denominations integrated their services. The individual denominations, however, continue to maintain their own centralized administrative offices and to exercise certain functions for which they are still totally responsible. Vocational, technical, and trade schools are nondenominational in character and are operated by the provincial government. Founded as Memorial University College in 1925, the Memorial University of Newfoundland, in St. John's, is the province's single university. It is nondenominational and managed by an autonomous board of regents.

Educational consolidation

Health and social services. The provincial government administers virtually all health services, which under the Medical Care Act are free to all residents. With few exceptions, members of the medical profession participate in the Medical Care Plan. Newfoundland differs from most other Canadian provinces in that few private agencies exist to provide services to the aged and children and to the variously disadvantaged. A notable exception is the International Grenfell Association, which since the beginning of the century has operated a hospital at St. Anthony and provided medical, educational, orphanage, and other social services on the Labrador coast and in the northern portions of the island.

Elsewhere, almost all such services are provided by the provincial government. This reflects both the very recent development of municipal institutions, the great dispersal of settlement, and the small population of most communities. Thus the police force that serves St. John's is not a municipal force but one that is, in theory, responsible for the whole of the province. In fact, however, they are assisted outside St. John's by the Royal Canadian Mounted Police, who operate under contract to the provincial government.

Quality of life. The same problems of population distribution and geography also combine to keep the cost of living somewhat higher and wage rates somewhat lower than in other parts of Canada. But statistics do not tell the whole story, for an inherent factor within a standard of living is the somewhat intangible notion of quality of life. Thus a cash income that statistically indicates subsistence below the poverty level might represent a comfortable living to a Newfoundland fisherman owning his own home, fishing equipment, and premises, as well as a few acres of cultivated land.

Cultural life and institutions. The insular location of Newfoundland and the internal isolation of hundreds of small communities have combined to preserve distinctive cultural characteristics that have long since disappeared in the land of their origin. Thus a most distinctive feature of Newfoundland is the local speech, which preserves much of the flavour of its West Country origins in 17th-century England. Over several centuries of comparative isolation, it has been modified only by the influence of the Irish who at one time constituted half the total population. But there is really no such thing as a Newfoundland accent, for there are almost as many dialects as there are bays. The differences between these dialects and "standard English" are not merely in accent but in vocabulary, in sounds, and in syntax.

Though isolation produced distinctive language patterns and usages, it created as well the necessity for local entertainment. This factor, perhaps in combination with the Irish traditions of bards and minstrels, has encouraged the development of the art of the storyteller and the preservation and creation of an immense body of folk music and song. Indeed, Newfoundland has been categorized as a living archive of folklore, folktale, and folk song. With a few notable exceptions, however, Newfoundlanders have not been exceptionally productive in the visual arts. Why the imaginative flair and inventiveness so apparent in the field of folk music has not carried over into other art forms can be no more than speculated upon. Professional theatre was driven out by television, but amateur groups have maintained the increasingly healthy theatre-going tradition.

Folk culture

St. John's has always been well served by newspapers, although the score or more that appeared during the 19th century and flourished for greater or lesser periods of time have now been replaced by two dailies, the *Evening Telegram* and the *Daily News*. A version of the *Evening Telegram* known as the *Western Star* is published in Corner Brook, and several gossipy weeklies serve other outpost regions.

But Newfoundland's distinctive culture has begun to recede before the impact of mainstream North American influences, which were first experienced acutely when World War II caused a large influx of military personnel. The increase in the amount of radio and television transmission and the influx of large numbers of professional people from elsewhere have hastened the process of change. A particular and pressing care of the university and of the Folk Arts Council is to prevent a total loss of the distinctive elements of the folk culture, to keep them not as museum pieces but as vital force in the province's life.

Prospect. By comparison with the materially more advanced provinces of Canada, Newfoundland is an underdeveloped region. But it is, nevertheless, a region rich in natural resources, not the least of which may lie beneath the sea that covers the extensive areas of continental shelf adjacent to the coasts.

The task of building the infrastructure of a modern society after centuries of neglect is costly in terms both of money and of time. Few peoples are more passionately attached to their homes than are Newfoundlanders to their ruggedly beautiful land. Few North Americans are more conscious of their history. They are now challenged by the task of preserving the distinctive elements of their identity while creating a modern and prosperous society. Progress has been rapid, however, and there is every reason to believe that continued sound economic policies and governmental stability will join with the determined individualism of the Newfoundlanders in bringing the richness of the past into an even more bountiful future.

BIBLIOGRAPHY. D.W. PROWSE, *A History of Newfoundland from the English, Colonial and Foreign Records* (1895), is the standard general history of Newfoundland but is now somewhat dated and is weakest in its treatment of the 19th century. J.D. ROGERS, *Newfoundland*, vol. 5, pt. 4 of *A Historical Geography of the British Colonies* (1911), is a delightfully written general history that complements Prowse. A.H. MCLINTOC, *The Establishment of Constitutional Government in Newfoundland, 1783–1832* (1941); and GERTRUDE GUNN, *The Political History of Newfoundland, 1832–1864* (1966), are detailed and scholarly works that capture the essence of Newfoundland politics during the struggle for and the early years of self-government. The historical beginnings of the fisheries, for many years the *raison d'être* of Newfoundland, are admirably covered in R.G. LOUNSBURY, *The British Fishery at Newfoundland, 1643–1763* (1934, reprinted 1969); while the economist's position is stated in H.A. INNIS, *The Cod Fisheries*, rev. ed. (1954), which is a veritable mine of information. Little of a definitive nature has been written to cover the 20th century, but ST. JOHN CHADWICK, *Newfoundland: Island into Province* (1967), is a useful survey while R.I. MCALLISTER (ed.), *Newfoundland and Labrador: The First Fifteen Years of Confederation* (1966); and the NEWFOUNDLAND GOVERNMENT, *Historical Statistics of Newfoundland and Labrador* (1970), provide essential basic data. R.A. MACKAY (ed.), *Newfoundland: Economic, Diplomatic and Strategic Studies* (1946), is a valuable collection of papers on miscellaneous aspects of Newfoundland's role in international affairs; and G.S. GRAHAM, *Empire of the North Atlantic* (1950), fits the island into its North Atlantic context. The development of educational services in Newfoundland is a fascinating story told in F.W. ROWE, *The History of Education in Newfoundland* (1952). Two dominant aspects of Newfoundland folk life are illustrated in KENNETH PEACOCK, *Songs of the Newfoundland Outports*, National Museum of Canada Bull. No. 197, 3 vol. (1965), a valuable collection of folksong; and H. HALPERT and G.M. STORY (eds.), *Christmas Mumming in Newfoundland* (1969), a group of fascinating essays by sociologists, historians, folklorists, and anthropologists.

(L.H.)

New Guinea, Mountain Ranges of

The mountains of New Guinea extend from northwest to southeast throughout the length of the world's second-largest island. They form a lofty and almost impenetrable cordillera (mountain chain) that constitutes an impediment to linguistic and cultural unity among its peoples. From 50 to 150 miles (80 to 240 kilometres) wide, and frequently higher than 12,000 feet, the massive complex of ranges runs approximately 1,500 miles (2,400 kilometres) through the Indonesian province of Irian Barat (West Irian) in the western half of the island, and, in the eastern half, along the frontier between the Australian-administered UN Trust Territory of New Guinea to the northeast and Australian-administered Papua to the southwest (jointly administered as Papua New Guinea).

Spectacularly steep, rain-drenched, and often crowned with snow, the mountains stand as a heavily-wooded fortress guarding impassable ridges, broad highland valleys, and plateaus that occur at altitudes of about 5,000 feet. On the northern coast another mountain complex that is nearly parallel but lower—reaching heights between 3,000 and 7,000 feet—rises abruptly from the Pacific Ocean. Large undulating, stable plains, part of the Australian continental shelf, dominate southern New Guinea.

The mountains. The Arfak Mountains in the northwestern Vogelkop Peninsula, which have a maximum elevation of about 10,000 feet, form the westernmost part of the system. Their southeastern foothills cross the isthmus between Geelvink and Bintuni bays to reach the main part of the island. Here the central cordillera rises to form a great mountain backbone that extends without interruption for 1,100 miles. From west to east the chain is composed of the Snow Mountains (Pegunungan Maoke, comprising the Nassau and Orange ranges), the Star Mountains, and the Hindenburg, Victor Emmanuel, Central, Bismarck, Kubor, Kuper, and Owen Stanley ranges. On approaching their New Guinea terminus at Milne Bay, the mountains diminish in height and width. A partly drowned extension between the Coral and Solomon seas appears to the southeast of New Guinea as the Louisiade Archipelago.

The western ranges

The western ranges are generally higher than those in the east and include many peaks that extend above the snow line, which occurs at about 14,000 feet. The highest point on the island is the summit of Mount Carstensz (Puntjak Sukarno), which reaches a height of 16,404 feet. Other important summits include Mount Idenburg (15,750 feet), Wilhelmina (15,585 feet), and Leonard Darwin (13,887 feet). An enormous glacier-capped ridge extends for approximately 8 miles between Mounts Carstensz and Idenberg. Gigantic cliffs border the ridge on both its northern and southern sides, presenting a breathtaking face of clear ice with a maximum vertical drop of 10,000 feet. Although the snowfield of Mount Carstensz was reached as early as 1909 by the Dutch explorer H.A. Lorentz, it was not until 1962 that its summit was climbed by an expedition led by the Austrian Heinrich Harrer.

In eastern New Guinea, Mount Wilhelm at 14,793 feet rises in the centre of the Bismarck Range. Extinct volcanoes, notably mounts Giluwe (14,330), Hagen (12,392), and Ialibu (11,369) and the Doma Peaks (11,700), are other dominant peaks. The Great Papuan Plateau, a plain made of volcanic debris, surrounds the smooth, symmetrical cone of Mount Bosavi (7,863).

The parallel northern system is—from west to east—composed of the Pegunungan Gauttier and Pegunungan Van Rees mountains, and the Torricelli and Finisterre ranges. The continuity is broken by deep troughs of these ranges, through which the inland lake-plain that lies between the central and coastal ranges can be reached. The plain, which is 100 feet above sea level, is drained by the Mamberamo, Sepik, and Ramu rivers; the river mouths are surrounded by broad swampy lowlands that contrast with the otherwise rugged character of the northern coast.

Geology. The cordillera lies in a zone of weakness in the earth's crust separating the relatively stable Australian continent from the basin of the Pacific Ocean. The ranges formed an axis of geological activity from 225,000,000 to 2,500,000 years ago. Formed of essentially young rocks that have been intensively folded, the ranges show evidence of violent movements of the earth's crust. Original folding that took place from 136,000,000 to 65,000,000 years ago was compounded by more severe folding that occurred from 65,000,000 to 2,500,000 years ago. Marine beds were uplifted to a maximum height of 16,000 feet, while at the same time, downward movements produced the depressions to the north and south of the central highlands; parts of the lands that had been raised also became isolated, forming island chains. Lakes Paniai, Tage, and Tigi, which were formed by folding that took place across the course of the Arabu River, lie at a height of about 5,750 feet in the region around Enarotali in the west.

With wind and water erosion occurring at a rapid rate, the mountains are much cut into hills and valleys; they have long, straight slopes and a thick vegetation cover. Chemical erosion is caused by the heat and rain and the rain forest cover encourages intense bacterial activity.

To the south the foothills are made of limestone, which permits water to filter through, thereby forestalling stream erosion; outcropping limestones consequently often form prominent ridges. Elsewhere the alluvium resulting from erosion is coarse; flooded streams occasionally transport boulders as large as 30 to 40 feet in diameter.

Although frosts occur above 10,000 feet, significant frost weathering is not usually observed below 12,000 feet. Glaciated landscapes indicate that a snow line existed at heights of 11,500 to 12,000 feet from 2,500,000 to 10,000 years ago; glaciation has nevertheless been a minor factor in landscape development.

Active volcanoes

Volcanic activity in recent times has been limited to the coastal region of the north and east; Mount Lamington, in the Owen Stanley range of eastern New Guinea, erupted in 1951, killing or injuring 4,000 people, and other eruptions have occurred on smaller islands in the Bismarck Sea. Volcanic activity has also recently occurred on the Vogelkop Peninsula.

Drainage. The central highlands form the watershed of a complex drainage system that includes several large rivers. From west to east these rivers are the northern

Tariku, the Taritatu, the Sepik, the Ramu, and the Markham rivers, and the southern Pulau, the Digul, and the Fly. From its source in the Star Mountains, the Sepik meanders for 700 or 800 miles to the northern sea. The Ramu and Markham rivers flow in opposite directions from a watershed in the middle of a valley between the Finisterre and Bismarck ranges. All these rivers carry huge amounts of alluvium. The largest Papuan river, the Fly, begins its more than 600-mile course in the Victor Emmanuel Range; at the river's mouth, which is 33 miles wide, a flood plain is being built up at a rate that approximately matches the rate at which the southern coast is subsiding.

Soil. Except for volcanically active areas, such as the Vogelkop Peninsula and the northeastern coast, the soils are heavily weathered and leached. Steep slopes usually support deep acid soils, commonly brown clays derived from plants. Peaty soils occur at high altitudes as well as in swampy areas. Shallow dark clays coat the steep limestone slopes; deep dark clays cover limestone floors packed with funnel-shaped cavities, and slumps (masses of material that have slipped down from high slopes). Olive ash soils, derived from plants, occur at moderate altitudes.

Climate. The climate is tropical and is dominated by two prevailing wind systems—the northwest monsoon wind, which blows from December to March, and the southeast trade wind, which prevails from May to October. The average site experiences ten inches of precipitation monthly; the amount is subject to only slight seasonal variation, although rainfall is heaviest during the northwest monsoon. In many districts higher than 5,000 feet, rain is practically incessant. Because the mountains occupy a central position between the moist air masses of the Bismarck Sea in the northwest and the Coral Sea in the southeast, every wind brings moisture that falls as rain when it reaches the mountains.

Rainfall and temperature

Temperature in the more-populated valleys varies between 50° and 80° F (10° and 27° C) with a daily range of 10° to 55° F (6° to 31° C). Intermontane plateaus, between the mountains, such as the upper Chimbu Plateau (which lies at an altitude of about 6,000 feet between the Finisterre Range to the east and various ranges to the west), average 10° F (6° C) less. Ground frosts are common at elevations above 10,000 feet.

The type of climate varies with altitude. A tropical monsoon climate predominates below 3,500 feet, developing into a wet, tropical mountain climate up to an altitude of 6,000 feet. A wet–cool mountain climate with persistent cloudiness occurs between 6,000 and 11,000 feet; above 11,000 feet a high mountain-steppe climate prevails.

Vegetation. With its dry plains, humid jungles, and freezing windswept peaks, New Guinea has a highly diversified plant life. An amazing wealth of approximately 3,000 different species of orchid, many of which are peculiar to the locality, enriches the vegetation. Near the foot of the mountains the undergrowth contains screw pine and creeping bamboo. Along the lower mountain valleys, at heights of from 3,000 to 5,000 feet both oaks and laurels mingle with valuable evergreen trees. Mountain forests of evergreen and oak are found at altitudes between 5,000 and 9,000 feet; the forest floor is carpeted with moss and fern, while Rhododendron and over 200 plants of the ginger family provide bright colours. Up to 8,000 feet hunters and cultivators have frequently reduced the forest to grassland. A moorland of low scattered bushes with fern and cycad is characteristic of the highest altitudes.

Animal and bird life. The diversity in climate and vegetation has produced a correspondingly varied animal life. Mammals are related to those of Australia, and consist of marsupials (animals carrying their young in pouches), rats, mice, bats, spiny anteaters (which lay eggs), crocodiles, and wild boars. Both venomous and harmless snakes occur. Rivers are poor in freshwater fish. Frogs and tree frogs abound. The most spectacular birds are the colourful birds-of-paradise, of which there are approximately 80 species. Formerly slaughtered for their feathers, they are now protected. The cassowary, a large, flightless bird, is domesticated for its feathers. Sulfur-crested cockatoos (*i.e.*, with lemon-coloured crests) are prominent among the 46 parrot species, and the crowned pigeon is also found.

Birds-of-paradise

The inhabitants. The mountain tribesmen are predominantly Negritos, who represent remnants of an early, widespread migration from southern Asia. The island probably first received human migrants about 25,000 years ago. Radio-carbon dating of stone tools found at Chimbu and Goroka has established that human settlement occurred in the highlands at least 10,000 years ago. The official population density is estimated at about 17 persons per square mile in the Australian territories, and about five per square mile in West Irian, but densities are undoubtedly highest in highland valleys.

Human adaptation to the environment, recurrent Melanesian intrusions, and genetic variations produced many different groups. Montane isolation in the mountains has led to the development of cultural and linguistic differences. Most of the people, however, are animists practicing agriculture on land tracts that are constantly changed, and cultivating bananas, sugarcane, taro (a stemless plant with a starchy edible root), and tobacco. Sweet potatoes, introduced in about 1600, permitted a more vigorous agriculture to develop at altitudes too high for the traditional crops grown at the 5,000-foot level. Skillful cultivation enables slopes standing at angles as steep as 70° to be used without causing surface erosion. Pigs are domesticated for religious ceremonies as well as to establish group prestige. The casuarina tree, which has a valuable hard wood, is planted for fuel and is also used for fencing in the few places where population pressure has resulted in semi-permanent cultivation. The main tribal groups are the Ekaris and Monis of the Arfak Mountains; the Tapiro, Pesechem, and Nogullo of the Nassau Range; and the Kukukukus and the Wabago of eastern New Guinea. The Enga language, spoken by 130,000 eastern highlanders, is the most widely used of more than 1,000 Papuan and Malayo-Polynesian languages that are spoken.

Mapping. Although progress has been hampered by persistent cloud cover, by 1970 about 85 percent of the mainland had been adequately photographed from the air for mapping purposes.

Transport. In West Irian few mountain passes occur below 13,000 feet. Rivers, gorges, and mountain barriers have formed serious obstacles to normal movement. Until 20 years ago, neighbouring groups were usually strangers and, therefore, enemies. The airplane has, to some extent, served to overcome both natural and human barriers to communication. There are approximately 60 airstrips on the island, some of which bear heavy traffic. At Goroka, in the U.N. Trust Territory of New Guinea, about 850 landings a month are recorded.

Overland transport remains poor, particularly in West Irian, where the only road networks are in the coastal lowlands. In eastern New Guinea most of the roads can only be traversed by four-wheel-drive vehicles or by tractors. The largest and best-maintained road system is the Highlands Highway, which extends 400 miles west from Lae on the east coast of the Trust Territory and crosses the 8,175-foot Daulo Pass.

Economic resources. While gold, oil, and copper have long been mined at coastal locations, the difficulty of access has delayed exploitation of the rich mineral, timber, and hydroelectric resources of the mountains. Small amounts of timber are obtained in the eastern districts. In the Trust Territory of New Guinea there is a small hydroelectric plant at Goroka, and another hydroelectric project is planned on the upper Ramu River.

BIBLIOGRAPHY. R. GERARD WARD and DAVID A.M. LEA *et al.*, *An Atlas of Papua and New Guinea* (1970), a comprehensive analysis of resources, production, and potential with illustrative maps; PETER RYAN (ed.), *Encyclopaedia of Papua and New Guinea*, 3 vol. (1972), done in collaboration with the University of Papua and New Guinea, the best recent work on the geography, history, society, and economy of the area, with valuable bibliographies; GAVIN SOUTER, *The Last Frontier* (1963); and OLAF RUHEN, *Mountains in the Clouds*

(1963), popular histories of the island; MOHAMMAD ANAS, "The Highlands of Australian New Guinea," *Geogrl. Rev.*, 50:467–490 (1960), a descriptive article; T.W.E. DAVID, *The Geology of the Commonwealth of Australia*, 3 vol. (1950), a standard work; H.A. HAANTJENS *et al.*, *Major Soil Groups of New Guinea and Their Distribution*, Communication no. 55 of the Dept. of Agricultural Research of the Royal Tropical Institute, Amsterdam (1967), explanatory material and photographs supplementing soil, rainfall, and climate maps; L.G. BRONGERSMA, *The Animal World of Netherlands New Guinea* (1958), a popular description, illustrated; H.D. BROOKFIELD and J. PETER WHITE, "Revolution or Evolution in the Prehistory of the New Guinea Highlands," *Ethnology*, 7:43–52 (1968), an ethnological description; JACK GOLSON, "Both Sides of the Wallace Line," in *Early Chinese Art and Its Possible Influence in the Pacific Basin*, ch. 17 (forthcoming), an analysis of cultural influences; H.A. JOHNSON, *The High Level Geodetic Survey of New Guinea*, a technical report of the Division of National Mapping of the Australian Department of National Development (1969).

(O.R.)

New Guinea, Trust Territory of

The United Nations Trust Territory of New Guinea, under Australian administration, includes the northeast quarter of the island of New Guinea (east of longitude 141° E), the islands of the Bismarck Archipelago, Bougainville and Buka islands, and about 600 small offshore islands. It is bounded to the west by West Irian (part of Indonesia), to the north and east by the Pacific Ocean, and to the south by Papua. The territory has a total area of 92,160 square miles (238,700 square kilometres) and a population of nearly 1,800,000 in 1970. The administrative capital, Port Moresby, is on the south coast of Papua, an Australian external territory. Papua and the Trust Territory of New Guinea are administered jointly. The territory is visited every three years by a UN visiting mission, which monitors Australia's fulfillment of its international trusteeship. (For Papua, see AUSTRALIAN EXTERNAL TERRITORIES; for an associated physical feature, see NEW GUINEA, MOUNTAIN RANGES OF; for further historical aspects, see OCEANIA, HISTORY OF.)

Relief, drainage, and soils. The most prominent physiographic feature of the territory is a complex system of ranges running roughly northwest to southeast through the centre of the island of New Guinea, approximately along the border between the trust territory and Papua. This Central Cordillera is broadest between the towns of Kainantu, to the east, and Wabag, to the west, where plains 4,000 to 6,000 feet high are flanked by high mountain ridges. The highest mountain is Mt. Wilhelm, in the Bismarck ranges; it rises to 14,793 feet (4,509 metres). North of the Central Cordillera is a fault trough through which drains the Sepik River (some 700 miles long and navigable for more than 300 miles) and the Ramu and Markham rivers. North of the fault trough, a northern chain of mountains runs parallel with the main cordillera, but it is not so wide, lofty, or continuous. Only in the Saruwaged Ranges of the Huon Peninsula do peaks exceed 13,000 feet. Raised coral reefs are a feature of the north coast of New Guinea. Some reefs on the Huon Peninsula rise to 2,000 feet above present sea level.

The main islands of New Britain, New Ireland, Manus, and Bougainville are hilly and mountainous, with valleys and gorges inland. Like the northern coast of New Guinea, some parts of their coastlines have been subject to uplifting during the last 10,000 years, but there are many coastal plains (for example, in West New Britain) and coastal mangrove swamps. The numerous lesser islands are of volcanic or coral origins.

Volcanic and seismic activity

The territory is geologically unstable and subject to much seismic and volcanic activity; 32 volcanoes (mainly in the Bismarck Archipelago) either have erupted in the last 80 years or are in a latent volcanic stage in which they produce hot vapours and gases.

Soils generally are very leached and infertile. Some exceptions are the volcanic soils in New Britain and Bougainville and the alluvial soils of the Markham and Ramu valleys.

Climate, vegetation, and animal life. About half of the territory receives more than 100 inches of rainfall a year, but some mountainous parts probably receive well in excess of 300 inches. Relatively dry zones receiving less than 70 inches annually occur in the lower Sepik and Markham valleys. Maximum rainfall typically occurs between January and April; between May and August, normally a relatively dry period, a markedly wet season nevertheless occurs in the southeast parts of the Huon Peninsula, New Britain, and Bougainville.

Temperature varies little throughout the year. In coastal areas average monthly maximum temperatures are about 86° F (30° C) and average minimum temperatures about 73° F (23° C). Daily temperature variation is about 10° to 15° F (6° to 8° C). In the highland valleys temperatures are lower and extremes are greater. On the tops of the highest mountains snow occasionally falls.

Lowland rain forest is the dominant vegetation cover. There is also a wide range of other plant communities, including coastal mangroves and strand forest, lowland grassland, woody and herbaceous swamps, moss forests in the mountains at about 9,000 feet, and tussock grasses on the peaks above 13,500 feet.

The animal life is closely related to that of Australia. Among mammals, marsupials such as the tree kangaroo and the phalanger (also a tree dweller) predominate. The spectacular birds-of-paradise occur in sparsely populated areas.

History. On the basis of associated evidence in Australia, which was joined to New Guinea at certain periods in the Pleistocene Period (which lasted from about 2,500,000 to about 10,000 years ago), it is probable that man entered the New Guinea region about 50,000 years ago. Diverse hunting and gathering activities persisted until about 6,000 years ago, when evidence of simple subsistence horticulture appears. Great differences in existing languages and agricultural systems indicate that coastal and inland groups were never closely related.

Portuguese and Spanish navigators sighted the north coast of New Guinea in the early part of the 16th century. Until German trading firms started operations in the New Guinea islands in the 1870s, only the coast was mapped, and very little was known about the interior of the mainland or of the islands.

German, Australian, and Japanese occupations

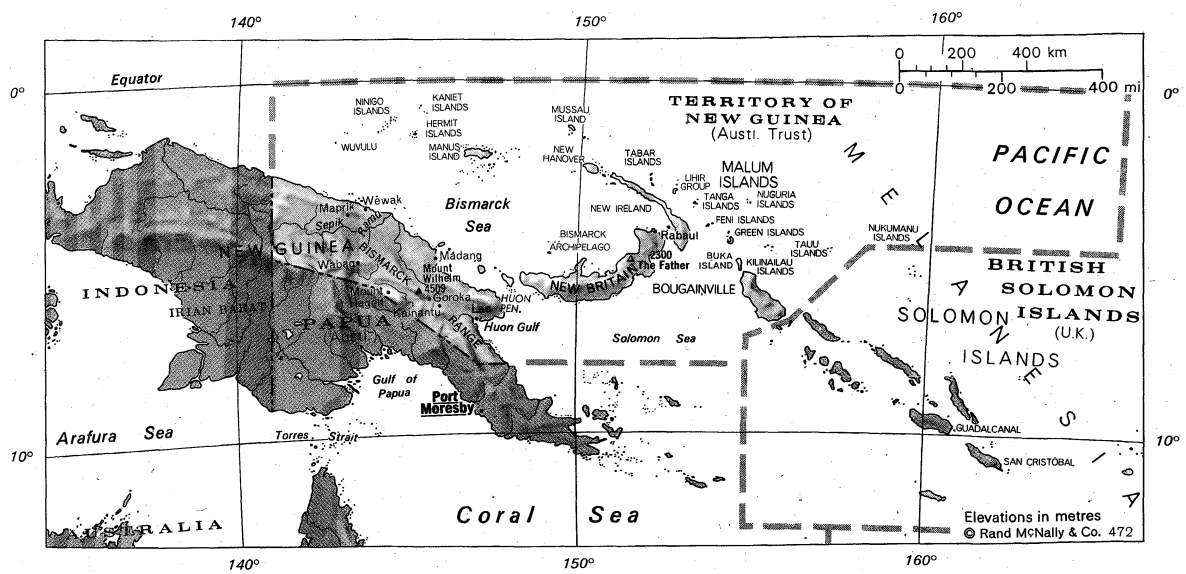
In 1884 Germany formally took possession of the territory, which was administered by the German New Guinea Company until 1899, when the German imperial government assumed control. In 1914, on the outbreak of World War I, the territory was occupied by Australian troops and remained under military administration until it became an Australian mandate of the League of Nations in 1921. By 1923 a strip extending only about 50 miles inland from the coast was under effective Australian control. It was not until the 1930s that the densely populated highland valleys were discovered. During the 1930s the foreign-owned coconut plantations were adversely affected by the economic depression, and the economy of the country was based largely on the export of gold from the Bulolo Valley.

Between 1942 and 1945, during World War II, the islands and lowland areas of the territory were occupied by Japan. Since 1946 the territory has been administered by Australia under the United Nations trusteeship system.

People, population, and culture. The indigenous people, officially "Australian protected persons," are ethnically diverse. Broadly, the so-called Papuan is representative of this interior, and the Melanesian is representative of the coasts and islands. There is, however, also some mingling with Micronesian and Polynesian peoples.

Two radically different indigenous language types exist —Austronesian or Melanesian, and non-Austronesian or Papuan. Some 200 related Austronesian languages occur, mainly in the islands and along the New Guinea coast. The 500 non-Austronesian languages have small speech communities, but the largest is the Enga, in the Wabag area, which has some 140,000 speakers. Because of the multiplicity of tongues, Pidgin English (Neo-Melanesian) has developed as an effective lingua franca.

In 1970 the estimated indigenous population was about 1,763,000 and was increasing at a rate of about 2.4 percent annually. The average population density of the



TRUST TERRITORY OF NEW GUINEA

whole territory is about 19 persons per square mile, but they are unevenly distributed. About 30 percent of the population are in the highlands, with some concentrations of between 200 and 500 persons per square mile. Other clusters occur in north New Britain and in the Maprik area of the Sepik District. The main towns

Principal towns

are Lae (population 35,000 in 1971), Rabaul, Wewak, Madang, Goroka, and Mount Hagen. Only about 5 percent of the total population are urban. The nonindigenous population, mainly Australian, numbered just over 20,000 in 1966, and about 70 percent lived in the towns. Many indigenous groups have rich and varied forms of artistic expression and an extensive ceremonial and ritual life.

New Guinea (UN Trust Territory of), Area and Population

	area*		population†	
	sq mi	sq km	1966 census	1970 estimate‡
Districts				
Bougainville	4,100	10,619	72,000	78,000
Chimbu	2,800	7,252	167,000	185,000
Eastern Highlands	5,000	12,950	204,000	229,000
East New Britain	7,000	18,130	110,000	99,000
East Sepik	10,500	27,195	158,000	200,000
Madang	10,800	27,972	152,000	175,000
Manus	800	2,072	21,000	22,000
Morobe	12,700	32,893	209,000	244,000
New Ireland	3,800	9,842	50,000	51,000
Western Highlands	9,200	23,828	291,000	329,000
West New Britain	7,100	18,389	44,000	55,000
West Sepik	19,700	51,023	100,000	97,000
Total New Guinea	93,500§	242,165‖	1,579,000¶	1,763,000¶

*Approximate areas. †Boundary changes since 1966 do not permit direct comparison of census data with estimates; adjusted figures not yet available. ‡Indigenous population only. §Total of approximate areas only; reported total area of New Guinea is 92,160 sq mi (238,700 sq km). ‖Converted figures do not add to total given because of rounding. ¶Figures do not add to total given because of rounding.
Source: Official government figures.

The economy. Agriculture is the most important element in both the subsistence and commercial sectors of the economy. Yams, taro, sago, and bananas are the lowland staple foods, and sweet potato is the main highland food. Commercial cash crops account for 85 percent of the foreign exchange earned by the territory. The crops are grown on plantations owned by the nonindigenous population or by indigenous smallholders. The main crops are coffee, cocoa, and copra, which together account for more than half of the total exports. Other important cash crops are pyrethrum (a plant used in the making of insecticides), tea, rubber, and oil palm. Gold, produced principally in the Wau-Bulolo area of the Morobe District, is declining in importance, but huge deposits of low-grade copper have been found in the Panguna area of Bougainville. Copper concentrate is expected to become a major source of income. It is estimated that there is a 900,000,000 ton reserve of copper ore at Panguna. Production began in 1972.

Besides revenue raised internally, mainly from import tariffs and direct taxation, the territory receives a non-repayable grant from the Commonwealth of Australia.

The main road system is the Highlands Highway, which runs for 400 miles westward from Lae. There are smaller road networks in the Sepik District and in New Ireland; limited road systems radiate out from the main towns. Coastal trade is important in moving agricultural produce to the main ports, but aircraft have played an important role in opening up and developing the territory. Aircraft are the main method of travel within the territory, and most of the inland centres and outstations are served by small planes.

Highways, coastal shipping, and air traffic

Administration. The territory consists of 12 districts and 50 subdistricts and is administered jointly with Papua by an administrator, appointed by the Australian government and based in Port Moresby. The main legislative body is the House of Assembly, which consists of 84 elected members and ten appointed; any ordinance passed by the House, however, must have the assent of the administrator and—in some cases—the governor general of Australia.

BIBLIOGRAPHY. COMMONWEALTH OF AUSTRALIA, *Territory of New Guinea: Report for 1968–1969* (1970), official report to the United Nations, containing numerous statistical tables; H.C. BROOKFIELD, *Melanesia: A Geographical Interpretation of an Island World* (1971), the best single interpretative description of New Guinea covering social and environmental conditions; P. HASTINGS, *New Guinea: Problems and Prospects* (1969); D.R. HOWLETT, *A Geography of Papua and New Guinea* (1967), a general work, somewhat outdated; INTERNATIONAL BANK FOR RECONSTRUCTION AND DEVELOPMENT, *The Economic Development of the Territory of Papua and New Guinea* (1965), a blueprint for development; D.A.M. LEA and P.G. IRWIN, *New Guinea: The Territory and Its People* (1967), a general school text; G. SOUTER, *New Guinea: The Last Unknown* (1963), a readable historical survey; J. TUDOR (ed.), *Handbook of Papua and New Guinea*, 5th ed. (1966); P.W. VAN DE VEUR, *Search for New Guinea's Boundaries from Torres Straits to the Pacific* (1966), an authoritative historical description of how New Guinea's irrational boundaries were drawn; R.G. WARD and D.A.M. LEA, *An Atlas of Papua and New Guinea* (1970), an excellent reference work dealing generally with exploration, population, migration, religion, language, soil, relief, economic conditions, and schools; PETER RYAN (ed.), *Encyclopaedia of Papua and New Guinea*, 3 vols. (1972), done in collaboration with the University of Papua and New Guinea, the best recent work on the geography, history, society, and economy of the area, with valuable bibliographies.

(D.A.M.L.)

New Hampshire

New Hampshire, one of the 13 original states of the United States, is situated with its five fellow New England states at the extreme northeastern corner of the nation. With an area of 9,304 square miles (24,097 square kilometres), it is bounded on the north by the Canadian province of Quebec, on the east by Maine and an 18-mile-long stretch of the Atlantic Ocean, on the south by Massachusetts, and on the west by Vermont. The capital is Concord, located in the south central part of the state.

Stability and progress

Though the essence of New Hampshire may be symbolized by its underlying granite bedrock, by the stern visage of the rock profile known as the "Old Man of the Mountain" looking out from Cannon Mountain in Franconia Notch, and by its more than three centuries of staunch Yankee tradition, New Hampshire contains within itself a number of contradictions. The 1970 census showing 737,681 inhabitants ranked it only 42nd among the 50 states, but a 21.5 percent population growth in the 1960s made it the third-fastest-growing state east of the Mississippi River. It is generally thought to be a rural state, but about 40 percent of its residents' wages and salaries were received from manufacturing in 1970, making it by this measure the fifth most industrialized state in the nation. In 1964 New Hampshire became a national pioneer when a statewide sweepstakes went into effect, its income earmarked for education. This innovation was necessitated, however, by the lack of either a general sales tax or a state income tax, two measures the citizens and state legislature have been loath to accept.

On the side of tradition, New Hampshire seems very much a region tied to the past, to the land, and to the conservative way of life. Many of its residents can trace their New Hampshire ancestry for generations. The names of many of its towns and 1,300 lakes and ponds—Sunapee, Winnipesaukee, and Kearsarge, for example—are picturesque reminders of the Indians who inhabited northern New England before the coming of the white man in the 17th century. Its towns are governed by the town meeting, a last bastion of direct democracy in which the citizens themselves vote on the measures that will direct local affairs for the coming year. Despite these echoes from the past, its industries produce space-age materials. For more information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; and NORTH AMERICA.

HISTORY

The English colony. The region that was to become New Hampshire was included in a series of grants made by the English crown to one John Mason and others in the 1620s. A fishing and trading settlement was established in 1623, and in 1629 the name New Hampshire, after the English county of Hampshire, was applied to a grant for a region between the Merrimack and Piscataqua rivers. The towns of Dover, Portsmouth, Exeter, and Hampton were the main settlements in the early decades.

From 1641 to 1679 the region was administered by the colonial government of Massachusetts. Following decades of territorial and religious dispute between Massachusetts and Mason's heirs, New Hampshire became a separate royal province in 1679. Bitter boundary feuds with Massachusetts and with New York over that part of the New Hampshire grant that became Vermont continued almost until the American Revolution. Most famous of the colonial governors was Benning Wentworth, who held the post from 1741 to 1767, the longest tenure of any royal governor in any of the colonies.

In 1767 the colony had its first census and counted about 52,700 inhabitants. In 1769 it was divided into five counties, to which five others have been added since 1800. New Hampshire soldiers played an active part in the colonial wars between Great Britain and France from 1689 to 1763. By the end of the colonial period the seat of government was at Portsmouth, and there were 147 chartered towns in the province.

Revolution and statehood. In December 1774 there was armed resistance to the British at New Castle, where Ft. William and Mary was seized by colonists. Gunpowder and other munitions were taken. New Hampshire people were overwhelmingly in sympathy with the aims of the revolutionary leaders. The state furnished two major generals to the Continental Army, three regiments of regular troops, and hundreds of short-term militiamen. Officially becoming a state in 1776, it issued its own Declaration of Independence some weeks before the national Declaration of July 4, 1776. Gen. John Stark of Derryfield led the Continental forces in a major victory at Bennington, Vermont, in 1777. New Hampshire's vote was the ninth and decisive one in ratifying the Constitution of the United States in 1788.

Role in forming the nation

Following the establishment of the new nation, the state grew rapidly. Agriculture, notably sheep raising, flourished and manufacturing developed along the fast-flowing rivers, particularly in Manchester. When the railroads appeared in the northeast, an extensive rail network was constructed. Portsmouth and its surrounding towns were shipbuilding centres. In 1846 Manchester became the first incorporated city in the state. On the national scene New Hampshire was the birthplace of such noted statesmen as Daniel Webster and Pres. Franklin Pierce.

Contributions to war and politics. New Hampshire played an active part in the Civil War and, following that conflict, voted regularly for Republican candidates until the New Deal of the 1930s. The progressive movement had an impact upon the state in the early 20th century, and the economic depression of the 1930s left scars that were only slowly effaced. In both world wars and in the Korean and Vietnamese conflicts, the state's men played significant roles, and the shipbuilding and naval facilities at Portsmouth were of special value.

Because it holds the earliest presidential primary in the nation, New Hampshire has furnished the first testing ground for many successful candidacies, including those of Dwight D. Eisenhower in 1952, of John F. Kennedy in 1960, of Lyndon B. Johnson in 1964, and of Richard M. Nixon in 1968. In 1905 Portsmouth was the site of the peace conference that ended the Russo-Japanese War, and in 1944 the state was host to the Bretton Woods Conference that founded the International Bank for Reconstruction and Development and the International Monetary Fund. The U.S. Air Force constructed a huge Strategic Air Command base at Newington in 1957, and the Portsmouth Naval Shipyard, in Kittery, has continued to specialize in building and repairing submarines.

THE NATURAL AND HUMAN LANDSCAPE

Physical environment. The basic physical features of New Hampshire are the result of the last glacial age (10,000 to 64,000 years ago), during which the "Wisconsin ice sheet" moved like a huge bulldozer across New England from the northwest to the southeast. Loose sand, silt, clay, and gravel were deposited as masses of glacial till that, near the town of Greenland, is 395 feet in depth. The mountain notches of New Hampshire—Carter, Crawford, Dixville, and Franconia—are the result of this glacial action, as are the potholes and cirques, amphitheatre-like recesses, found in the state. The great glacier left deposits in the form of many deltas and hillocks of stratified deposits. New Hampshire's lakes are also the results of glacial action.

Glacier-formed topography

Rivers and soils. New Hampshire has five main river-drainage basins. The largest is that of the Merrimack River in the central part of the state. Second in size is the Connecticut River drainage basin along the western border. The remainder of the state feeds the Saco and the Androscoggin rivers, as well as several smaller streams, including the Piscataqua, that are known collectively as the coastal rivers.

Mountains. The mountains are the most striking feature of New Hampshire's landscape. There are about 1,500 classified elevations, including eight mountains rising more than a mile in altitude, 61 rising more than three-quarters of a mile in height, and 157 with an elevation of a half-mile or more. The best known is Mt. Washington at 6,288 feet, the third-highest peak in the U.S. east of the Mississippi River. The mean elevation of the state is about 1,000 feet above sea level.

Climate. New Hampshire's climate is highly varied. In the winter the temperatures may drop below 0° F (−17° C) for days at a time. Summers are relatively cool, and the mean annual temperature is about 40° F (4.4° C). The annual precipitation is approximately 40 inches and is rather evenly distributed over the four seasons.

Vegetation and animal life. Even today, more than 80 percent of New Hampshire remains under forest cover. The majority of the trees found in the eastern U.S. are indigenous to New Hampshire. The most valuable single species of tree has always been the white pine, which has been characterized as "nature's chief gift to New England." Invaluable for masts in the era of sailing ships, these great trees also furnished useful lumber.

The wooded areas also support a flourishing wildlife. Whitetail deer are numerous, and moose are occasionally seen. There is a hunting season for deer each year, but the killing of moose is prohibited. Beavers, once almost exterminated, have been protected in recent years and are making a comeback. Black bears are relatively common, while smaller mammals like rabbit, squirrels, raccoons, foxes, and mink are still plentiful. There is an abundance of bird life, including such species as grouse, woodcock, pheasants, and ducks.

Until recent years, many varieties of fish were available in both coastal and interior waters. State rearing stations keep the interior lakes and rivers well stocked for fishing. During the 1960s, however, there was much concern about the effects of pollution on the under-water world, and strenuous efforts, both public and private, are underway to prevent further contamination of lakes, streams, and coastal waters.

Patterns of human settlement. Six quite distinctive regions exist, each with its roots deep in New Hampshire's history. The heavily forested White Mountain area in the north is popular with mountaineers and tourists in summer and winter alike. The lakes region centring on Lake Winnipesaukee is a favoured locality for summer camps and resorts and for water sports. The seacoast region around Portsmouth, Dover, Exeter, and Hampton has many maritime activities. The south central, or Merrimack, region, surrounds Manchester, the most heavily industrialized section of the state. The Dartmouth–Sunapee Lake region in the west central portion is dotted with educational institutions and summer homes. Finally, the area around Mount Monadnock, in the southwestern corner of the state, is noted for many small industries and such attractions as the MacDowell Colony, a residential retreat in Peterborough for artists in many fields, and for the Cathedral of the Pines in Rindge, an outdoor shrine dedicated to America's war dead. Each of these regions is officially organized and finances its own promotional activities.

Historic regions

The population of the state in 1972 was distributed among 13 cities, 224 incorporated towns, and 24 unincorporated places. The 11 percent of the state's area that comprises the massive White Mountain National Forest is almost uninhabited.

THE PEOPLE OF NEW HAMPSHIRE

Demographic trends. In the first U.S census, in 1790, New Hampshire had a population of 141,885. Since then each decennial count recorded a growth, except for that of 1870, when there was an extensive post-Civil War exodus to the new states of the Middle West. In the first census the urban population of the state was 3.3 percent, but in 1970 almost 60 percent of the state's 737,681 inhabitants were classified as urban. The people of the state were concentrated to a large extent in the southern and southeastern regions, and the larger urban centres were all located south of the White Mountains. In the area nearest to Boston, Massachusetts, New Hampshire was becoming a kind of "bedroom suburb" for thousands of persons who commuted to that metropolis each day. In the expanding population, persons over 65 years of age totalled 11 percent of the citizenry in 1970, a figure among the highest in the nation.

The vital statistics of the state revealed that the lowest birthrate in modern times was in 1933 and that the highest was in the years just after World War II. The highest death rate in the 20th century was in 1918, the year of the worldwide influenza epidemic.

Ethnic and religious origins. In the colonial period, the majority of the people were of English origin, but in 1719 there began a significant influx of Scotch-Irish who were largely Presbyterian in faith. Settling in the south central and southwestern portions of New Hampshire, they named their principal towns Derry, Londonderry, Antrim, and Dublin. These Scotch-Irish were energetic pioneers, excellently suited for fighting Indians, the British, or any other enemy who confronted their adopted country. They gave yeoman service in the Revolution and all subsequent wars.

Immigration and religion

New Hampshire never had a state church, even in colonial times. Rather, it had a unique system of "town churches" in which any officially recognized denomination could be designated at the annual town meeting as the one to receive public tax support. Prior to the American Revolution, there were five possible denominations from which to choose: Congregational, Baptist, Presbyterian, Quaker, and Church of England. In 1819 this system was discarded by the Toleration Act passed by the legislature. Since then all churches have been privately supported, and any denomination may function freely.

In the 19th and 20th centuries, new waves of immigration came into the state from central Europe, from the Balkans, and from countries peripheral to Russia, such as Poland and Finland. These newcomers brought with them their native customs and varied religious faiths. The first Roman Catholic congregation was established in 1828, the first parochial school in 1859, and a statewide diocese in 1884. The first Jewish congregation was organized in 1892, and the first Greek Orthodox Church in 1905. One of the only two Shaker communities still functioning in the U.S. is located in Canterbury. The Cathedral of the Pines attracts thousands of visitors each year to its impressive site. The New Hampshire Council of Churches, organized in 1945, developed broad ecumenical policies in which many faiths take part.

From the beginning, the people of New Hampshire have been mostly white. In 1970 black citizens made up less than 0.4 percent of the population, and citizens of Asian origin were even fewer. The largest group not directly descended from origins in the British Isles was the French-Canadian, or "Canado-Americaine," who first began to arrive in the years immediately after the Civil War, chiefly from Quebec. They are attracted mainly to such industrial cities as Manchester, Nashua, Laconia, and Berlin. By the end of World War I, New Hampshire had the largest percentage of citizens of French-Canadian origin of any of the states.

French-Canadian population

THE STATE'S ECONOMY

New Hampshire manifested some striking economic trends in the 1960s. The historic shoemaking, woodworking, apparel, and textile industries showed a marked drop in productivity and employment opportunities, while space-age industries grew rapidly. Among nonmanufacturing industries that also grew significantly were communications and public utilities, insurance and banking, medical and health-service establishments, miscellaneous business services, and wholesale and retail trades. Employment in New Hampshire was less affected by the recession that began in 1969 than were most other parts of the nation, and at the end of 1970 the state's unemployment rate was the lowest in New England. Per capita personal income in the state in 1970 was $3,608, ranking New Hampshire 26th in the nation in the average earning power of its people.

Agriculture, forests, and fisheries. Agricultural acreage decreased by two-thirds between 1910 and 1970, but, with an even greater decrease in the number of farms, the acreage per farm has increased. Poultry and poultry products are the chief source of income, followed by dairying and vegetable farming. Two-fifths of the harvested timber was for sawlogs, one-fifth for pulpwood to supply paper and newsprint industries, and most of the remainder for firewood. With only Rockingham County

on the coast, the state ranks low in commercial fishing yields but has marketable catches of lobsters, oysters, clams, and deep-sea fish.

Manufacturing, mining, and tourism. Electrical- and electronic-goods manufacturing has largely supplanted the textile and garment industry, much of which after 1945 either closed or moved to the South in search of cheaper labour and greater proximity to raw materials. Paper and wood products comprise another important industry. The only large industrial centre north of the White Mountains is Berlin, with large paper and pulp mills.

There are many rock quarries, chiefly of granite, throughout New Hampshire, but modern construction methods have reduced the demand for stone and depressed the industry. Other native minerals that have commercial value include garnet, zircon, beryl, and bog iron. In 1955 the state legislature established an Industrial Park Authority to attract new business into New Hampshire. Supplemented by the endeavours of individual cities and towns and by various private organizations, the authority has been successful in modernizing the industrial and business life of the state. An organized effort began in 1925 to publicize New Hampshire as a tourist area, and this effort has steadily expanded since that date. The great popularity of winter sports after World War II made New Hampshire as attractive to visitors in the cold months as it has been in the summer. Tourism became a major economic aspect of life in the Granite State.

Transportation. At its maximum extent, in 1920, the New Hampshire railroad network comprised 1,252 miles of trackage; 50 years later it was barely half that figure. Hundreds of miles of right-of-way had been abandoned, and there was no passenger service except on a six-mile-long cog railway on Mount Washington. This line had opened in 1869, the first such construction of its kind in the world. A century and more later it was still carrying tourist passengers behind steam locomotives during the summer season.

As the railroad disappeared, the bus and highway systems of the state greatly increased. In 1970 state highways totalled more than 4,100 miles; local roads, over 10,000 miles; and the interstate-highway system, over 210 miles, some of it still under construction. Air service to the chief cities of the state was maintained at a decreasing level. Most of the nearly 40 airports were privately owned, and only a few of the principal cities had either regularly scheduled airline service or adequate runways and landing aids.

ADMINISTRATION AND SOCIAL CONDITIONS

Constitution and governmental structure. The constitution of New Hampshire, the second oldest among the 50 states, was adopted in 1784. Every seven years the voters of the state may vote on the question of holding a convention to consider modifications of the basic law. Proposals that pass these conventions must be approved by two-thirds of the voters at a popular referendum. Fifteen such conventions have been held in the state's history, and a large number of modifications have been adopted, but a considerable portion of the original constitution remains in effect.

Executive power. Since 1784 New Hampshire has had 72 governors, some of whom in the early years were re-elected repeatedly. The normal incumbency in modern times has been two terms of two years each, but a determined effort is under way to establish a four-year gubernatorial term. The governor is assisted in his duties by a five-man council, an institution that is found elsewhere in the United States only in Maine and Massachusetts. Members, elected every two years from five geographical districts in the state, must approve all appointments to state offices and to judicial posts.

Legislature. The state legislature, the General Court, comprises at least 400 members elected every two years. The lower body, the House of Representatives has between 375 and 400 delegates and in the English-speaking world is believed to be exceeded in size only by the United States House of Representatives and the British Parliament. The state Senate has a modest membership of 24. In 1971 the Citizens Conference on State Legislatures, a privately funded study group, ranked the New Hampshire legislature 39th from the top in terms of effectiveness as a lawmaking body.

Judiciary and police. The state's judicial functions take place in three levels of courts, municipal and district courts, county superior and probate courts, and the state Supreme Court. All justices are appointed by the governor and his council and serve during good behaviour or until they are 70. The state police, in addition to their usual duties, are at the disposal of the governor for emergency assignments. Local police are under the jurisdiction of the municipal authorities.

Local government. The traditional town meeting, held annually on the first Tuesday in March, is the keystone of local government in New Hampshire. One of the last vestiges of pure democracy in the world, the town meeting is basic to community life in New Hampshire. Presided over by an elected moderator, the voters ballot on each article in the town warrant for the year. If adopted, these articles become the guide to town action for the ensuing 12 months. The meeting chooses the executive of the town and one of three selectmen, each of whom serves for three years. The meeting also selects such departmental heads as fire chief, treasurer, town clerk, road agent, and members of various special boards and commissions. In some towns the selectmen are empowered to choose a town manager, who conducts the day-by-day affairs of the community. Cities are governed by mayor–council, council–manager, or commission systems.

The annual town meeting

Politics. New Hampshire has always been a two-party state, and third-party candidates traditionally have had little success with Granite State voters. For decades prior to the Civil War, New Hampshire was overwhelmingly Democratic, but from then to the depression of the 1930s it was dominated by the Republicans. In recent years, however, the voters have opted on several occasions for Democratic presidential candidates—Woodrow Wilson, Franklin D. Roosevelt, and Lyndon B. Johnson—and have sent several Democrats to Congress.

Patterns of taxation and state aid. Taxes in New Hampshire present a confusing picture. New Hampshire clung tenaciously to its distinction of having neither a general sales tax nor a comprehensive state income tax. More than 60 percent of state income was derived from taxes on alcoholic beverages, tobacco, legalized racetrack betting, and the sweepstakes that went into effect in 1964. Such taxes admittedly were designed to draw heavily on the immense volume of tourism. The overwhelming burden of local income was derived from property, with non-property local taxes trivial.

The resultant problem of state aid to local units of government was an increasing and an abrasive one. Whereas, nationally, state aid provided nearly 30 percent of all local general revenue, in New Hampshire it provided only 7 percent. This resulted in such figures as New Hampshire's 1970 rank of 32nd in state aid to welfare services, 42nd in state aid for local education, 47th in support of its state university, and 49th in programs of state aid for vocational rehabilitation. Its health and welfare services, especially to the mentally retarded, were augmented in the 1960s, but there was agitation for further improvement.

Financial difficulties

Despite intricate and complex formulas for state aid devised by successive legislatures, the absence of any broad-based changes in the tax structure of the state kept the fiscal picture gloomy in the early 1970s. A 3 percent general sales tax proposed by the governor in 1971 polarized opinion into grudging support and vehement opposition, but it failed to pass the legislature.

Education. New Hampshire has had a public-school system since 1647, when, as a part of Massachusetts, it was required to provide different kinds of schools depending on community size. The statewide system is under a board of education headed by a commissioner. Each town is constituted a school district, with its main support coming from local property taxes. The land-grant college that became the state university was found-

ed in 1866 in Hanover and moved to Durham in 1893. Affiliated colleges are located at Plymouth and Keene. During the 1960s six state vocational colleges and a technical institute were established; an educational-television system based at the state university went into operation in 1959.

Best known among the private institutions of higher learning is Dartmouth College at Hanover, a liberal arts college that observed its bicentennial in 1969. Other private colleges include St. Anselm's College (1889) at Manchester, Mount Saint Mary College (1934) at Hooksett, and Rivier College (1933) at Nashua, the last two for women only. Since 1945, eight new four-year liberal arts colleges, all privately financed, have been founded in the state. In addition to these, there are several private junior colleges, the oldest of which is Colby College-New Hampshire (1928) at New London.

CULTURAL LIFE AND RECREATION

The arts. New Hampshire has several outstanding cultural institutions. The MacDowell Colony was founded as a memorial to the composer and Peterborough resident Edward MacDowell (1861–1908) by his widow in 1907, and it since has been a mecca for musicians and writers. In 1964 the 86-acre former home of the noted sculptor Augustus Saint-Gaudens (1848–1907) at Cornish was designated as a National Historic Site. Manchester has the Currier Gallery of Art and an institute of arts and sciences; Dartmouth College a distinguished fine arts centre; Nashua an arts and science centre; and several towns have smaller galleries, art centres, and museums. Summer theatres flourish in a dozen or more of the state's resort areas.

Sporting and historic sites. Recreation and tourism are important parts of the cultural pattern of the state. New Hampshire has more than 30 ski areas, many of which operate their lifts in the summer for sightseers. There are more than 200 youth camps and more than 60 golf courses. Numerous historical sites throughout the state include many restored colonial homes and Ft. William and Mary at New Castle. There are more than 40 state parks and over 800,000 acres of publicly owned forest preserves.

PROSPECTS

Despite its conservatism, New Hampshire has shown evidence of facing the future with a willingness to seek new ways. By the early 1970s, more than half the towns in the state had established conservation commissions to protect the ecology. The New Hampshire Council on the Aging sought to assist in geriatric matters. Hundreds of "tree farms" total more than 500,000 acres. There was a Center for Continuing Education at the university in Durham and a State Council on the Arts.

One of the assets with which New Hampshire can face the future is the stabilizing force of historical continuity, which may allow it to evaluate the direction of needed innovation. In almost every town and city there are families whose successive generations have lived on the same property, even in the same house, for a century or more. There are other less concrete traditions of the land that modernity and sophistication cannot erode. It was not by accident that Thornton Wilder's famous play *Our Town* (1938), depicting many universal aspects of human experience, was set in rural New Hampshire. It was not wholly fortuitous that the noted American poet Robert Frost (1874–1963) lived most of his life in the state. It is not coincidence that "Old Home Week" has been proclaimed by the governor every year since 1913.

BIBLIOGRAPHY

General: J.D. SQUIRES, *The Granite State of the United States*, 4 vol. (1956), the only large-scale history of the state in the last 40 years; see also the same author's *The Story of New Hampshire* (1964).

Other works: F.A. BURT, *The Story of Mount Washington* (1960), a vivid account of the state's highest mountain; C.E. CLARK, *The Eastern Frontier: The Settlement of Northern New England 1610–1763* (1970), an admirable review of the frontier era; M. CLEVELAND, *New Hampshire Fights the Civil War* (1969), definitive work on this topic; D.B. COLE, *Jacksonian Democracy in New Hampshire, 1800–1851* (1971), an excellent analysis; J.W. GOLDTHWAIT, *The Geology of New Hampshire* (1925), readable and helpful; W.R. HARD, *The Connecticut* (1947), and R.P.HOLDEN, *The Merrimack* (1958), two fine surveys of these river valleys; C.B. KINNEY, *Church and State: The Struggle for Separation in New Hampshire, 1630–1900* (1955), the most complete review of this intricate subject; M. MELCHER, *The Shaker Adventure* (1941), the best work on this theme; W.G. SALTONSTALL, *Ports of Piscataqua* (1941), an admirable survey of maritime New Hampshire; E.C. TEAGUE, *Mount Washington Railway Company* (1970), a recent history of the oldest mountain cog railway in the world; M. WADE, *The French Canadians, 1760–1945* (1955), a comprehensive survey of an important ethnic minority in New Hampshire; H.F. WILSON, *The Hill Country of Northern New England: Its Social and Economic History, 1790–1930* (1936), a classic work.

(J.D.S.)

New Jersey

Although it is a major social, economic, and political force in its own right, New Jersey often is looked upon as a stepchild among the heavily industrialized and populated Eastern Seaboard states of the United States. Only four states in the nation are smaller than its 7,836 square miles (20,295 square kilometres), but only seven had more inhabitants in the 1970 census than the nearly 7,170,000 New Jerseyites counted. As a geographical entity and as a human collective, New Jersey nonetheless suffers from a massive lack of identity among American states. For some 300,000 of its citizens who commute to New York, on the north and northeast, and to Pennsylvania, on the west, New Jersey is a vast "bedroom state." For its transportation system, one of the most extensive and busy in the world, it is primarily a funnel for goods and people moving between New York City and points north and Philadelphia, Delaware, and points south. For hundreds of thousands of out of staters it offers long stretches of fine beaches along the Atlantic Ocean on the east and south, although the name and image of the resort Atlantic City may be better known than those of the state itself.

Contradictory nature of the state

Above all, New Jersey is rife with contradiction and anomaly. Its people fiercely fight off the challenges of state government to complete home rule by powerful municipal administrations, which have contributed heavily to the almost uniform depression and decay of its cities. Whereas it has produced some of the most able and respected American governors in the 20th century, its politics have been astonishingly corrupt, and, in recent years, it has achieved notoriety as a major haven for organized crime. It is a wealthy state but beset with fiscal problems, and its citizens tax themselves heavily for public education but consistently refuse to spend money on higher education.

New Jersey is called the Garden State because of its many prosperous truck farms, yet it is among the most urbanized and crowded of states. The urban glut of its northeast contrasts sharply, however, with the rugged hills of the northwest and the enormous stretches of pine forest and rolling, lush horse country. It is one of America's most important industrial centres, but it has paid the price in environmental pollution, in dirt and noise, and in congested roads and slums. In sum, New Jersey is a curious amalgam of the urban and rural, the poor and wealthy, the progressive and backward, the parochial and cosmopolitan. Although it was one of America's 13 original states and the third to ratify the constitution, the New Jersey of the late 20th century has many earmarks of a political body unable to come to terms with itself. (For information on related topics, see UNITED STATES; UNITED STATES, HISTORY OF THE; and NORTH AMERICA.)

THE HISTORY OF NEW JERSEY

The colony. Before the Europeans came, the Lenni Lenape (or Delaware) Indians spent the summer at the shore feasting on fish and collecting the shells they made into wampum. This is what the Florentine explorer Giovanni da Verrazano found in 1524, when he became the first white man to touch shore in New Jersey. Almost a century passed before colonization began with the arrival in 1609 of the English navigator Henry Hudson, who

sent a party to explore Newark Bay. The first settlers were English, Scots–Irish, Swedes, and Dutch. The colony was brought under English rule in 1664, although for the next five years the Dutch disputed that claim. In 1676 the province was divided into East and West Jersey, the former going to Sir George Carteret and the latter to a group of Quakers. The division continued until 1702, when all of the province reverted to James II.

Unlike other colonists, who suffered from the harshness of English rule, the early New Jerseyites were of such an independent and misanthropic nature that it was the royal governors who did much of the suffering. Until 1738, New Jersey and New York were ruled by a single governor. When Lewis Morris took office as the first governor of New Jersey alone, one member of the Assembly advised his colleagues on how governors should be treated: "Let us keep the doges poore and wee'll make them do as we please."

Revolution and statehood. Typically, considerable division of sentiment occurred within the state over the American Revolution, and Tory activity was heavy. The most significant battle of the conflict was fought in New Jersey on December 26, 1776. Gen. George Washington and his hungry, ragged troops crossed the Delaware River from Pennsylvania in longboats, surprised the garrison of German mercenaries in Trenton, and captured the city. A week later, Washington won another vital battle at Princeton, routing Lord Cornwallis, commander of the British army. The victories breathed new life into the Revolution, and an army of colonials near despair was transformed into an effective fighting force. Between the Revolutionary and Civil wars, New Jersey underwent tremendous industrial development, largely abetted by the construction of canals and, later, railroads. The railroads, particularly the Camden and Amboy (forerunner of today's Penn Central), played a crucial role in the state's political life, dominating and controlling legislators and governors during the "robber baron" era of industrial expansion in the 19th century. Accommodating tax laws of that era gave New Jersey the epithet "The Mother of Trusts": one-half of the nation's largest corporations made their headquarters in the state. Public dissatisfaction with the power of the trusts and public utilities helped pave the way for the election of Woodrow Wilson as governor in 1910.

During the Civil War, Southern sympathizers operated almost openly in New Jersey, although the state contributed nearly 90,000 troops to the Federal army. Industrial growth continued through and after both world wars, and, until after the middle of the 20th century, the growing decay of the cities was lost sight of amid the general prosperity that industry brought with it.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Topography.* New Jersey comprises four distinct physical regions—the rugged hill country of the northwest, where the Appalachians slice across; the rolling horse country just below the hills; the central plain of the northeast, where the major cities and suburban towns lie; and the coastal plain of the south. In the thinly populated northwest, the soil is heavy and contains much limestone and sandstone. The highest elevation, at High Point, is 1,803 feet (550 metres) above sea level. Southward, the hills become lower and gentler, and the soil contains considerable granite; this gives way to soil dominated by shale. The central plain is a mixture of loams and marl, while the coastal plain is composed mainly of sands and loam. The most striking features of the state are its beaches, the hills of the northwest, and the marshy meadowlands of the northeast.

The diverse physical regions

Lakes and ponds cover about 315 square miles of the state's surface. New Jersey's major river, which it shares with Pennsylvania, is the Delaware. The Hudson River separates the state from New York. Other major rivers are the Passaic and the Hackensack, both in the northeast, and the Raritan, which runs west to east and is generally regarded as the boundary between North and Central Jersey. Hopatcong Lake in Sussex and Morris counties is the state's largest.

Climate. The climate is moderate, although there is a marked difference in temperatures between the northwest and the southeast. In the north, the average temperature is 50° F (10° C)—in winter, 28° F (−2° C) and in summer, 70° F (21° C). In south, the annual average is 55° F (13° C)—in winter, 35° F (2° C) and in summer, 71° F (22° C). Rainfall averages 46 inches a year—36 inches (914 millimetres) in South Jersey and 50 inches (1,270 millimetres) in North Jersey.

Vegetation and animal life. Virtually all of the plant life common to the northeastern United States can be found in New Jersey, and in the marshes and Pine Barrens of south New Jersey many rare plant species grow, including some insect-ingesting plants. The Barrens are dominated, however, by many varieties of pine and white cedar. Oak, elm, birch, ash, sweetgum, maple, walnut, and chestnut trees are found in river valleys. Plants common to the state include wild azalea, rhododendrum, honeysuckle, mountain laurel, wintergreen, and cardinal flower.

The Hackensack Meadows and the Great Swamp in Morris County are dominated by grasses. Whereas the ecology of the Swamp is sound, centuries of dumping, filling, and dredging have destroyed much of the plant and animal life of the Meadows. Plume grass is the main plant in the Meadows, but it is of negligible value to fowl or fish. Muskrats abound in both marshes. The increasing suburban development of New Jersey has encroached on wildlife habitats, although bear and wildcats can still be found in the Barrens and mountains. Deer are plentiful, as are opossums and muskrats. Other animals encountered include rabbits, squirrels, chipmunks, woodchucks, and many varieties of snakes. All of the birds common to the Northeast are found in New Jersey, as are many migratory varieties.

Traditional regions. *The Jersey Shore.* The most distinctive of New Jersey's regions is its 127-mile-long shore, much of it composed of long and narrow sandbars separated from the mainland by bays and inlets that form part of the Intracoastal Waterway, an interconnected water highway that stretches all the way to Florida. Cape May, at the southern tip of the state, was America's first summer resort, and both that community and Long Branch in Monmouth County were known as the playgrounds of presidents during the 19th century. Today, the quality of the Shore ranges from the urban garishness of Asbury Park to the opulence of Deal and Mantoloking. In such resorts as Wildwood and Atlantic City, the night life plies its revels until dawn, whereas such other towns as Avalon, Ocean City, and Beach Haven are basically family resorts. The Jersey Shore at its best can be found in its two state parks, Sandy Hook in the north and Island Beach in the south. There, the dunes are still topped with coarse but fragile grass, and the osprey still build their nests. The marshes teem with wildlife, and the trees are bent and twisted by wind and salt spray.

Human uses of the land

The metropolitan area. Five northeastern counties in the New York metropolitan area—Essex, Hudson, Passaic, Bergen, and Union—contain nearly one-half of New Jersey's population. Four of the state's six largest cities —Newark, Jersey City, Paterson, and Elizabeth—are located here. The Newark–Hudson County–Elizabeth complex appears to many travellers as one endless, smoky, industrial city: dingy, smelly, but throbbing with commercial life.

Beyond the cities lie the suburbs. Most are pleasant and prosperous, but some are old and showing the first signs of the blight that began to eat away at the cities decades ago. Industrial construction in suburban communities was increased tremendously in recent years, but many suburban towns, especially in Bergen County, remain bedroom communities of New York City and of the New Jersey cities. Newark's population of almost 400,000 doubles every day as the work force pours in. The state remains dominated, however, by the two giants just beyond its borders. An estimated 160,000 New Jerseyites cross the Hudson to New York by rapid-transit trains, by the main line of the Penn Central, or by bus and automobile through the Holland and Lincoln tunnels or across

the George Washington Bridge. North Jerseyites watch New York television, root for New York athletic teams, and patronize New York theatres and restaurants. The same situation exists in Camden, Burlington, Gloucester, and Salem counties where 95,000 persons cross the Delaware to jobs in Philadelphia and environs.

For all of this, however, New Jerseyites have an antipathy for these two big cities. Many citizens of the Garden State grew up in New York or Philadelphia, and, although they retain a lively interest in events there, they profess a profound gladness at being out of the city, with its crime and its financial and housing problems. This attitude is shared by those who never lived in the cities but who grew up in their shadows.

South and Central Jersey. Just below Trenton, the state capital, begins South Jersey, comprising all or most of eight counties, roughly one-half of the state's area but less than one-fourth of the population. Although the climate is drier than in the north, the loamy soil is well suited to vegetable farming, and most of the land not covered by forest or marsh is farmed. Less than 1 percent of the state's population is engaged in farming, but farm income per acre is the highest in America.

Central Jersey, all of five counties and part of three, is largely plain, but hilly areas occur in Hunterdon and Somerset counties. Middlesex and Mercer counties, especially the former, are heavily industrialized. Princeton University is located in Mercer County and the borough of Princeton, which combines the charm of the campus with a rich colonial past to create one of the nation's loveliest towns. Rutgers, the State University is in nearby New Brunswick. Hunterdon and Somerset counties are a pleasant mixture of farm and woodland.

The northwest. The four counties of the northwest take in a mixture of small town, affluent suburb, and rugged countryside, although two major cities, Passaic and Paterson, are in Passaic County. The area contains numerous dairy farms and parks and recreation areas.

The Meadowlands. Although relatively small, the New Jersey Meadowlands loom large in the state's future. Stretching from Perth Amboy to Hackensack, they cover some 19,000 acres (8,000 hectares), 12,000 of which are undeveloped. The lack of development results from disputes between the state and private interests over ownership and from the many problems of construction in a marshy area. Located as it is in the centre of the world's busiest metropolitan area, however, the value of the Meadowlands is immense. In the late 1960s, New Jersey put together the machinery to develop this area through the New Jersey Meadowlands Commission. Initial plans call for a sports complex and ultimately a new city between the Passaic and Hackensack rivers. In 1971 the New York Giants professional football team signed contracts to play in the new arena by 1975, and the possibility existed that the New York Yankees baseball franchise would follow them.

Disputes over land development

THE PEOPLE OF NEW JERSEY

Ethnic composition. New Jersey's first settlers were the English, Scots–Irish, Swedes, and Dutch, but its population today more nearly reflects the immigration patterns of the 19th and 20th centuries: Germans and Slavs, Russian and European Jews, Irish and Italians. New Jersey was a prime destination for the waves of blacks that left the South during and after World War II; by 1970 they made up 10.7 percent of the population. The state has a sizable Puerto Rican population, and Cubans who left their country after the Castro revolution and settled in North Jersey, mainly Hudson County, comprise the latest group to make its presence felt.

Italian-Americans are now the state's largest ethnic group. They are the predominant Caucasian bloc in the cities, although the cities also contain sizable Polish, Hungarian, and other eastern European groups. The Italian-Americans and blacks dominate the political and cultural life of the cities—a situation that often has brought the two groups into competition and conflict.

Demography. The population is unequally distributed. Roughly three of every four New Jerseyites live north of Trenton, and 60 percent of the population lives within 30 miles of New York City. The 18.2-percent growth during the 1960s was a drop from the 25.5-percent figure recorded in the 1950s. Population density was, however, the greatest in the nation.

The most striking demographic trend in New Jersey, as in the nation as a whole, is the movement of the white population away from the cities and the concurrent proportional growth of the urban black population. Between 1960 and 1970, the populations of Newark, Jersey City, Trenton, and Camden showed marked decreases, while those of Paterson and Elizabeth inched upward. Urban Hudson County was the only county to post a loss in population, though slight. Newark, with a population of some 380,000, remains the state's largest city and one of only a handful of American cities with a black majority; in 1970 Newark became the first major city in the Northeast to elect a black mayor. The exodus from the cities is directed, as elsewhere in the nation, toward suburban and rural regions. Areas only a few years ago considered rural are becoming suburban or even urban in character. In the 1960s the population of rural Ocean County jumped more than 90 percent, and other rural counties experienced growth of close to or more than 50 percent. This process is solidifying New Jersey's position as the most urbanized of states after California: about 89 percent of New Jerseyites live in areas designated by the Bureau of the Census as urban. Three counties—Essex, Union, and Hudson—are 100 percent urban, and four more—Bergen, Camden, Middlesex, and Passaic—are more than 95 percent urban.

Patterns of suburban and rural migration

The flight of the white middle class from the cities has been accompanied by an emigration of industry and commerce. Old, outmoded factories are left behind for sleek new buildings out of town, and huge suburban shopping centres have replaced the downtown department stores. This shift means more jobs in the suburbs and rural areas, and the jobs create a commensurate demand for housing. In the end, the homes and apartments probably will go where the jobs are—in the suburbs. The cities and towns bordering these areas will become increasingly black, a situation that is potentially explosive because of the continuing decay of the cities and the poor quality of housing and services the cities can supply. Clearly, the future promises great pressures for, as well as great resistance to, low-cost, integrated housing in the suburban counties.

THE STATE'S ECONOMY

Alexander Hamilton's attempt in 1791 to build the nation's first industrial town at Paterson was a failure. He had the right idea, however, for New Jersey was destined to become an industrial giant. The state now ranks seventh nationally in per capita income and seventh in value added by manufacturing.

Between 1960 and 1970, total buying income in New Jersey increased by 90 percent and retail sales by 12 percent. The work force increased by 22 percent, or 548,000 people, and value added by manufacture by 65 percent. New Jersey's gross state product in 1970 was $40,000,000,000, and total personal income was $33,000,000,000. Average income per household was just under $15,000. One of three New Jersey workers is in manufacturing, compared with a national average of 25 percent. The New Jersey worker earns about 2 percent more than the national average.

Economic activities. One of New Jersey's largest and most prestigious industries is research. New Jersey has the highest number of engineers and scientists per capita in the nation, and New Jersey firms spend 10 percent of the nation's research dollars. Although it is not reflected in the manufacturing statistics, a close correlation exists between research and heavy industry. America's greatest inventor, Thomas Edison, was a New Jerseyite, and he established the nation's first research laboratory in Menlo Park in 1876. There he created the electric light, the phonograph, and all manner of electrical gadgets, as well as pioneering the technology of the motion-picture industry. Edison's primitive laboratory of a century ago has been succeeded by the David Sarnoff Research Center,

Leadership in scientific research

the Bell Telephone and Western Electric laboratories, the Institute for Advanced Study at Princeton, the James Forrestal Research Center, and many more sleek, modern centres dotting the landscape of Central Jersey. New Jersey's major manufacture is electronics products, closely followed by chemicals, in which it leads the nation. It holds high national ranking in the production of clothing, electrical machinery, toys, sporting goods, and stone, glass, and clay products. The state is not noted for its mineral deposits, but the iron found in the South Jersey bogs provided much of the shot and cannon used during the Revolution. Minerals mined today include trap magnetite, iron, uranium, and zinc, as well as sand and gravel.

The spread of industry and housing is costing New Jersey much of its farmland, the most valuable per acre in the nation. Slightly more than 1,000,000 acres remain under cultivation, and gross farm income runs well over $250,000,000 a year. The resort industry is a big factor in New Jersey's economy, especially in the south, where a bad year at the Shore hurts the economic well-being of the entire region. Resorts account for some $2,000,000,000 a year, most of this coming during the summer months at the Shore. But Atlantic City, with its huge Convention Hall and fine old hotels, does a thriving winter convention business. Because of its many industries, the New Jersey economy is highly diversified and usually reflects the health of the national economy. At the end of the 1960s, the unemployment rate was climbing rapidly, reaching over 7 percent by the summer of 1971. The long-dormant housing industry was showing signs of recovery, however, and the New Jersey Economic Policy Council predicted slow but steady economic growth.

The State Division of Economic Development, along with the major utilities and business organizations, conducts an effective program of selling New Jersey to industry. During the late 1960s and early 1970s, the state attracted many new industries, especially corporate headquarters from New York City, largely through its greater space, better transportation, and favourable tax rates.

Transportation. Since colonial days, when New Jersey's toll roads linking Philadelphia and New York were a major industry, transportation has been the lifeblood of New Jersey's economy. The role of transportation in New Jersey can best be appreciated in the Newark area. There, 12 lanes of the New Jersey Turnpike, the busiest in the world, converge with the main line of the Penn Central Railroad, Newark Airport, Port Newark, and Port Elizabeth to provide a steady stream of goods and humanity in a hurry. It is perhaps the busiest transportation centre in the world. The economy of northern New Jersey is bound tightly to that of New York City, and the commercial traffic between the two states is the heaviest in the nation.

Port of New York Authority

In 1921 the states of New York and New Jersey formed the Port of New York Authority—a bistate commission empowered to finance and operate transportation facilities in the New York metropolitan area. The Port Authority, as it is known, is a public corporation operating Newark and Teterboro airports in New Jersey and La Guardia and Kennedy in New York, as well as the Lincoln and Holland tunnels, three bridges, huge piers, bus and rail lines, truck terminals, and two giant towers of the World Trade Center rising above Lower Manhattan.

A similar but much smaller transit complex exists in the Camden area, linking the South Jersey area with Philadelphia. There is a deepwater port at Camden on the Delaware and high-speed transit to Philadelphia. In addition to the turnpike, which travels the length of the state from the George Washington Bridge in the north to the Delaware Memorial Bridge in the south, the Garden State Parkway stretches 173 miles north to south, and the Atlantic City Expressway links that city with the Camden area. Before World War II, New Jersey had one of the finest road networks in the nation. A pay as you go fiscal policy kept the state from building roads fast enough, however, and the highways in North Jersey grew crushingly congested during the 1950s and 1960s. A bond issue approved in 1968 has helped ease the traffic problem, but it seems unlikely that the state can ever catch up with demands. There are about 43,000 miles of railroad track in New Jersey, the most per square mile in the nation. Ten railroads and seven trunk lines cross the state. New Jersey also has more airports per square mile than any other state, with 3,000 aircraft based at its 83 airports and 194 heliports. The largest and busiest field is Newark Airport, which services more than 7,000,000 passengers a year.

New Jersey's transportation future appears to lie in the area of mass transit, a theme that dominated the 1969 gubernatorial election campaign. In 1971 the Governor realized his first breakthrough when the Port Authority agreed to build a new rail line joining suburban Union County with Penn Station in Newark and with Newark Airport.

ADMINISTRATION AND SOCIAL CONDITIONS

New Jersey, on December 18, 1787, became the third state to ratify the United States Constitution. In the early days of the republic it maintained a tradition of legislative superiority, a condition that was somewhat modified in favour of the governor when a new constitution was adopted in 1844. Even then, however, New Jersey's chief executive was one of the weakest in America. Its people were distrustful of big state government, fearful that a strong central government would trample on municipal powers. By the end of World War II, the people were ready for a change. The result was the constitutional convention of 1947, which gave the governor strong new powers and created what is generally regarded as the strongest court system in the United States, with enormous powers vested in the chief justice.

Constitutional framework of government

Governmental structures. New Jersey governors serve terms of four years, and they are permitted re-election to a second term. No other state official runs on a statewide basis, so the governor has no rivals in the executive branch. He appoints, with the advice and consent of the Senate, virtually all top state officers and members of state boards, authorities, and judiciary, and he has the authority to supercede county prosecutors. By virtue of his broad executive and administrative powers, the vast patronage at his disposal, and his unequalled access to the press, radio, and television, the governor of New Jersey is as strong a chief executive as there is in the nation. A two-thirds majority of both legislative houses is necessary to override his veto. The governor can be checked and occasionally defeated by a rebellious legislature or political leaders, but, in the final analysis, governors seldom lose wars, only battles. In the state legislature, reapportioned in 1966 to conform to the "one man, one vote" principle, the Senate is composed of 40 members and the General Assembly of 80. Assemblymen are elected every two years, while senators run for one two-year term and two four-year terms each decade. As late as 1971 the state Supreme Court had failed to approve the reapportionment plan, although it was permitted to stand for the election of that year.

Local government

New Jersey's 21 counties are administered by boards of freeholders, elected on a county-wide basis. The boards vary in size from five to nine members, depending on the size of the county. In addition to these elected officials, the various levels of local government are supplemented by service commissions, boards, and authorities, many of which enjoy wide independence and even autonomy. Attempts to merge municipalities, reduce the number of school districts by consolidation, or strengthen county government have failed because of the fear of compromising local power and prerogatives.

Political life. In national elections, New Jersey is a "swing" state. It leans Republican, and it was the only major industrial state in the Northeast carried by Richard M. Nixon in 1968. In 1969 New Jersey elected its first Republican governor in 16 years, William T. Cahill; he succeeded Richard J. Hughes and Robert B. Meyner, each of whom served eight years in office. In only two of these years, 1966 and 1967, the Democrats controlled both houses of the legislature, a feat they last accomplished more than 50 years earlier when Woodrow Wilson was governor, a post he used as his springboard to the

presidency. Although there were signs that civic bossism was in the first stages of decline as the 1970s opened, New Jersey's political system was long dominated by powerful county leaders who drew their power from the patronage and contracts that they dispensed through control of the courthouse or city hall. The most notorious of these bosses was Frank Hague, who operated as the absolute ruler of Jersey City and Hudson County from 1917 to 1947. For three decades Hague dominated the Democratic Party and heavily influenced the Republicans. His philosophy of government was best summed up in his famous reply to critics: "I am the law in Jersey City."

Financial policies. For decades New Jersey politicians boasted that the state spent less per capita than any other in America. Whether this can be attributed to frugality or to the ignoring of social needs is a matter of debate. New Jersey collected the fourth-lowest per capita tax in the United States as late as 1966, when a 3 percent sales tax was enacted that in 1970 was raised to 5 percent. By 1971 informed opinion was virtually unanimous that New Jersey would have to enact a personal-income tax within a few years. Since New Jersey consistently has avoided the imposition of a broad-based tax for so long, it has been forced to rely heavily on the real-estate tax levied by municipalities to cover their own expenses and those of school districts and counties. At the same time, the state paid comparatively little aid to the local school districts. The result is a property tax in some communities, especially in the cities, that is among the highest in the nation. The state showed signs in the early 1970s of plunging into a tax-reform program, but then there was the difficult question as to whether enough time remained to halt the downward plunge of such major cities as Newark and Trenton.

Social services. Most of the services rendered the citizens of New Jersey come from the state, although most of the major counties maintain institutions of one kind or another, and much funding comes through federal agencies. The state has for several years faced a chronic shortage of hospitals, prisons, and various other institutions, although much has been done to correct the situation since the 1960s.

Higher education

New Jersey has long had a reputation for shirking its responsibilities to higher education, sending 55 percent of its college students out of the state for an education during the 1960s. Indeed, there was no state university at all until New Jersey took over full responsibility for Rutgers University in 1945. Under the leadership of Governor Hughes, the process of converting the six state colleges from strictly teacher-preparation institutes to full-fledged liberal arts schools was begun in the late 1960s. A medical school was begun in Newark during the decade, and in the 1970s construction was scheduled to begin on two new universities. New Jersey is also the home of Princeton University, one of the nation's most prestigious institutions.

Housing, especially in the core cities, is one of New Jersey's most difficult problems. In 1970 it was estimated that some 500,000 housing units in the state were substandard, and, in common with most Northern states, New Jersey was saddled with staggering and spiralling welfare costs.

CULTURAL LIFE AND INSTITUTIONS

The living arts. New Jersey has never been noted as a cultural giant because the vast majority of its citizens are well served by such attractions as the New York theatre or the Philadelphia Orchestra. Still, there are 18 symphony orchestras in the state, all but the New Jersey Symphony Orchestra small and community-oriented. In 1968, however, New Jersey helped correct its reputation as a cultural backwater with the opening of a beautiful new amphitheatre at the Garden State Arts Center, home of the New Jersey Symphony Orchestra. Located on Telegraph Hill off the Garden State Parkway in Monmouth County, the facility has proved to be an extraordinary success, and its programming of music, drama, and dance has been exceptionally well received.

There are 11 summer theatres in New Jersey, most of them located near vacation areas. The McCarter Theater, on the Princeton University campus, is open year round and offers high-quality plays and musical presentations.

Artistic and historical preservations. New Jersey has an unusually high number of museums, 120 in all, many operated in conjunction with historic sites or buildings. The New Jersey State Museum (with a planetarium), located in the state capitol complex in Trenton, the Rutgers Art Gallery, the Newark Museum, and the Princeton University Art Museum are among the better known.

New Jersey's rich traditions are manifested in such historic homes and sites as Washington's winter headquarters in Morristown, where he wrote his farewell address to the Continental Army; Morven, the official governor's residence in Princeton; the restored colonial villages of Batsto and Allaire; the Camden home of poet Walt Whitman; and many others that attract thousands of tourists each year.

Communications. New Jersey depends entirely on New York and Philadelphia network stations for its national television and local news. A state-operated public television system, on ultrahigh frequency, was initiated in 1971. The state's largest newspaper is the *Newark Star-Ledger*. The *Newark Evening News* and the *Newark Sunday News* were the most influential journals, but a long strike in 1971 forced the sale of the Sunday paper and a cutback in daily operations. Other papers include the *Record* of Hackensack and the *Sunday-Record-Call*, the *Trenton Times* newspapers, and *Courier-Post* of Camden.

PROSPECTS

The major challenge facing New Jersey in the years ahead is the regulation of growth so that pollution, congestion, and noise can be controlled and the dwindling supply of land utilized widely. The heavy concentration of population in the northeast makes water supply a special problem, and reservoir sites are scarce.

Problems of growth

Strong leadership will be needed at the state and local levels if the mistakes made in the cities are not to be repeated in the suburbs. If its cities are decaying, facing bankruptcy, and torn by racial turmoil, the situation is not unique to New Jersey, and only national action is likely to cure their ills.

Perhaps New Jersey's most severe and immediate problem is housing. Low- and middle-income housing in the suburbs is badly needed, but restrictive zoning laws and building codes prevented such construction in the 1960s. State and local officials seemed to achieve a new awareness of this problem in the early 1970s, however, and a series of court cases pointed to an end to these practices.

Most students of government and many political and governmental leaders believe that New Jersey will have to face up to governmental reorganization in the years ahead. They argue that there are too many individual municipalities and school districts and that the result is waste, duplication, and inefficiency. Leaders of county government say that the counties are the logical place to concentrate services on a regional basis, but state laws permit the counties little flexibility or responsibility.

In the early 1970s all of these problems were being debated in the state legislature, and the needs were clearly recognized. Local opposition to change, however, remained strong.

BIBLIOGRAPHY

Geography: *The Geology of New Jersey*, bulletin no. 14 of the state Department of Conservation and Economic Development (1915), covers the geography and geology of the state in great detail, as does the *Topographic Atlas of New Jersey*, which contains large maps showing railroads, canals, highways, mines, and quarries.

History: THOMAS FRANCIS GORDON, *The History of New Jersey, from Its Discovery by Europeans, to the Adoption of the Federal Constitution* (1834), gives the best account of the state's role in the Revolutionary War. FRANCIS B. LEE, *New Jersey As a Colony and State*, 4 vol. (1902); IRVING STODDARD KULL (ed.), *New Jersey: A History*, 6 vol. (1930–32); and WILLIAM STARR MYERS (ed.), *The Story of New Jersey*, 5 vol. (1945), are all thorough and detailed works.

Population: The best recent study of the state's ethnic background is JOHN E. BRUSH, *Study of the Population of New Jersey*, 2nd ed. (1958). See also HERSCHEL LEE SCHENECK, *Indians of New Jersey* (1951).

Economy: ADELAIDE ROSALIA HASSE, *Index of Economic Material in Documents of the States of the United States, New Jersey, 1789–1904* (1914), is a prime source for information on New Jersey's industrial economy during the 19th century. HOMER HOYT, *An Economic Survey of the State of New Jersey* (1950); and JOHN T. CUNNINGHAM, *Made in New Jersey* (1954), are also useful. For a current picture of the state's economy, the DEPARTMENT OF LABOR AND INDUSTRY, *Facets of New Jersey*, a periodically updated series of pamphlets, is the best and most concise account.

Culture: The FEDERAL WRITERS' PROJECT, *New Jersey: A Guide to Its Present and Past* (1939), contains an excellent section on the arts. LOLITA L.W. FLOCKHART, *Art and Artists in New Jersey* (1938), also contains valuable information.

Government: The NEW JERSEY LEAGUE OF WOMEN VOTERS, *Spotlight on Government* (1969), is a comprehensive, one-volume study of the state government. *Fitzgerald's Legislative Manual*, published annually, contains a wealth of biographical information on governmental leaders plus election returns, the organization of state departments, lists and memberships of agencies, authorities and commissions, the text of the state constitution, and many other important pieces of information. The New Jersey Tercentenary Commission in 1964 published a 26-volume *New Jersey Historical Series* covering a wide variety of topics. DUANE LOCKHARD, *The New Jersey Governor: A Study in Political Power* (1964), part of the series, is an especially readable and analytical source of information on modern government.

(J.Ke./J.McL.)